remote sensing

Special Issue Reprint

Ocean Remote Sensing Based on Radar, Sonar and Optical Techniques

Edited by
Xuebo Zhang, Haiyong Xu, Jingyue Lu, Lei Zhang,
Marc Pinto and Farman Ali

mdpi.com/journal/remotesensing

MDPI

Ocean Remote Sensing Based on Radar, Sonar and Optical Techniques

Ocean Remote Sensing Based on Radar, Sonar and Optical Techniques

Guest Editors

Xuebo Zhang
Haiyong Xu
Jingyue Lu
Lei Zhang
Marc Pinto
Farman Ali

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editors

Xuebo Zhang
College of Physics and Electronic Engineering
Northwest Normal University
Lanzhou
China

Haiyong Xu
Faculty of Information Science and Engineering
Ningbo University
Ningbo
China

Jingyue Lu
National Key Laboratory of Radar Signal Processing
Xidian University
Xi'an
China

Lei Zhang
School of Electronics and Communication Engineering
Sun Yat-sen University
Guangzhou
China

Marc Pinto
EXAIL ROBOTICS
Toulon
France

Farman Ali
Department of Computer Science and Engineering
Sungkyunkwan University
Seoul
Republic of Korea

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Remote Sensing* (ISSN 2072-4292), freely accessible at: https://www.mdpi.com/journal/remotesensing/special_issues/U1E907WT8I.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-3841-7 (Hbk)
ISBN 978-3-7258-3842-4 (PDF)
https://doi.org/10.3390/books978-3-7258-3842-4

© 2025 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

About the Editors . . . . . **vii**

Juan Lei, Huigang Wang, Zelin Lei, Jiayuan Li and Shaowei Rong
CNN–Transformer Hybrid Architecture for Underwater Sonar Image Segmentation
Reprinted from: *Remote Sensing* **2025**, *17*, 707, https://doi.org/10.3390/rs17040707 . . . . . **1**

Fei Li, Li Wan, Jiangbin Zheng, Lu Wang and Yue Xi
Contrastive Feature Disentanglement via Physical Priors for Underwater Image Enhancement
Reprinted from: *Remote Sensing* **2025**, *17*, 759, https://doi.org/10.3390/rs17050759 . . . . . **33**

Yuanjie Zhang, Ting Gao, Hongtu Xie, Haozong Liu, Mengfan Ge, Bin Xu, et al.
Narrowband Radar Micromotion Targets Recognition Strategy Based on Graph Fusion Network Constructed by Cross-Modal Attention Mechanism
Reprinted from: *Remote Sensing* **2025**, *17*, 641, https://doi.org/10.3390/rs17040641 . . . . . **53**

Wenjing Cao, Shiliang Fang, Chuanqi Zhu, Miao Feng, Yifan Zhou and Hongli Cao
Three-Dimensional Non-Uniform Sampled Data Visualization from Multibeam Echosounder Systems for Underwater Imaging and Environmental Monitoring
Reprinted from: *Remote Sensing* **2025**, *17*, 294, https://doi.org/10.3390/rs17020294 . . . . . **75**

Shoudong Wang, Haozhong Wang, Zhaoxiang Bian, Susu Chen, Penghua Song, Bolin Su and Wei Gao
A Robust Direction-of-Arrival (DOA) Estimator for Weak Targets Based on a Dimension-Reduced Matrix Filter with Deep Nulling and Multiple-Measurement-Vector Orthogonal Matching Pursuit
Reprinted from: *Remote Sensing* **2025**, *17*, 477, https://doi.org/10.3390/rs17030477 . . . . . **100**

Tian Rong, Yuhang Wang, Qiguang Zhu, Chenxu Wang, Yanchao Zhang, Jianfeng Li, et al.
Sequential Multimodal Underwater Single-Photon Lidar Adaptive Target Reconstruction Algorithm Based on Spatiotemporal Sequence Fusion
Reprinted from: *Remote Sensing* **2025**, *17*, 295, https://doi.org/10.3390/rs17020295 . . . . . **124**

Zhaohua Su, Jie Zhuo and Chao Sun
Modeling and Data Analysis of Bistatic Bottom Reverberation from a Towed Horizontal Array
Reprinted from: *Remote Sensing* **2025**, *17*, 192, https://doi.org/10.3390/rs17020192 . . . . . **151**

Kun Ye, Lang Zhou, Shaohua Hong, Xuebo Zhang and Haixin Sun
Research on 2-D Direction of Arrival (DOA) Estimation for an L-Shaped Array
Reprinted from: *Remote Sensing* **2024**, *16*, 4787, https://doi.org/10.3390/rs16244787 . . . . . **175**

Jieyi Lu, Long Yang, Yixin Yang and Lu Wang
A Fast Two-Dimensional Direction-of-Arrival Estimator Using Array Manifold Matrix Learning
Reprinted from: *Remote Sensing* **2024**, *16*, 4654, https://doi.org/10.3390/rs16244654 . . . . . **188**

Yuan Cao, Tianjun Zhou and Qunfei Zhang
Gridless DOA Estimation Method for Arbitrary Array Geometries Based on Complex-Valued Deep Neural Networks
Reprinted from: *Remote Sensing* **2024**, *16*, 3752, https://doi.org/10.3390/rs16193752 . . . . . **207**

Zhichao Lü, Yaqian Shi, Liangang Lü, Dongyue Han, Zhengkai Wang and Fei Yu
Dual-Feature Fusion Learning: An Acoustic Signal Recognition Method for Marine Mammals
Reprinted from: *Remote Sensing* **2024**, *16*, 3823, https://doi.org/10.3390/rs16203823 . . . . . **228**

Penghua Song, Haozhong Wang, Bolin Su, Liang Wang and Wei Gao
Source Range Estimation Using Linear Frequency-Difference Matched Field Processing in a Shallow Water Waveguide
Reprinted from: *Remote Sensing* **2024**, *16*, 3529, https://doi.org/10.3390/rs16183529 . . . . . **246**

Sheng Feng, Shuqing Ma, Xiaoqian Zhu and Ming Yan
Artificial Intelligence-Based Underwater Acoustic Target Recognition: A Survey
Reprinted from: *Remote Sensing* **2024**, *16*, 3333, https://doi.org/10.3390/rs16173333 . . . . . **266**

Yuzheng Li, Bo Gao, Zhuo Chen, Yueqi Yu, Zhennan Wang and Dazhi Gao
Target Motion Parameters Estimation by Full-Plane Hyperbola-Warping Transform with a Single Hydrophone
Reprinted from: *Remote Sensing* **2024**, *16*, 3307, https://doi.org/10.3390/rs16173307 . . . . . **299**

Mingqiang Ning, Heping Zhong, Jinsong Tang, Haoran Wu, Jiafeng Zhang, Peng Zhang and Mengbo Ma
A Novel Chirp-Z Transform Algorithm for Multi-Receiver Synthetic Aperture Sonar Based on Range Frequency Division
Reprinted from: *Remote Sensing* **2024**, *16*, 3265, https://doi.org/10.3390/rs16173265 . . . . . **317**

Zhiqiang Lin, Zhouyan He, Chongchong Jin, Ting Luo and Yeyao Chen
Joint Luminance-Saliency Prior and Attention for Underwater Image Quality Assessment
Reprinted from: *Remote Sensing* **2024**, *16*, 3021, https://doi.org/10.3390/rs16163021 . . . . . **340**

About the Editors

Xuebo Zhang

Xuebo Zhang (Senior Member of IEEE) earned a Ph.D. in Underwater Acoustic Engineering in 2014. He is passionate about all aspects of underwater acoustics. He has authored over 90 papers. He previously served as the Co-Chair of the IEEE OES Underwater Acoustic Technical Committee. He is currently an Associate Editor for *CSSP* and *IET RSN*. He is also a Guest Editor for *IEEE J-STARS*. His contributions have earned him recognition in the form of the Science and Technology Progress First Prize and other honors.

Haiyong Xu

Haiyong Xu received a Ph.D. from Ningbo University in 2020. He is currently a Professor at the School of Mathematics and Statistics, Ningbo University. His research interests include underwater image processing and deep learning.

Jingyue Lu

Jingyue Lu received a Ph.D. from Xidian University in 2021. He is currently a Prospective Associate Professor at the School of Computer Science and Technology, Xidian University, Xi'an, China. His research interests include forward-looking SAR imaging.

Lei Zhang

Lei Zhang received a Ph.D. from Xidian University in 2011. He is currently a Professor at Sun Yat-sen University, Guangzhou, China. His research interests include radar imaging (SAR/ISAR) and motion compensation.

Marc Pinto

Marc Pinto received a Ph.D. from the University of Paris in 1991. From 1997 to 2010, he worked with the NATO SACLANT Undersea Research Centre. From 2010 to 2014, he was Chief Sonar Scientist at Marport, France. In 2014, he joined the ECA GROUP as Programme Directo. Since 2019, he has been a Scientific Advisor.

Farman Ali

Dr. Farman Ali is an Associate Professor in Sungkyunkwan University, South Korea. He received a Ph.D. from Inha University in 2018. His research interests include the IoT and remote sensing. From 2021 to 2024, he was named among the "Top 2% Scientists in the World" by Stanford University on a yearly basis. He serves as an Associate Editor for *Scientific Reports*, *CMC*, and *CMSE*.

remote sensing

MDPI

Article

CNN–Transformer Hybrid Architecture for Underwater Sonar Image Segmentation

Juan Lei [1], Huigang Wang [1,2,*], Zelin Lei [3], Jiayuan Li [4] and Shaowei Rong [5]

[1] School of Marine Science and Technology, Northwestern Polytechnical University, West Youyi Road, Beilin District, Xi'an 710072, China; enya@mail.nwpu.edu.cn
[2] Research and Development Institute of Northwestern Polytechnical University in Shenzhen, Nanshan District, Shenzhen 518057, China
[3] School of Artificial Intelligence, Xi'an Jiaotong University, Xianning West Road, Beilin District, Xi'an 710049, China; thedieisnotcast@stu.xjtu.edu.cn
[4] School of Computer Science, Northwestern Polytechnical University, DongXiang Road, Chang'an District, Xi'an 710114, China; lijiayuan1488@mail.nwpu.edu.cn
[5] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Jingming South Road, Chenggong District, Kunming 650500, China; rsw1986@mail.nwpu.edu.cn
* Correspondence: wanghg74@nwpu.edu.cn; Tel.: +86-186-0294-7903

Abstract: The salient object detection (SOD) of forward-looking sonar images plays a crucial role in underwater detection and rescue tasks. However, the existing SOD algorithms find it difficult to effectively extract salient features and spatial structure information from images with scarce semantic information, uneven intensity distribution, and high noise. Convolutional neural networks (CNNs) have strong local feature extraction capabilities, but they are easily constrained by the receptive field and lack the ability to model long-range dependencies. Transformers, with their powerful self-attention mechanism, are capable of modeling the global features of a target, but they tend to lose a significant amount of local detail. Mamba effectively models long-range dependencies in long sequence inputs through a selection mechanism, offering a novel approach to capturing long-range correlations between pixels. However, since the saliency of image pixels does not exhibit sequential dependencies, this somewhat limits Mamba's ability to fully capture global contextual information during the forward pass. Inspired by multimodal feature fusion learning, we propose a hybrid CNN–Transformer–Mamba architecture, termed FLSSNet. FLSSNet is built upon a CNN and Transformer backbone network, integrating four core submodules to address various technical challenges: (1) The asymmetric dual encoder–decoder (ADED) is capable of simultaneously extracting features from different modalities and systematically modeling both local contextual information and global spatial structure. (2) The Transformer feature converter (TFC) module optimizes the multimodal feature fusion process through feature transformation and channel compression. (3) The long-range correlation attention (LRCA) module enhances CNN's ability to model long-range dependencies through the collaborative use of convolutional kernels, selective sequential scanning, and attention mechanisms, while effectively suppressing noise interference. (4) The recursive contour refinement (RCR) model refines edge contour information through a layer-by-layer recursive mechanism, achieving greater precision in boundary details. The experimental results show that FLSSNet exhibits outstanding competitiveness among 25 state-of-the-art SOD methods, achieving **MAE** and E_ξ values of 0.04 and 0.973, respectively.

Keywords: salient object detection; FLS images; CNN; transformer; Mamba; hybrid architecture

Academic Editors: Xuebo Zhang, Lei Zhang, Jingyue Lu, Haiyong Xu, Marc Pinto and Farman Ali

Received: 31 December 2024
Revised: 7 February 2025
Accepted: 13 February 2025
Published: 19 February 2025

Citation: Lei, J.; Wang, H.; Lei, Z.; Li, J.; Rong, S. CNN–Transformer Hybrid Architecture for Underwater Sonar Image Segmentation. *Remote Sens.* **2025**, *17*, 707. https://doi.org/10.3390/rs17040707

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Underwater SOD aims to automatically identify and locate the most visually prominent and attractive objects in underwater images. It has widespread applications in underwater resource exploration [1], underwater rescue [2], and underwater hull inspection fields [3]. Unlike traditional underwater object detection [4], underwater SOD focuses on identifying objects that are visually distinct from their surrounding environment, rather than predefined specific categories. Current underwater SOD technologies typically rely on forward-looking sonar (FLS) sensors, which generate real-time image data through the reflection of sound waves. However, despite the potential of existing SOD methods for processing FLS images, inherent noise interference and intensity non-uniformity in FLS images pose significant challenges to prediction accuracy, which somewhat hinders the further development and optimization of SOD technology for FLS images.

The FLS sensor is a device that uses sound waves to detect underwater environments, usually installed in front of ships or submersibles. FLS has a narrow imaging angle and fast signal attenuation speed. In addition, SOD in FLS images faces the following challenges: (1) Resolution limitation: due to the relatively long wavelength of sound waves, the resulting images have low resolution, leading to significant loss of salient feature information. This requires the model to have the capability to extract target information as effectively as possible from limited semantic information in order to accurately identify and locate underwater salient objects. (2) Intensity non-uniformity: as sound waves propagate, their energy attenuates due to absorption, reflection, and scattering by different media, resulting in uneven intensity across the image. As shown in the purple area of Figure 1a, even within the same salient target or surroundings, the pixel intensity in different areas will exhibit varying brightness [5]. (3) Multipath effect: when sound waves propagate underwater, they may reach the receiver through multiple paths. The signals from these different paths can overlap, potentially causing ghosting or aliasing effects in the image [6]. As illustrated in the blue area of Figure 1b, multipath noise results in ghost images of non-salient objects in the FLS image. (4) Shadow noise interference: during sound wave transmission, obstacles can easily block the waves, leading to target shadows in FLS images, as shown in the yellow area of Figure 1c. Due to the unique characteristics and complex scene patterns of FLS images, the existing SOD methods designed for optical images often struggle to generate satisfactory saliency maps when directly applied. Consequently, there is an urgent need to develop SOD techniques for FLS images, addressing their inherent challenges and complexities. Traditional SOD in FLS images is often coupled with segmentation techniques. For instance, the Markov algorithm [7] can be used to segment sonar images into shadow and seabed reverberation regions, enhancing the segmentation performance by estimating noise distribution parameters and Markov prior parameters. In addition, techniques such as the Bayesian algorithm [8] and wavelet transform [9], which combine local spatial information and statistical information [6], have also been widely applied to optimize the SOD and segmentation performance of FLS images. However, these methods have limitations in accuracy, efficiency, and generalization ability and rely on manually extracted prior features such as color, texture, etc. These features are typically transmitted to classifiers such as support vector machines (SVM) and AdaBoost for recognition. However, the complex underwater environment often results in poor accuracy and generalization ability of manual features. Compared with the traditional methods, deep learning methods have demonstrated excellent feature extraction and semantic learning capabilities in SOD of FLS images, significantly promoting the development of this field. The traditional SOD method based on the convolutional neural network (CNN) mainly explores the contextual semantic relationships and shared features between image pixels through convolution operations in order to identify salient objects. However, local sliding windows are mainly

used for systematic modeling of the local correlation between adjacent pixels in the image, gradually aggregating high-level semantic features through multi-layer convolution, and are good at capturing local features of salient objects [10]. In addition, CNN also has the drawback of insufficient receptive field, especially at deeper levels, where its actual receptive field is significantly smaller than the theoretical receptive field [11]. In recent years, several studies have developed various methods targeting the long-range correlations of contextual semantic information. For example, Jia et al. [12] significantly improved the expression of semantic information by introducing interactions between local multi-scale features and cross-layer features. Liu et al. [13] integrated CNNs with capsule networks (CapsNets) to extract discriminative cues and ensure consistency between parts and the whole. Xia et al. [14] employed dilated convolutions and feature reconstruction techniques to effectively extract and reconstruct multi-scale contextual semantic information, enhancing the diversity and richness of feature representation. Chen et al. [15] employed stepwise aggregation of features at different levels to capture long-range contextual information. Although these methods improve the ability to extract global contextual features, they often confine attention learning to high-level features of small-scale inputs, thereby limiting performance in complex underwater environments. Recently, vision transformers (ViTs) [16], with their superior self-attention mechanisms, have been widely applied to dense pixel prediction tasks, including image classification [17], object detection [18], image segmentation [19], and SOD [20]. Compared to traditional CNN-based SOD methods, Transformers, through stacking self-attention layers, can provide global context information, thereby constructing global semantic representations to enhance SOD. Yuan et al. [20] explored the uniqueness and interdependencies of features at different layers of Transformers, enhancing these relationships through layer-wise feature interactions. To address the issue of object structural information loss due to prior assumptions of visual markers, Chen et al. [21] reconstructed patches and superpixels as visual markers to better preserve structural information. Liu et al. [22] used T2T-ViT [23] as the base network to construct a unified SOD architecture for pure Transformers targeting RGB and RGB-D. Although these methods effectively model global visual structures, they exhibit limitations in local representation.

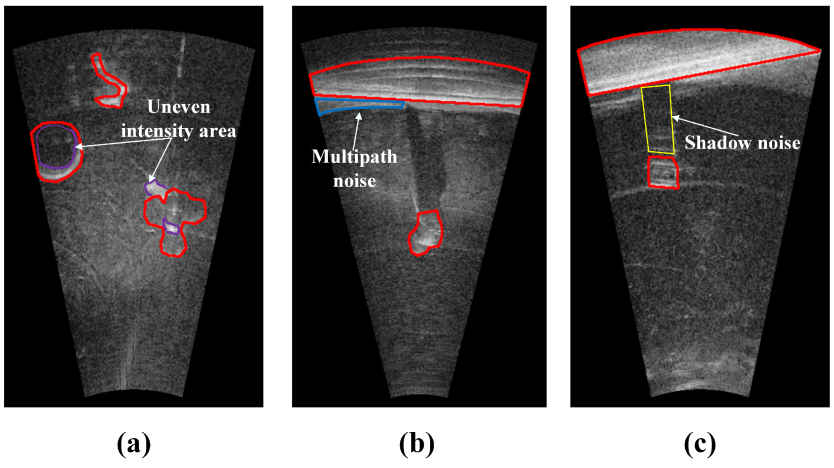


Figure 1. Examples of FLS images containing various noise sources. The red region indicates the ground truth of the salient target, the purple region represents areas with intensity inconsistency, the blue region indicates multipath noise, and the yellow region represents shadow noise.

As mentioned above, the existing pure CNN and pure Transformer SOD methods have gone from one extreme to another, that is, from lacking global modeling ability to insufficient local representation ability [24]. To address this challenge, the recent research has attempted to combine the strengths of CNNs and Transformers to achieve complemen-

tary performance. Qiu et al. [24] embedded the Transformer architecture PVT [25] into an asymmetric U-Net architecture for SOD. Yuan et al. [26] achieved deep optimization of information representation by iteratively integrating and enhancing CNN local information and Transformer global features through a multi-stage fusion mechanism. Ge et al. [10] employed a parallel interaction strategy to effectively integrate multimodal features, enabling the model to accurately capture salient objects at various scales. Zeng et al. [27] proposed a serial integration framework for Transformer and CNN and introduced a mutual interaction module to facilitate efficient interaction of different modal information while enhancing the regularization of relevant features through edge supervision mechanism. These methods can integrate the advantages of different modes; however, by indiscriminately fusing different modal features, the potential of the model to explore specific modal advantages may be limited. In addition, the loss of long-range dependencies during feature fusion is also a key issue that urgently needs to be addressed. In recent years, structured state space models (SSMs), inspired by classical Kalman filtering models, have integrated multiple advantages of sequence models at a theoretical level due to their outstanding sequence transformation capabilities. This includes features from continuous time models (CTMs) and recurrent neural networks (RNNs). This framework not only inherits the core concepts of traditional sequence modeling paradigms but can also deeply embed them into neural network structures, greatly expanding the application potential of deep learning in the field of complex sequence data processing. A recent work by Mamba [28] integrated time-varying parameters into a state space model (SSM) and proposed a hardware aware algorithm that enhances the ability to model pixel long-range correlations. Although some research has begun to explore the potential of Mamba in visual applications, its application potential in complex underwater imaging environments and in FLS images with significant category differences remains to be further investigated and explored.

Based on the above discussion, we believe that the complementary patterns between CNN, Transformer, and Mamba have great potential in improving model performance. The core task of complementary patterns is to suppress the differences between heterogeneous cross modal cues and enhance the high-order unity of features. Therefore, introducing different levels of feature extraction modules into the SOD model can significantly improve detection performance. This article proposes a multimodal parallel hybrid network called FLSSNet, which mainly consists of the following modules: asymmetric dual encoder–decoder (ADED) module, transformer feature converter (TFC) module, long-range correlation attention (LRCA) module, and recursive contour refinement (RCR) module. Specifically, the ADED module integrates the contextual information of CNN with the spatial structure information of Transformer and performs saliency mapping. The TFC module captures and shares cross modal information by combining the dual self attention mechanism within branches and the local feature extraction of convolutional neural networks (CNN), effectively reducing the problem of feature misalignment between different modal features. In addition, in order to preserve the global detail features of the Transformer, TFC also introduces a residual channel attention mechanism to improve the global consistency and detail fidelity of the features. Given the limitations of convolution in capturing long-range dependencies, the LRCA module combines multi-directional stripe convolution and the Mamba module to model the local and global long-range correlations of pixels from a dual perspective of convolution and sequence. The RCR module adopts a layer by layer recursive mechanism to gradually refine the edge contour of the target and further optimize the structural details. Finally, in order to achieve optimal performance of the model, we adopted a hybrid loss function. The contributions are as follows:

(1) We propose a SOD network for FLS images, called FLSSNet. To our knowledge, this is the first SOD network for FLS images based on a CNN–Transformer–Mamba hybrid

architecture, which combines CNN's fine local feature extraction ability, Transformer's global spatial structure modeling ability, and Mamba's advantage in pixel level remote dependency modeling.

(2) The design of the TFC module aims to accurately reconcile the misalignment problem between CNN and Transformer feature expressions. In the subsequent feature fusion stage, the LRCA module further enhances CNN's ability to capture long-range dependencies, ensuring deep fusion and consistency of cross modal information.

(3) To further avoid insufficient multimodal feature fusion and improve the precision of target boundaries, the RCR module introduces a multi-layer recursive mechanism to optimize feature interaction and boundary details layer by layer, ensuring accurate description of the target and deep integration of multimodal information.

(4) Compared with the existing pure CNN, pure Transformer, and their fusion hybrid methods, FLSSNet demonstrates its highly competitive detection performance.

2. Related Works

In the field of SOD, researchers have proposed a large number of SOD methods, which can be roughly divided into traditional methods and deep-learning-based methods. The traditional SOD method relies on manually designed features such as texture [29], color [30], and contrast [31] or employs a bottom-up or top-down approach for saliency mapping. In addition, some methods often rely on heuristic clues [32]; please refer to [33,34] for more details. However, the computational accuracy of these methods is often limited, making it difficult to cope with complex scenarios, especially in underwater environments, and exhibiting poor generalization ability. This article mainly focuses on reviewing SOD methods based on deep learning, covering pure CNN, pure Transformer, and CNN–Transformer hybrid models. In addition, this article also reviews the latest applications and developments of visual state space models in this field.

2.1. SOD Methods Based on CNN

In recent years, significant progress has been made in SOD methods based on convolutional neural networks (CNN). These methods typically rely on pre trained networks (such as ResNet [35], VGG [36], etc.) as backbone networks or adopt different architectural designs, such as U-shaped encoding–decoding structures [37] and bottom-up [38] and top-down structures [39]. Although these methods have made significant contributions to the field of SOD, the inherent local receptive field of CNN limits its ability to capture long-range dependencies between target pixels, which are crucial for capturing the global structure and spatial layout of the target. Therefore, some studies aim to break through the limitations of CNN and attempt to enhance the network's ability to perceive the overall contour of the target by enhancing global dependency modeling. Some research methods focus on the implementation of multi-scale feature fusion. For example, Zhu et al. [40] used a feature aggregation module to capture the complementarity and differences between features at different levels and integrated a feature refinement module to effectively suppress the noise interference generated during the feature aggregation process. Xia et al. [14] introduced a diamond-shaped hierarchical structure that skillfully combines local convolutions and dilated convolutions to integrate features at different receptive field scales, enabling efficient fusion and representation of multi-level features. Zhu et al. [41] introduced a two-stage residual nested aspp module (Resaspp2) in a dual encoding–decoding network, which maximizes the receptive field while reducing the number of convolutional layers. In order to learn more feature clues, some research methods have proposed multi-scale attention mechanisms. Li et al. [42] introduced a purification mechanism that enhances the focus of the promotion sub network on salient regions, while using the correction sub

network to correct errors in salient regions. In addition, to address the critical issue of edge information loss caused by ignoring global information, Tu et al. [43] enhanced the effective feature representation of non-local features by introducing edge-guidance mechanisms and non-local operations. In the video SOD task, Kousik et al. [44] combined the excellent feature extraction ability of CNN with the advantages of the convolutional recurrent neural network (CRNN) in sequence data processing, effectively overcoming the information loss problem that traditional CNNs may cause in complex scenes due to receptive field limitations. Go et al. [45] explored the strategy of introducing spatial bias in non-local neural networks by stacking parallel convolutions on feature maps to adjust the spatial structure of convolutional features. Although the above methods have improved the long-range correlation of target pixels to some extent, their essence is still based on local operations and extensions, making it difficult to adaptively adjust feature relationships at different scales and ranges. In addition, some research methods aggregate different hierarchical features layer by layer through multi-level refinement, aiming to gradually fuse local features into a more global representation. For example, Wu et al. [46] cascaded multiple cross refinement networks, systematically integrating the saliency of cross layer features and edge features, thereby achieving effective collaboration and reinforcement between the two. Zhang et al. [47] iteratively optimized the initial rough saliency mapping image through cross level hierarchical loop connections, while introducing pyramid pooling modules and channel attention modules to capture global contextual information. Wan et al. [48] used context-driven focusing modules and semantic guided localization modules to accurately locate and enhance feature expressions at different levels. Subsequently, clear boundary clues and final fine saliency maps are further generated through the edge refinement module. These methods effectively validate the excellence of multi-stage refinement strategies. Inspired by this, we propose the recursive contour refinement (RCR) model, which aims to further explore the hidden edge information within features to achieve more refined SOD.

2.2. SOD Methods Based on Transformer

Transformer was first introduced into natural language processing by Vaswani et al. Its multi-head self-attention mechanism provides powerful support for capturing target spatial structural information, enabling the architecture to quickly expand into the field of computer vision. The existing Transformer-based SOD methods can be divided into pure Transformer structure and hybrid Transformer structure. In a pure Transformer structure, the input image is first divided into several fixed-sized blocks, and the correlations between these image blocks are then modeled by stacking multiple self-attention layers. Vit first applied Transformer to the field of image classification in pixel level dense prediction tasks, laying the foundation for subsequent improvements [49] in ViT and the application of Transformer in computer vision. VST [22] proposes a unified SOD model for RGB and RGB-D based on the pure Transformer architecture T2T-Vit [23]. Gao et al. [50] proposed the ASTM module and embedded it into the Transformer backbone network, effectively reducing the redundancy of invalid tokens. The proposal of Swin-Transformer [51] marks the further maturity of Transformer architecture in the field of computer vision. On the basis of inheriting the advantages of Swin-Transformer structure, M3Net [20] focuses on the mutual correlation between multi-level features and achieves efficient fusion and interaction of cross level features by introducing multi-layer interaction modules as decoders. These pure Transformer-based SOD methods can quickly establish global dependencies between pixels but often lack the ability to capture local details. Compared to Transformer, CNN has significant advantages in local feature extraction and spatial invariance. Therefore, some studies attempt to integrate Transformer and CNN architectures to merge the advantages between the two, namely, hybrid Transformer architecture. For example, ASNet [52] em-

ploys a dual-modal encoder to adaptively integrate multi-level global structural features and local fine-grained cues. Additionally, the network introduces an adaptive semantic matching module (ASMM), which enables precise matching and alignment of feature representations across different levels. TranSalNet [53] simulates human visual characteristics by integrating Transformers into the CNN architecture, overcoming the inherent biases present in traditional CNN architectures. Wang et al. [54] employed a hybrid cascaded CNN and Transformer architecture as the encoder for feature extraction, fully leveraging the strengths of both in capturing local details and modeling global dependencies.

2.3. Visual State Space Models

The state representation of state space models (SSMs) typically requires a significant amount of computation and memory resources. Although structured SSMs alleviate this limitation to some extent by parameterizing the state matrix, their constant sequence structure still has limitations in modeling pixel long-range correlations. Mamba [28] achieved sequential scanning of SSM along specific directions by integrating selection mechanisms and further developed a hardware aware algorithm, significantly accelerating the training and inference process of the model. With the success of natural language processing methods in cross modal learning in the field of computer vision, many studies have attempted to explore the application and improvement of Mamba. VMamba [28] proposed a 2D scanning (SS2D) module based on visual state space (VSS) blocks that unfolds images into sequences along both horizontal and vertical directions and performs selective scanning. MambaVision [55] proposed a Mamba–Transformer hybrid architecture network that models remote spatial dependencies through multiple self-attention blocks. Medmamba [56] constructed the SS-Conv-SSM module, which combines local features extracted by convolution with remote dependencies captured by SMM to achieve efficient classification of medical images. RS-Mamba [57] focuses on extracting spatial features and uses an omnidirectional state space block (OSSM) to scan the input image in multiple directions, accurately capturing long-range correlations between pixels. Inspired by the concept of multimodal feature fusion, we have designed an SOD model for FLS images that aims to combine the complementary advantages of CNN, Transformer, and Mamba. Firstly, to address the issue of feature misalignment, we proposed the Transformer feature converter module and integrated it into the CNN architecture. Additionally, we introduce the long-range correlation attention (LRCA) module, which boosts the CNN's capability to capture long-range dependencies both locally and globally by utilizing directional strip convolutions and comprehensive all-direction scanning with Mamba. Finally, to address potential contour information loss resulting from insufficient multimodal feature fusion, we introduce the recursive contour refinement model (RCRM). This module progressively optimizes and refines the information representation during the feature fusion process, further uncovering and enhancing the latent value of each feature.

3. Methodology

We first describe the overall framework of FLSSNet in Section 3.1. Subsequently, we introduce the asymmetric dual encoder–decoder, transformer feature converter module, long-range correlation attention module, recursive contour refinement model module, and the hybrid loss function in Sections 3.2, 3.3, 3.4, 3.5, and 3.6, respectively.

3.1. Overall Architecture

Extracting more contextual semantic and spatial structural information has always been an important research challenge in SOD of FLS images. The contextual information profoundly affects the accuracy of SOD, while spatial information determines the accuracy

of object localization and boundary segmentation. Convolutional neural networks (CNNs), with their powerful feature extraction capabilities, can capture rich contextual semantic information and model local fine-grained features. Transformer can achieve modeling of global spatial structural features by stacking a large number of self-attention layers. In addition, Mamba achieves a balance between remote dependency modeling and computational complexity by linearly scaling sequence length through sequence transformation and selective scanning [58]. In order to effectively explore the application of a hybrid network based on CNN, Transformer, and Mamba in FLS image saliency dense prediction, we propose FLSSNet. This network aims to enhance pixel long-range correlation modeling and solve the precise SOD problem of FLS images by integrating the performance and synergy between different modes, as in [59]. As shown in Figure 2, the FLS image is input into the asymmetric multimodal salient object detection (AMSOD) model for multimodal feature extraction. Subsequently, the recursive contour refinement (RCR) model further refines the local and contour regions of the target. The AMSOD architecture consists of three modules: the asymmetric dual encoder–decoder (ADED) module, transformer feature converter (TCF) module, long-range correlation attention (LRCA) module, and recursive contour refinement (RCR) model. The ADED module uses a hybrid architecture of CNN and Transformer to extract both local and global contextual semantic information. Unlike most SOD methods that rely on simple feature concatenation, we introduce the TCF module to delve into the complementarity and consistency of different modal features. Additionally, because features at different levels contain rich details and semantic information, we have designed the LRCA Module to further balance the local and global pixel associations across different feature levels. Considering that a single U-Net decoder may struggle to fully express the fused feature information and effectively leverage the advantages of different modalities, we use the RCR Model to further enhance the network's precision and extract edge information from the modal features.

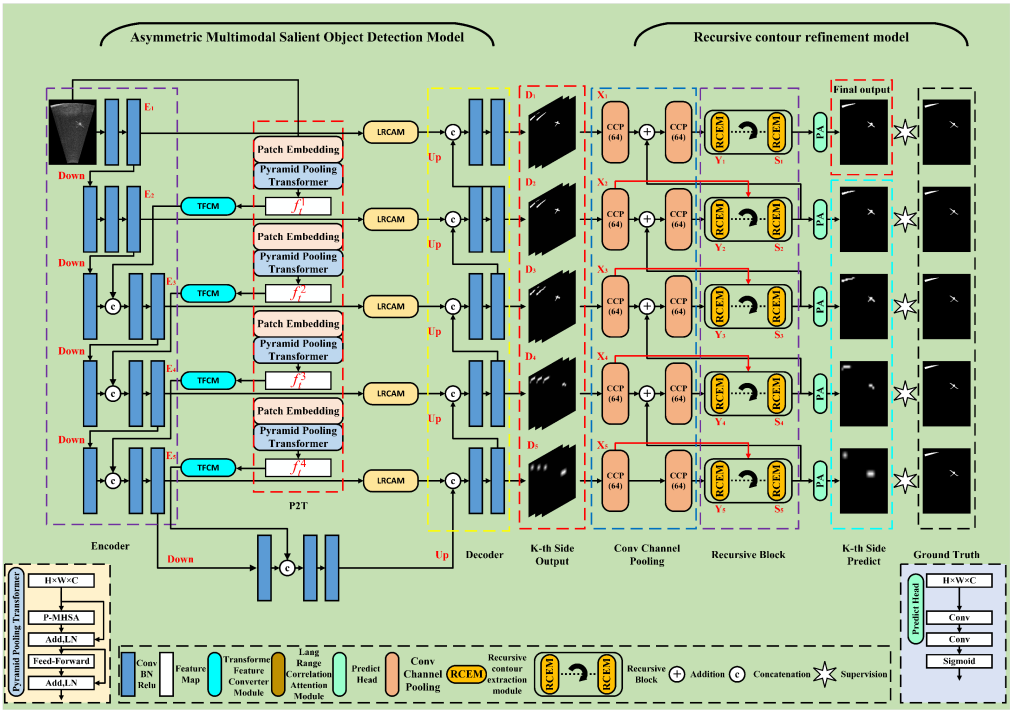


Figure 2. The overall structure of the proposed FLSSNet. The method employs a two-stage strategy: the first stage utilizes an asymmetric dual encoder–decoder structure for saliency feature extraction,

while the second stage further refines the feature maps using the recursive refine module. In the first stage, the image is input into two encoders to obtain feature information from different modalities. Simultaneously, the Transformer feature converter module is used to transform and compress information from these modalities. Next, the long-range correlation attention module integrates multi-level features and reduces feature redundancy. Finally, the recursive refine module is employed to further enhance the precision of feature prediction.

3.2. Asymmetric Dual Encoder–Decoder

Firstly, we need to briefly explain the concept of asymmetric dual encoder–decoder architecture: (1) FLSSNet uses different modes of paths (CNN and Transformer) as encoders. (2) The number of stages for the two paths is different (CNN uses six stages, Transformer uses four stages). (3) Different encoder paths are responsible for different tasks (contextual semantic information modeling and spatial structural feature modeling). (4) The decoder path adopts the CNN architecture uniformly.

As shown in Figure 2, the core of the asymmetric dual encoder–decoder consists of U-Net and a pre trained P2T [60] network. P2T is a Transformer network that integrates pyramid pooling, effectively alleviating the issue of excessively high quadratic complexity caused by long sequence lengths through multi-level pooling operations. The P2T network consists of four stages, each adopting a unified architectural design, primarily composed of patch embedding blocks and pyramid pooling transformer blocks. Specifically, given an FLS image $I \in \mathbb{R}^{H \times W \times 3}$, the image is first divided into $\frac{H}{4} \times \frac{W}{4}$ blocks, with each block flattened into 48 pixels. Subsequently, the feature dimension is gradually expanded from 48 to 64, 128, 320, and 512 through the linear projection layer and learnable positional encoding block in the patch embedding block. In the pyramid pooling transformer block, the feature maps are progressively scaled by a factor of 2 at each stage, ultimately resulting in a set of feature maps $\{f_t^1, f_t^2, f_t^3, f_t^4\}$, with sizes corresponding to $1/4, 1/8, 1/16,$ and $1/32$ of the original input image, respectively. Benefiting from the self-attention mechanism, P2T explores the spatial relationships of the target through multiple self-attention layers. Although this approach effectively promotes SOD in FLS images through global context representation, it sacrifices the modeling of local correlations to some extent. Therefore, we introduce the classic U-Net network to supplement it.

The encoder of U-Net stacks five convolutional layers, and their output feature images are represented as $\{E_1, E_2, E_3, E_4, E_5\}$. In stages 3 to 5, the input features not only come from the output of the previous stage of the U-Net encoder but also integrate the output features of each stage of P2T. It is worth noting that directly concatenating the output features of the two networks along the channel dimension may lead to feature misalignment and redundancy. To address this, we use the TFC module (will be described in Section 3.3) to perform feature transformation and channel compression on $\{f_t^1, f_t^2, f_t^3, f_t^4\}$. As a result, FLSSNet achieves dual modeling of both local and global features. Additionally, FLSSNet incorporates the LRCA Module (will be described in detail in Section 3.3), which aims to enhance both local and global long-range correlations between salient object pixels and alleviate noise accumulation issues.

3.3. Transformer Feature Converter Module

Clearly, the features of different modes (CNN and Transformer) have different properties, and further processing is needed to address the issue of feature misalignment [10]. Before introducing the TCF modules, we need to explain why feature transformation is only performed on $\{f_t^1, f_t^2, f_t^3, f_t^4\}$ instead of $\{E_1, E_2, E_3, E_4, E_5\}$. The reasons are as follows: (1) Determined by network architecture: to progressively restore the image size while minimizing the loss of local details, we chose the classic U-Net architecture for SOD. The U-Net decoder is more suited to handling convolutional features compared to other

architectures. (2) Complexity of feature fusion: Transformer typically processes features in a sequential manner, making multi-scale feature fusion more challenging. In contrast, CNN offers greater flexibility and advantages, allowing feature fusion through addition, multiplication, or concatenation. (3) Quadratic computational complexity: Transformer requires calculating correlations between sequence positions and performs extensive matrix operations within the self-attention layers. In comparison, CNN's convolutional operations are local, avoiding excessive computational complexity. We aim to strike a balance between performance and computational complexity, leveraging the strengths of features from different modalities.

As shown in Figure 3, the TFC module adopts a dual-branch strategy, primarily consisting of the residual channel attention module (RCAM) and the multi-scale dual self-attention mechanism module (MDSAM). For a given input feature map $F \in \mathbb{R}^{H \times W \times C}$, the output feature maps F_a and F_s are obtained through the RCAM and MDSAM, respectively. The final output feature map F_l is then produced by linearly adding these two feature maps. RCAM adopts the channel attention mechanism and weight layer to selectively integrate key salient information in the global context, enhancing the expressive power of feature maps. MDSAM adaptively integrates the common saliency features of CNN and Transformer through multi-scale convolution. By using a dual multi-head self-attention module (MHSA), the consistency between features of different scales is captured, thereby achieving deep fusion and optimization of feature information.

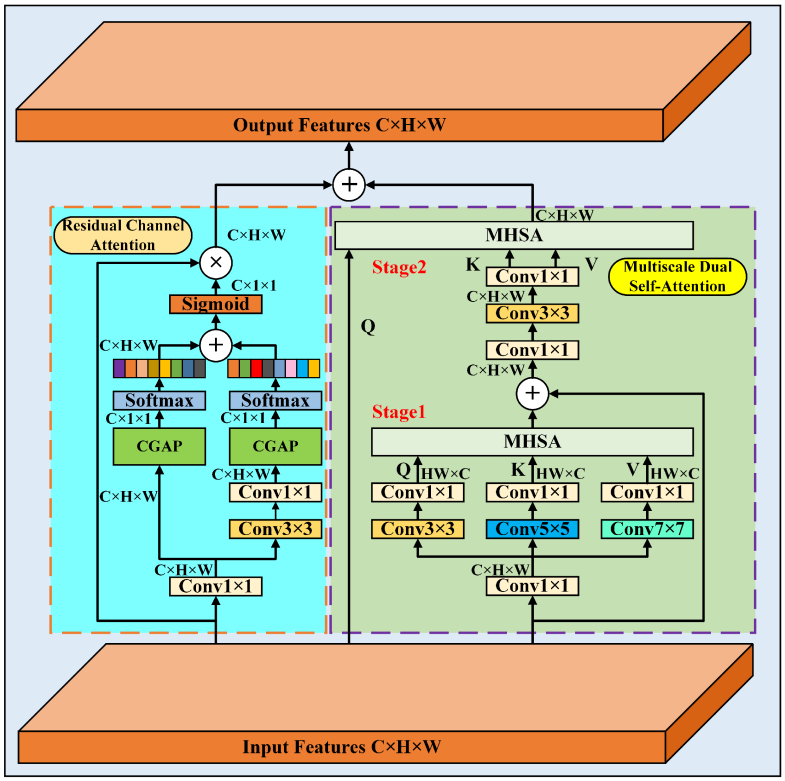


Figure 3. The overall structure of the proposed Transformer feature converter (TFC) module. The TFC module mainly consists of the residual channel attention module (RCAM) and the multi-scale dual self-attention mechanism module. MHSA stands for multi-head attention mechanism.

(1): Residual channel attention module: Firstly, perform a 1 × 1 convolution operation on the input feature F to achieve cross channel feature recombination. This means that each layer in the Transformer features has undergone independent interaction between channels. Next, the reassembled features are processed by separate branches to extract different channel attention weights. Each branch utilizes global adaptive average pooling

to accurately capture channel attention weights, while an additional convolution operation is introduced in one branch to further enhance the semantic representation depth of the Transformer features. Finally, the mixed channel attention weight A_c was generated through linear addition and sigmoid function processing. The entire process can be expressed as

$$A_c^1 = \text{Softmax}\{G_c^a(C_1(F))\}. \tag{1}$$

$$A_c^2 = \text{Softmax}\{G_c^a(C_{1,3,1}(F))\}. \tag{2}$$

$$A_c = \text{Sigmoid}(A_c^1 + A_c^2). \tag{3}$$

where $\text{Softmax}(*)$ and $\text{Sigmoid}(*)$, respectively, represent the Softmax and Sigmoid functions. C_1 represents a convolution operation with a kernel size of 1×1. $C_{1,3,1}$ represents the sequential execution of convolution operations with kernel sizes of 1×1, 3×3, and 1×1. G_c^a represents performing global adaptive average pooling operations on the channel dimension. Finally, by performing element-wise multiplication between A_c and F, the output feature map F_a with hybrid channel attention weights is obtained.

(2): Multi-scale dual MHSA module: The module employs a dual-stage multi-head attention mechanism(MHSA) and a multi-scale approach. Similarly, F first undergoes a 1×1 convolution operation, followed by three parallel multi-scale convolution layers to extract the common salient features from both CNN and Transformer. The convolution kernel sizes for the multi-scale convolution layers are R_i $(i \in \{1, 2, 3\})$ and are set to 3, 5, and 7, respectively. Subsequently, cross-channel feature reorganization is performed, and the multi-scale output features F_i are used as input to the first stage of the multi-head self-attention (MHSA) module. For simplicity, the description of the flattening operation has been omitted.

$$F_1^i = C_1(C_{R_i}(C_1(F))). \tag{4}$$

$$A_{s1}\left(Q_{F_1^i}^{s1}, K_{F_1^i}^{s1}, V_{F_1^i}^{s1}\right) = \text{Softmax}\left(\frac{Q_{F_1^i}^{s1}\left(K_{F_1^i}^{s1}\right)^T}{\sqrt{d_k}}\right) V_{F_1^i}^{s1}. \tag{5}$$

where $Q_{F_1^i}^{s1}$,$K_{F_1^i}^{s1}$, and $V_{F_1^i}^{s1}$ represent the query, key, and value generated from F_1^i in the first stage, respectively. d_k is the channel dimension K_{s1}, which can be used as a learnable scaling parameter and is approximately considered a normalization process. Compared to traditional MHSA mechanisms, we employ a multi-scale approach (including convolutional local attention and cross-attention) to model the co-salient features of CNN and Transformer. Subsequently, feature representation is enhanced through residual connections, resulting in the feature $F_{s1} = \left\{A_{s1}\left(Q_{F_1^i}^{s1}, K_{F_1^i}^{s1}, V_{F_1^i}^{s1}\right) + F\right\}$, which is then used as the input for the subsequent stages. The process is represented as follows:

$$F_2 = C_1(C_3(C_1(F_{s1}))). \tag{6}$$

$$F_s = A_{s2}\left(Q_F^{s2}, K_{F_2}^{s2}, V_{F_2}^{s2}\right) = \text{Softmax}\left(\frac{Q_F^{s2}\left(K_{F_2}^{s2}\right)^T}{\sqrt{d_k}}\right) V_{F_2}^{s2}. \tag{7}$$

where Q_F^{s2}, $K_{F_2}^{s2}$, and $V_{F_2}^{s2}$ represent the query, key, and value generated from F and F_2, respectively. Subsequently, F_a and F_s are combined through a linear summation to obtain the final output feature map $F_l = F_a + F_s$.

3.4. Long-Range Correlation Attention Module

Low-level features contain rich structural information, while deep features carry abundant semantic information, both of which play a crucial role in SOD [20,26]. Traditional methods typically fuse multi-level features through skip connections, but they often overlook several key issues: (1) how to eliminate noise interference within multimodal features; (2) how to establish long-range dependencies between the local pixels. To address these issues, we propose an LRCA module. As shown in Figure 4, this module consists of three main components: a multi-directional strip convolution module, an attention module, and an omnidirectional selective scan module (OSSM) based on Mamba. These components collaboratively enhance long-range correlations between pixels and mitigate noise interference through the directionality of convolutional kernels, selective sequence scanning, and attention mechanisms.

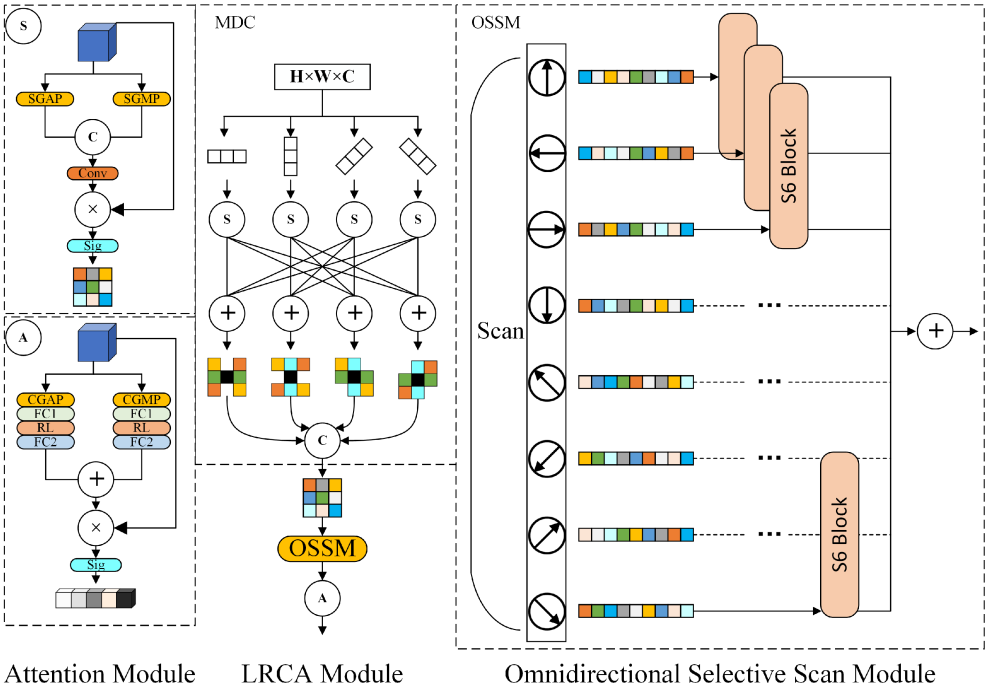


Figure 4. The overall structure of the proposed long-range correlation attention (LRCA) module. The LRCA module primarily consists of a multi-directional convolution module (MDC) and omnidirectional selective scan module (OSSM), an attention module.

Inspired by Zhang et al. [61,62], we integrate features extracted from convolutions in different directions to capture the long-range correlations between local pixels. First, the given input features are divided into four groups along the channel dimension, denoted as $x_i \in \mathbb{R}^{H\times W\times\left(\frac{C}{4}\right)}$ $(i = 1,2,3,4)$. Each feature group x_i undergoes convolution in different directions to generate the saliency map x_d^i. Let $\omega \in \mathbb{R}^{2m+1}$ represent a 1D transposed convolution kernel of size $2m+1$. Let $l = (l_h, l_w)$ represent the directional vector of the 1D convolution kernel, where the values $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$ correspond to horizontal, vertical, positive diagonal, and negative diagonal transposed convolutions, respectively. The specific representation of directional convolution is as follows:

$$y_l[i,j] = \left(x * \omega^T\right)_l = \sum_{t=-r}^{r} x[i + l_h t, j + l_w t] \cdot \omega[r - t]. \tag{8}$$

where $y_l[i,j] \in \mathbb{R}^{H\times W}$ represents the result of the directional convolution between the input feature $x \in \mathbb{R}^{H\times W\times\frac{C}{4}}$ and ω. By setting $r = 4$, we ensure that each 1D convolution kernel

has the same nine parameters as the 3×3 convolution kernel. Since the 3×3 transposed convolution layer is replaced by four concatenated sets of 1D convolution kernels in different directions, the total number of network parameters and computational complexity remain unchanged. For more details, please refer to paper [63]. Next, the directionally convolved features x_d^i are processed through the following spatial attention module to capture the importance of spatial position pixels:

$$\mathrm{SA}_i = \mathrm{Sigmoid}\left\{ C_1\left(\mathrm{SA}(x_d^i)\right) \times x_d^i \right\}. \tag{9}$$

where $\mathrm{SA}(*)$ represents the spatial attention operation, defined as

$$\mathrm{SA}(x) = \mathrm{Concat}\{G_s^a(x), G_s^m(x)\}. \tag{10}$$

where $\mathrm{Concat}\{x, y\}$ represents the operation of concatenating features x and y along the channel dimension. $G_s^a(x)$ and $G_s^m(x)$ represent global average pooling and max pooling operations along the spatial dimensions, respectively. Secondly, SA_i is fused along different directions, and the original feature channel count is restored, resulting in the final spatial attention feature map I_s:

$$I_s = \mathrm{Concat}\left\{ \sum_{i\in\{1,3,4\}} \mathrm{SA}_i, \sum_{i\in\{2,3,4\}} \mathrm{SA}_i, \sum_{i\in\{1,2,3\}} \mathrm{SA}_i, \sum_{i\in\{1,2,4\}} \mathrm{SA}_i \right\}. \tag{11}$$

Although directional convolution combines the powerful feature extraction ability of convolution with the ability to model long-range dependencies in specific directions, its effectiveness is still limited by the size of the convolution kernel, which can only capture specific long-range dependencies within a local range of pixels. Inspired by the application of Mamba in computer vision, we attempt to use OSSM to model global long-range dependencies. Thirdly, we model the global long-range dependency of spatial feature I_s and generate global dependency aware feature I_o. OSSM can scan input images from different directions, and its core component is SSM. SSM typically uses algorithms such as Kalman filtering or particle filtering for state estimation, consisting of the state equation X_t and observation equation Y_t.

$$X_t = F_t X_{t-1} + G_t U_t + \omega_t. \tag{12}$$

$$Y_t = H_t X_t + v_t. \tag{13}$$

where X_t represents the hidden state at time t, F_t is the state transition matrix, G_t is the control matrix, U_t is the control input, and ω_t denotes the process noise (usually assumed to be mean Gaussian white noise). Y_t is the observation at time t, H_t is the observation matrix, and v_t represents the observation noise. For more details, please refer to papers. Fourthly, we process I_o through the following channel attention module to capture channel importance, obtaining the channel-weighted feature vector I_c.

$$I_c = \mathrm{Sigmoid}(\mathrm{CA}(I_o) \cdot I_o). \tag{14}$$

where $\mathrm{CA}(*)$ represents the channel attention operation, defined as

$$\mathrm{CA}(x) = F_c^2\left[R_u\left\{F_c^1(G_c^a(x))\right\}\right] + F_c^2\left[R_u F_c^1(G_c^m(x))\right]. \tag{15}$$

where F_c^1 and F_c^2 represent fully connected operations of different dimensions. G_c^a and G_c^m denote global average pooling and maximum pooling along the channel dimension, respectively. R_u represents the ReLU activation function.

3.5. Recursive Contour Refinement Model

In the decoding phase of U-Net, contour information is often utilized for saliency fusion and final loss supervision. To further refine the network model and extract edge details from multimodal features, inspired by paper [64], we propose a recursive contour refinement module (RCRM). The RCRM uses a layer-by-layer recursive extraction method, progressively enhancing the precision of edge through recursive processing. As shown in Figure 2, the RCRM consists of multiple conv channel pooling (CCP) modules, a recursive block, and a prediction head (PA) module. In RCRM, the CCP module is first used to perform channel conversion on the feature images D_m (m = 1, 2, 3, 4, 5) at different levels of the decoder, uniformly adjusting them to 64 dimensions to simplify subsequent operations. The CCP module consists of three sets of serial convolutions with kernel sizes of 3 × 3. Subsequently, the features are input into a recursive block to extract edge contour information from multimodal features. Inspired by U2-Net, we employ a multi-layer side supervision strategy, where the output feature maps from each layer of the recursive block are processed through the PA module for saliency mapping and supervision. The PA module consists of two 3 × 3 convolutional layers followed by a Sigmoid activation function.

The recursive block is the core of the entire RCRM, and its structure is shown in Figure 5a. The recursive block uses multiple recursive contour extraction module (RCEM) blocks to extract local and contour features from the decoder's hierarchical output features. We represent N = 1, 2, 3, ..., n as the total number of RCEM in a single-layer recursive block and m = 5, 4, 3, 2, 1 as the hierarchical arrangement number of the recursive block. S_m represents the feature output of the m-th layer recursive block. The formula for recursive blocks at different levels is expressed as

$$S_m = R_m^n\Big(\dots\Big(R_m^2\Big(R_m^1(Y_m) + Y_m\Big) + Y_m\Big) + \dots\Big) + Y_m + X_m. \tag{16}$$

where R_m^1, R_m^2, ...,R_m^n represent the different stages of recursive contour extraction module operations at the m-th level. It is worth noting that within the recursive block, we utilize both local and global residuals. However, the global residual is applied only to the final layer of the recursive block, with the input feature map coming from the original input feature map X_m. Although the recursive block consists of multiple RCEMs, these RCEMs share the same weights during the recursive process, only with different subscripts used to represent them in the equations. Next, we will provide a detailed explanation of the design of the RCEM. As shown in Figure 5b, for the input feature X_{n-1}, we first use multi-scale and different modality convolutions to extract local features X_{n-1}^l and contour features X_{n-1}^c. To reduce the computational complexity caused by multiple recursion, depthwise separable convolution was used as the convolution for different modes. Additionally, we employ dilated convolutions to expand the receptive field to capture fine local details. The process is as follows:

$$M_1 = C_3^{d=3}[C_3\{C_1(X_{n-1})\} + \mathrm{DC}_3\{C_1(X_{n-1})\}]. \tag{17}$$

$$M_2 = \mathrm{DC}_3^{d=3}[\mathrm{DC}_3\{C_1(X_{n-1})\}]. \tag{18}$$

$$M_3 = C_3^{d=5}(M_1) + X_{n-1}^l. \tag{19}$$

$$M_4 = \mathrm{DC}_3^{d=5}\big(M_2 + M_1 + X_{n-1}^c\big). \tag{20}$$

$$X_{n-1}^{l} = C_3(M_1). \tag{21}$$

$$X_{n-1}^{c} = C_3(M_2). \tag{22}$$

$$X_n = \text{Softmax}\{G_c^a(M_3, M_4)\} * M_4. \tag{23}$$

where M_1, M_2, M_3, and M_4 represent intermediate feature outputs, and $C_3^{d=3}$ and $C_3^{d=5}$ represent 3×3 convolutions with dilation rates of 3 and 5, respectively. $\text{DC}_3^{d=3}$ and $\text{DC}_3^{d=5}$ represent depthwise separable convolutions with dilation rates of 3 and 5, respectively.

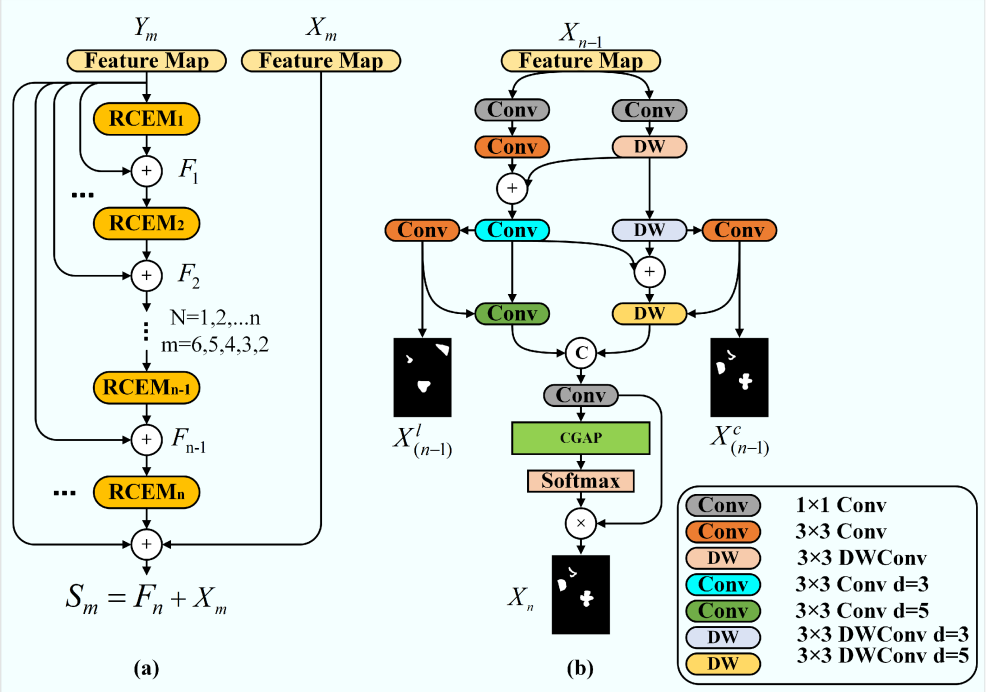


Figure 5. The overall structure of the proposed recursive block (**a**) and recursive contour extraction (RECM) module (**b**). N denotes the number of RECMs, and m represents the sequence number of recursive blocks in the hierarchy of RCR.

3.6. Hybrid Loss

Our training loss is defined as the sum of the RCR module loss L_r. L_r represents the loss of the k-th side output, and there are a total of five loss terms. In the RCR module, we extract five levels of feature maps $M_{r(i)}^{(H\times W\times 1)}$ for loss supervision, where $H = W = \frac{224}{2^{i-1}}$, and $1 \le i \le 5$.

$$L_r = \sum_{i=1}^{5} L\left(M_{r(i)}^{(H\times W\times 1)}\right). \tag{24}$$

To further enhance the model's sensitivity and precision in capturing fine details, we abandoned the traditional single BCE loss and instead adopted the hybrid loss function L proposed in Basnet [65]:

$$L(k) = L_{\text{BCE}}(k) + L_{\text{SSIM}}(k) + L_{\text{IoU}}(k). \tag{25}$$

where $k \in [0,1]^{(H\times W\times 1)}$ represents the loss supervision mapping image. $L_{\text{BCE}}(k)$, $L_{\text{SSIM}}(k)$, and $L_{\text{IoU}}(k)$ represent the application of binary cross-entropy loss (BCE), structural similarity loss (SSIM), and intersection-over-union loss (IoU) to k, respectively. The BCE loss is widely used in binary classification tasks and can effectively handle imbalanced data:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]. \tag{26}$$

where N is the number of samples, y_i represents the true label of the i-th sample (either 0 or 1), and p_i denotes the probability predicted by the model that the i-th sample belongs to the positive class. SSIM is a metric used to measure the similarity between two images. It takes into account not only pixel-level similarity but also factors such as brightness, contrast, and structural information. The formula for calculating SSIM is expressed as

$$L_{\text{SSIM}} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{27}$$

where μ_x and μ_y represent the means of x and y, respectively. σ_x^2 and σ_y^2 denote the variances of x and y, respectively. σ_{xy} represents the covariance between x and y. C_1 and C_2 are small constants to avoid division by zero. Typically, these constants are set to $C_1 = 0.01^2$ and $C_2 = 0.03^2$. IOU mainly measures the overlap between the predicted bounding box and the ground truth bounding box and is commonly used to evaluate a model's ability to accurately locate objects within an image. The IOU is calculated as follows:

$$L_{\text{IoU}} = 1 - \frac{\sum_{x=1}^{H}\sum_{y=1}^{W} S(x,y)G(x,y)}{\sum_{x=1}^{H}\sum_{y=1}^{W}(S(x,y) + G(x,y) - S(x,y)G(x,y))}. \tag{28}$$

where $S(x,y) \in \{0,1\}$ represents the predicted saliency probability for the pixel (x,y), and $G(x,y) \in \{0,1\}$ denotes the ground truth label for the pixel (x,y).

4. Experiments and Analysis

In this section, we first provide a detailed overview of the FLS image dataset, experimental parameters, and evaluation metrics. We then conduct extensive experiments based on the FLS image benchmark dataset, comparing and analyzing various SOD methods. Finally, we further validate the effectiveness and advantages of the proposed method through ablation experiments.

4.1. FLS Image Dataset

We evaluated the feasibility of the proposed method on the marine debris forward-looking dataset (MDFD) [66] benchmark dataset. Given the sparse distribution of marine debris in natural environments, the MDFD utilized a high-frequency ARIS Explorer 3000 sensor to simulate a realistic marine environment in a 3 × 2 × 4 m tank. The dataset includes 11 prominent target objects and one background object. The MDFD comprises 1868 images, each with a resolution of 480 × 320 pixels, and each image contains at least one prominent target. The target objects are primarily composed of household items and typical marine debris. Household items include beverage cartons, tin cans, plastic bottles, etc., while other types include chains, fish hooks, propellers, rubber tires, and more, comprehensively reflecting the diverse characteristics of marine debris. Figure 6 shows the real target types.

The total number of pixels for each object, the number of occurrences, and their corresponding proportions are shown in Table 1. It is worth noting that in the MDFD, apart from walls, tires have the highest proportion, while most other targets have proportions below 10%. This indicates a significant class imbalance in the MDFD, which may affect the model's accuracy in detecting rare targets. Additionally, due to the complex underwater environment, FLS images often come with a lot of noise and may easily overlook small targets with lower prominence, as shown in Figure 7. These challenges severely impact the accuracy and robustness of SOD in FLS images.

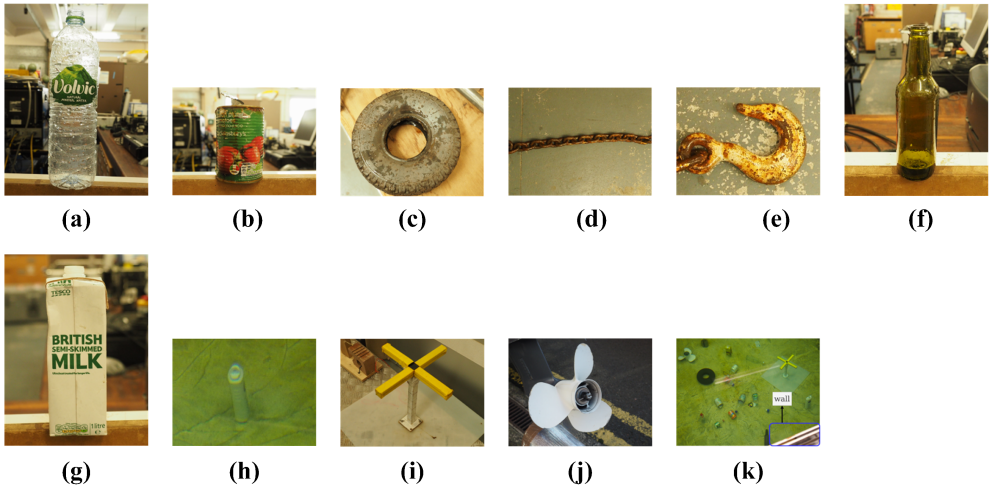


Figure 6. Contains different types of samples. (**a**) Bottle; (**b**) can; (**c**) tire; (**d**) chain; (**e**) hook; (**f**) standing bottle; (**g**) drink carton; (**h**) shampoo bottle; (**i**) valve; (**j**) propeller; (**k**) wall.

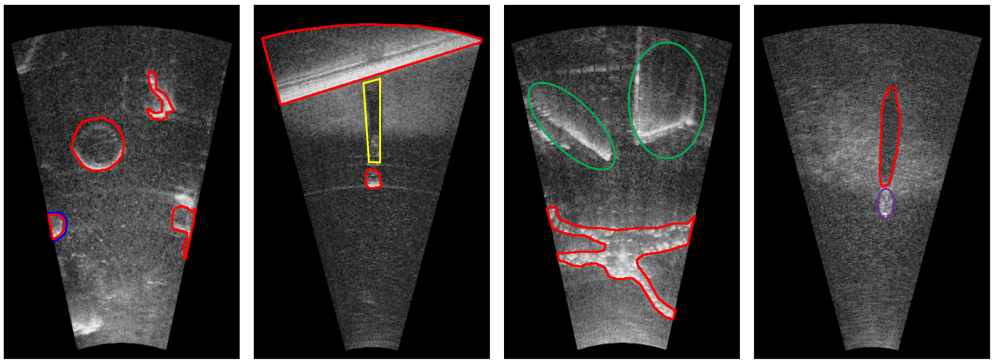


Figure 7. Examples of noise in FLS images are shown. The red areas indicate the ground truth. The blue areas represent small targets that are easily lost due to spatial positioning. The yellow areas show shadow noise caused by occlusions. The cyan areas depict scattering noise caused by sound waves encountering suspended particles, bubbles, and other media. The purple areas illustrate pseudo-target noise caused by reflection noise from water waves.

Table 1. The statistical information of MDFD. TT: object type; TP: total number of object pixels; TPP: total proportion of object pixels; TOC: number of target occurrences; PTO: proportion of object occurrence times.

TT	TP (10^6)	TTP (%)	TOC	PTO (%)
Bottle	0.7	6.6	400	10
Can	0.7	6.6	380	9.5
Tire	1.6	15	620	15.6
Chain	1.4	13.2	370	9.3
Hook	0.2	1.8	180	4.5
Standing Bottle	0.1	0.94	150	3.7
Drink Carton	0.3	2.8	380	9.5
Shampoo Bottle	0.3	2.8	160	4
Valve	0.1	0.94	240	6
Propeller	0.5	4.7	210	5.2
Wall	4.7	44	880	22.1

4.2. Evaluation Metrics

To quantitatively analyze the effectiveness of the proposed method, we used seven commonly used evaluation metrics in SOD tasks: precision–recall (PR) curve, F-measure

curve, mean absolute error (MAE), F-measure (F_β), E-measure ($E_{\tilde{\xi}}$), structural similarity measure (S_α), and boundary displacement error (BDE).

(1) MAE: Measures the average absolute error between the model's predictions and the ground truth. Generally, a lower MAE value indicates more accurate predictions by the model.

$$\mathrm{MAE} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} |P(x,y) - G(x,y)|. \tag{29}$$

(2) F-measure: F_β is a commonly used metric to evaluate the performance of binary classification models, taking into account both the accuracy and recall of the model.

$$F_\beta = \frac{(1+\beta^2) \times \mathrm{Precision} \times \mathrm{Recall}}{\beta^2 \times \mathrm{Precision} + \mathrm{Recall}}. \tag{30}$$

(3) E-measure: $E_{\tilde{\xi}}$ is commonly used to simultaneously evaluate the local and global similarity between predicted images and real images.

$$E_{\tilde{\xi}} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi_s(x,y). \tag{31}$$

(4) Boundary displacement error: BDE is commonly used to measure the average displacement error of boundary pixels between predicted targets and real labels.

$$\mathrm{BDE}_{(R,G)} = \frac{\sum_x f_{y \in G}^{\min}(d(x,y))}{2N_x} + \frac{\sum_y f_{x \in R}^{\min}(d(x,y))}{2N_y}. \tag{32}$$

where R and G represent the boundary pixel sets of the predicted image and the ground truth, respectively. x and y denote the pixels in R and G. N_x and N_y are the number of pixels in R and G, respectively. $f^{\min}(*)$ indicates the minimum value operation, and $d(*)$ represents the Euclidean distance between pixels.

(5) structural similarity measure: S_α is a metric used to assess the perceptual similarity of object structures.

$$S_\alpha = \alpha \times S_0 + (1-\alpha) \times S_r. \tag{33}$$

where S_0 and S_r represent the structural similarity of the object perception and the structural similarity of the region perception, respectively. $\alpha \in [0,1]$, based on recommendations from the literature [67], we set α to 0.5.

4.3. Implementation Details

In our experimental setup, we utilized the PyTorch library and accelerated computations with the RTX 4090 GPU. We divided the MDFD, which contains 1868 images, into training, validation, and test sets with a ratio of 7:2:1. To prevent overfitting, we applied data augmentation techniques, including random cropping, flipping, and shifting. Before training the model, we resized the input images from 480 × 320 to 320 × 320, and then randomly cropped them to 224 × 224. During training, we used the Adam optimizer with a cosine annealing learning rate scheduler. The learning rate was set to 1×10^{-5}, with a cycle length of 50, a minimum learning rate of 1×10^{-6}, and decay parameters β_1 and β_2 set to 0.9 and 0.99, respectively. Weight decay was set to 0. The entire model training took approximately 40 h with a batch size of 4. After 300 iterations, the model loss gradually converged. During testing, input images were resized to 288 × 288 to generate saliency maps, which were then restored to the original image size using bilinear interpolation.

4.4. Comparison with State-of-the-Art Methods

In this section, we compare our proposed method with 25 state-of-the-art methods, including 7 hybrid CNN-Transformer models, 16 pure CNN architecture models, and 2 pure Transformer models. To ensure fairness and comparability, we strictly used the publicly available code and default parameter settings provided by the authors of each model, and standardized the batch size to 4 for retraining. It is important to note that different models have varying input image size requirements (e.g., the Abiu-Net model uses an input size of 384×384). To preserve the original design of each model, we retained their respective default input size settings during the experiments to ensure evaluation under their optimal configurations. The models compared are: Amulet [68], BASNet [65], CPDNet [69], DASNet [70], EGNet [71], F3Net [72], GATENet [73], GCPANet [15], PGNet [74], PICANet [75], POOLNet [76], R3Net [77], RASNet [78], SCRNet [46], U2Net [79], ABiU-Net [24], CTIF-Net [26], M3Net [20], CMTFNet [80], MINet [81], SAM2-UNet [82], MDSAM [83], TSCNet [84], VST [22], and Gelenet [85].

(1) Model complexity analysis: Table 2 shows the comparison between the proposed model and other model parameter quantities. From the table, it can be seen that the parameter count of the hybrid CNN–Transformer architecture model is generally higher than that of the pure CNN structure models and the pure Transformer architecture models. It is noteworthy that ABiU-Net, CMTFNet, and Gelenet achieve relatively lower parameter counts by fusing features through concatenation or using a small amount of convolutional parameters. In addition, MINet focuses on lightweight SOD tasks, resulting in a significant reduction in parameter count. Compared to other hybrid CNN–Transformer architecture models (CTIF-Net, SAM2-UNet), our model maintains a moderate balance in complexity with a moderate number of parameters.

Table 2. Comparison of computational complexity of different methods. C: CNNs; T: Transformer; C-T: CNNs + Transformer.

Methods	Year	Type	Param (M)
Amulet	2017	C	33.1
BASNet	2019	C	87.0
CPDNet	2019	C	47.8
DASNet	2020	C	36.6
EGNet	2019	C	100.9
F3Net	2020	C	25.5
GATENet	2020	C	128.6
GCPANet	2020	C	67.0
PGNet	2022	C	72.7
PICANet	2018	C	47.2
POOLNet	2019	C	68.2
R3Net	2018	C	56.1
RASNet	2018	C	24.7
SCRNet	2019	C	25.2
U2Net	2020	C	44.0
ABiU-Net	2023	C-T	35.8
CTIF-Net	2023	C-T	331.4
M3Net	2023	T	30.3
CMTFNet	2023	C-T	30.0
MINet	2024	C	2.7
SAM2-UNet	2024	C-T	216.5
MDSAM	2024	C-T	100.4
TSCNet	2023	C-T	103.5
VST	2021	T	44.5
Gelenet	2023	C-T	24.4
Ours	–	C-T	128.3

(2) Quantitative analysis: In order to visually present the overall performance of the proposed model, we first plotted the PR curve and F-measure curve. As shown in Figure 8, the PR curve of FLSSNet significantly approaches the upper right, indicating that the model can maintain a balance between accuracy and recall at different thresholds. At the same time, the F-measure curve also exhibits significant advantages, demonstrating its robustness and generalization ability in complex scenarios. Table 3 visually displays the quantitative results of different models in MAE, E_ξ, S_α, F_β, and BDE. The experimental data results show that FLSSNet has a competitive advantage in these key performance indicators. Compared to pure CNN-based models, FLSSNet leverages the RCR module's strength in edge contour feature extraction, resulting in a 63.2% reduction in the boundary displacement error (BDE) metric compared to the best-performing DASNet. Additionally, although FLSSNet has 1.27 times more parameters than EGNet, the BDE metric of EGNet is 2.57 times higher than that of FLSSNet. Even with GATENet having nearly the same number of parameters as FLSSNet, FLSSNet outperforms GATENet across all key metrics, including MAE, E_ξ, S_α, F_β, and BDE. In comparison to pure Transformer-based models such as VST and M3Net, FLSSNet achieves a 0.02 improvement in the MAE metric while also showing gains of 0.012 and 0.035 in the S_α metric, respectively. Compared to the CNN–Transformer hybrid module, FLSSNet still demonstrates strong competitiveness. It should be noted that the parameter count of SAM2-UNet is 1.68 times that of FLSSNet, but the increase in E_ξ, S_α, and F_β is only 0.003, 0.021, and 0.008, respectively.

Table 3. Quantitative results across five evaluation metrics for different methods. Both MAE and BDE are better when lower, while E_ξ, S_α, and F_β are better when higher.

Methods	MAE	E_ξ	S_α	F_β	BDE
Amulet	0.07	0.916	0.856	0.756	9.7801
BASNet	0.10	0.936	0.850	0.786	9.8883
CPDNet	0.05	0.988	0.909	0.878	4.9946
DASNet	0.04	0.982	0.921	0.875	4.4328
EGNet	0.07	0.974	0.893	0.852	4.1996
F3Net	0.05	0.972	0.915	0.862	4.795
GATENet	0.07	0.971	0.908	0.842	4.453
GCPANet	0.15	0.938	0.872	0.792	4.6254
PGNet	0.25	0.897	0.809	0.775	5.9908
PICANet	0.10	0.912	0.863	0.723	3.7405
POOLNet	0.05	0.965	0.912	0.842	3.87
R3Net	0.06	0.953	0.915	0.820	5.3555
RASNet	0.06	0.969	0.920	0.854	2.6697
SCRNet	0.19	0.900	0.812	0.790	9.2854
U2Net	0.06	0.947	0.916	0.808	3.3527
ABiU-Net	0.06	0.976	0.901	0.805	3.7601
CTIF-Net	0.06	0.977	0.896	0.801	5.3970
M3Net	0.06	0.966	0.889	0.853	3.9688
CMTFNet	0.10	0.943	0.841	0.760	9.2087
MINet	0.14	0.872	0.788	0.661	11.2171
SAM2-UNet	0.03	0.976	0.945	0.880	1.4227
MDSAM	0.12	0.903	0.811	0.701	14.6836
TSCNet	0.17	0.914	0.759	0.715	53.4374
VST	0.06	0.906	0.912	0.747	3.7963
Gelenet	0.08	0.971	0.909	0.843	29.5761
Ours	0.04	0.973	0.924	0.872	1.631

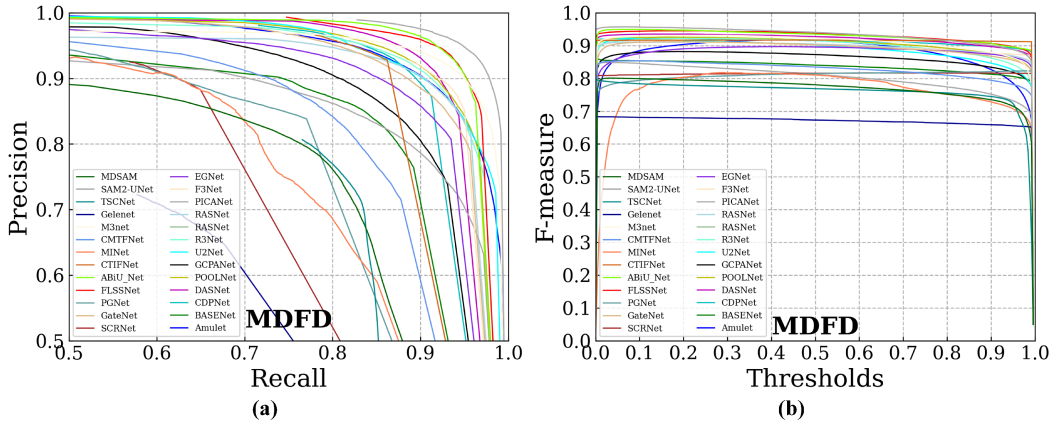


Figure 8. The comparison between FLSSNet and its comparison model on the PR curve (**a**) and F-measure curve (**b**). Please zoom in to view.

(3) Qualitative analysis: As shown in Figure 9, the pure CNN method faces some shortcomings and deficiencies in processing SOD tasks in FLSS images. Firstly, when the image contains multiple salient targets, the pure CNN model has shortcomings in capturing the global spatial structure information of the targets, resulting in unclear distinction between the targets and the background, which, in turn, leads to partial loss of salient targets. For example, Basnet (line 10) failed to fully predict all salient targets. Relatively speaking, although pure Transformer models can effectively capture global information, they have limitations in handling local details. For example, VST (lines 1, 5, 7) and M3Net (line 11) models can recognize all salient targets but perform poorly in detail preservation. Although the CNN–Transformer hybrid architecture model has achieved good overall accuracy, most models have not been able to finely process the edges of salient targets, for example, CTIF-net (line 5) and Mdsam (lines 10, 12). In contrast, FLSSNet has certain advantages in pixel long-range correlation modeling and edge refinement. Taking the fifth row as an example, FLSSNet demonstrates excellent pixel long-range correlation modeling ability, which can accurately capture the correlation between distant pixels. Additionally, as observed from column 11, FLSSNet excels in the refinement and accuracy of edge contours, delivering clearer and more precise object boundaries.

Figure 10 clearly illustrates the dynamic evolution of feature transformation layer by layer during the decoding process of FLSSNet. We can observe that as features are progressively restored from S5 to S1, the edge contours and detail information become increasingly refined and enhanced. The visualization results of the deep feature S5 show that the model is able to accurately capture the spatial structural information of salient objects and effectively differentiate between the salient object and the background, eliminating non-salient feature redundancy.

4.5. Ablation Experiment

To further analyze the performance and characteristics of the proposed modules, we conducted extensive ablation experiments for both qualitative and quantitative evaluation. The proposed model primarily consists of the TFC module, LRCA module, and RCR module. We choose the method of single module ablation, and the backbone network still adopts the CNN–Transformer hybrid architecture. At the same time, we delete the TFC module, LRCA module, and RCR module in sequence to obtain three variant models: FLSSNet-T, FLSSNet-L, and FLSSNet-R. Finally, to validate the impact of the Transformer on extracting global spatial structural information, we removed the P2T backbone network, resulting in the pure CNN-based variant model, FLSSNet-C. Additionally, we evaluated

the roles of the LRCA and RCR modules within the FLSSNet-C model, producing two further variants: FLSSNet-CL and FLSSNet-CR.

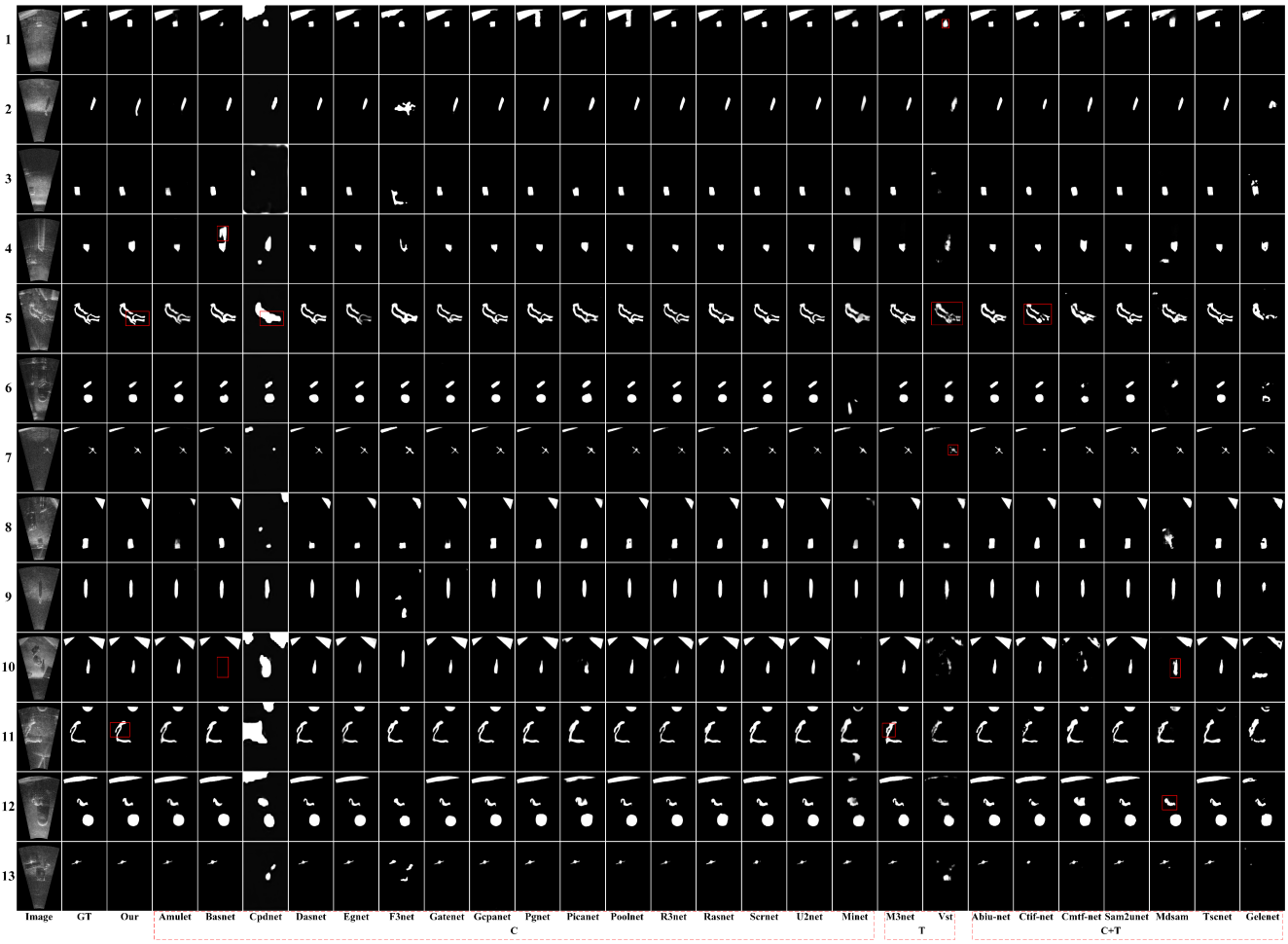


Figure 9. Visual display of FLSSNet and comparison models. The red box indicates significant differences.

(1) Effectiveness of individual modules within the CNN–Transformer hybrid backbone: To validate the effectiveness of the proposed modules, we conducted ablation experiments by sequentially removing individual modules. Additionally, when the RCR module was further ablated, the variant model was trained using an end-to-end training approach.

(a) Quantitative analysis: As shown clearly in Table 4, removing any single module results in a significant decline in the overall performance of FLSSNet. This is particularly evident in the boundary displacement error (BDE) metric, where the BDE increases to 4.8 times that of the original model upon removal of the RCR module. Relatively speaking, due to the retention of the RCR module in the FLSSNet-T and FLSSNet-L variant models, their BDE metrics are at an approximate level. From the MAE values in Table 4, it can be further observed that FLSSNet-L, due to the absence of the LRCA module, experiences significantly weakened long-range pixel correlation modeling. Consequently, its MAE value is the highest among FLSSNet and its variants. This indicates that the LCRA module plays an indispensable role in capturing remote dependencies and improving prediction accuracy, and its absence directly leads to severe degradation of model performance. In addition, as shown in Figure 11, we can see from the PR curve and F-measure curve that ablating the TFC module also significantly reduces the performance of the model, further indicating the importance of addressing feature misalignment. The ablation experiment results emphasize the core role of collaborative effects between different modules in improving model performance, as well as the critical impact of global and local feature fusion on SOD.

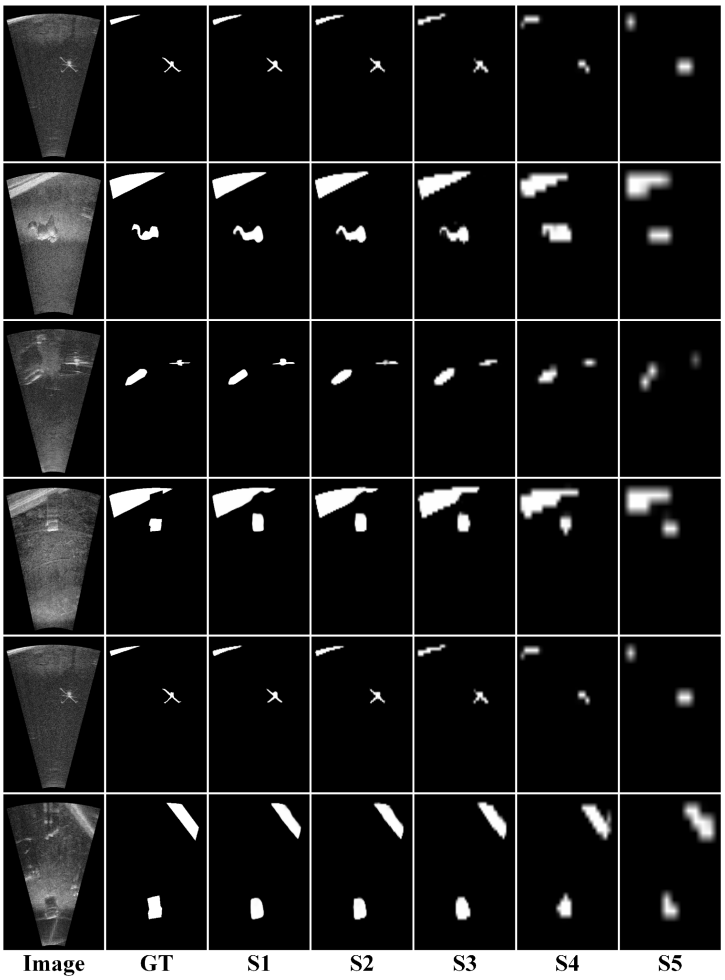


Figure 10. Visualization results of side outputs at different levels of the recursive block.

Table 4. Evaluation results of variant models for ablating individual modules in CNN–Transformer hybrid architecture.

Methods	TFC	LRCA	RCR	MAE ↓	E_ξ ↑	S_α ↑	F_β ↑	BDE ↓
FLSSNet-T		✓	✓	0.09	0.924	0.816	0.759	6.5030
FLSSNet-L	✓		✓	0.1	0.904	0.802	0.736	6.0296
FLSSNet-R	✓	✓		0.08	0.926	0.876	0.797	7.8640
FLSSNet	✓	✓	✓	0.04	0.973	0.924	0.872	1.631

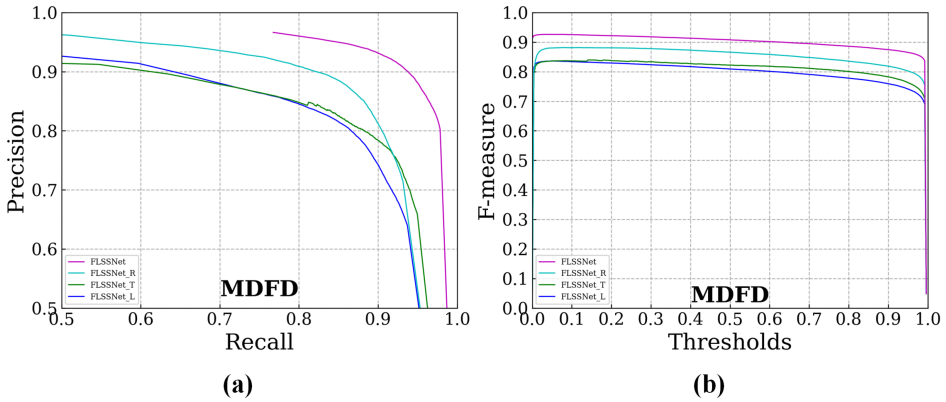


Figure 11. (**a**,**b**) Quantitative comparison of the variant models within the CNN–Transformer hybrid backbone architecture in the PR (precision–recall) and F-measure curves.

(b) Qualitative analysis: Figure 12 shows the visual comparison of variant models after ablating different modules. In order to further analyze the role of each module in the model, we conducted visual comparative analysis from dimensions such as local details, global spatial structure, and contour edges. From the visualization results in columns 1 and 10 of Figure 12, it can be clearly seen that ablation of either the TFC or LRCA module will lead to a significant decline in the overall performance of the model, which is consistent with the trend shown in the PR curve and F-measure curve in Figure 11. Only when both the TFC and LRCA modules are retained can the model achieve high prediction accuracy, as demonstrated by the performance of FLSSNet-R in the visualized images. However, although FLSSNet-R can accurately capture salient objects, it still exhibits shortcomings in perceiving boundary contours. The boundary predictions are relatively coarse and lack clarity, as demonstrated in the visualized results of columns 7 and 9. Additionally, we further found that due to the absence of the TFC module in FLSSNet-T, the model performs poorly in learning the synergistic consistency between CNN and Transformer features. This results in weaker predictions of the target's spatial structure, as evidenced by the visualized results of FLSSNet-T in columns 2 and 8. FLSSNet-L, due to the lack of modeling capability for long-range correlations between target pixels, leads to missed detections and false positives of salient objects, as shown in the visualized results of columns 4 and 9.

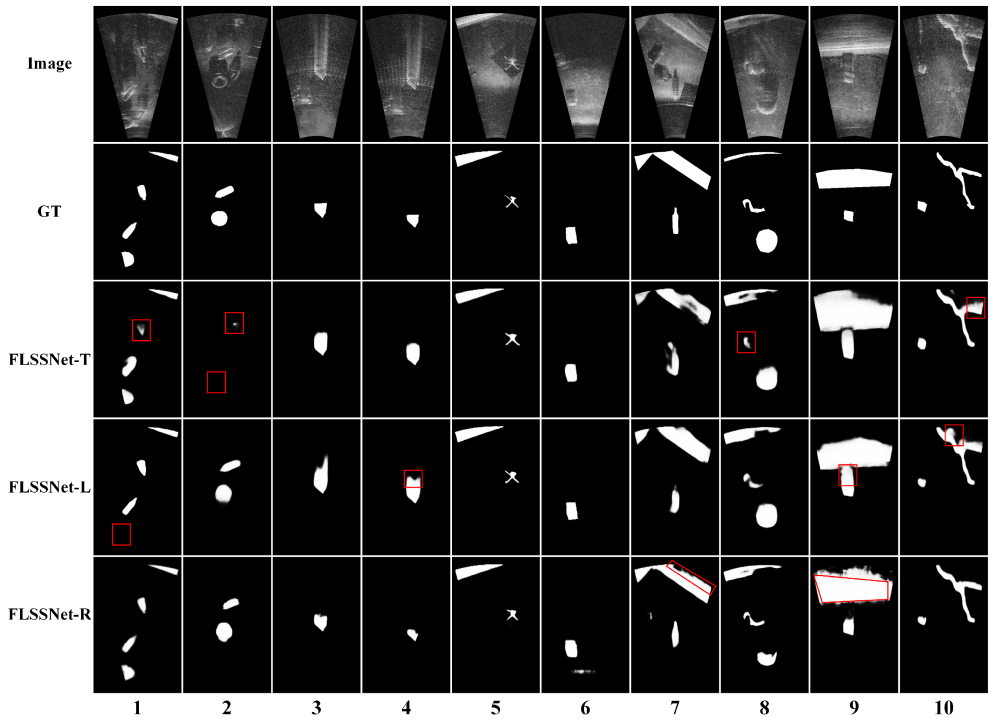


Figure 12. Visualization results of variant models in CNN–Transformer hybrid backbone architecture.

(2) Effectiveness of individual modules within a pure CNN backbone architecture: To evaluate the performance of the modules in a pure CNN structure, we removed the Transformer backbone network, P2T. Notably, with the removal of the Transformer backbone, the TFC module is no longer required. We continued to use both quantitative and qualitative analysis methods to assess the overall performance of individual modules in different architectures.

(a) Quantitative Analysis: As shown in Table 5, after the ablation of the P2T backbone from FLSSNet, there is a sharp increase in the MAE. Additionally, when the LRCA module is further ablated from FLSSNet-C, the MAE does not exhibit significant changes and remains at a relatively high level. We believe that the primary function of the LRCA module is to

address the limitations of CNNs in capturing long-range dependencies during multimodal fusion. Since FLSSNet-C's backbone relies on CNN for feature extraction and lacks the Transformer's capability to model global dependency features, its overall performance is constrained. This limitation results in similar trends across the five evaluation metrics: MAE, F_β, E_ξ, S_α, and BDE. Comparing FLSSNet-C with FLSSNet-CR, it can be further found that after ablation of the RCR module, the BDE index of FLSSNet-CR reached the highest value, mainly due to the significant loss of edge contour information, which led to the MAE index also being at the highest level.

(b) Qualitative analysis: As shown in the visualization results of lines 1 and 3 of Figure 13, the variant models FLSSNet-C, FLSSNet-CL, and FLSSNet-CR have shortcomings in capturing spatial structural information of significant targets, resulting in overall performance degradation and widespread missed detections. In addition, the visualization results in the second row indicate that FLSSNet-CR performs the weakest in edge perception, resulting in significant deviations in predicting edge contour regions, leading to blurry and unclear contours.

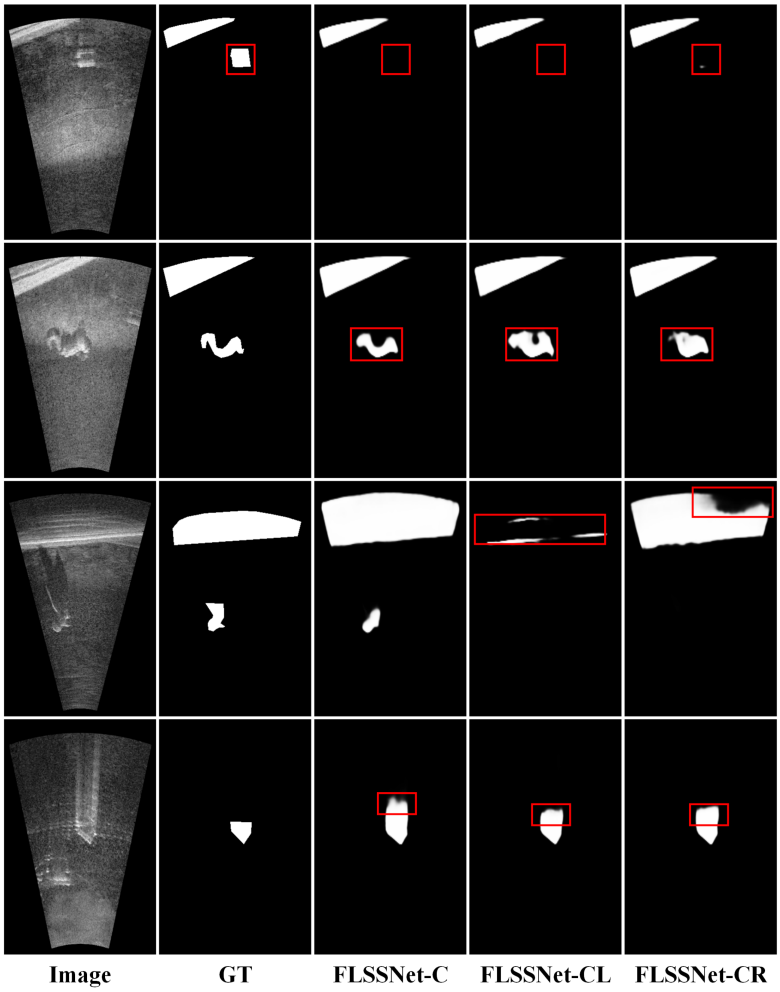


Figure 13. The visualization results of a single module in a pure CNN backbone architecture.

Table 5. Evaluation results of variant models with single module ablation in pure CNN backbone architecture.

Methods	P2T	TFC	LRCA	RCR	MAE ↓	E_ξ ↑	S_α ↑	F_β ↑	BDE ↓
FLSSNet-C			✓	✓	0.14	0.925	0.801	0.743	5.3618
FLSSNet-CL				✓	0.14	0.936	0.809	0.746	5.6788
FLSSNet-CR			✓		0.15	0.937	0.827	0.752	6.0104
FLSSNet	✓	✓	✓	✓	0.04	0.973	0.924	0.872	1.631

(3) The synergy between strip convolution module and OSSM: In order to further analyze the impact of the multi-directional convolution(MDC) module and omnidirectional selective scan module (OSSM) on the long-range correlation of salient target pixels, we conducted ablation experiments in FLSSNet and FLSSNet-C separately, removing MDC and OSSM from LRCAM and constructing variant models such as FLSSNet-Ld, FLSSNet-Lo, FLSSNet-Cd, and FLSSNet-Co to systematically evaluate the contribution of these key modules to model performance.

(a) Quantitative analysis: As shown in Table 6, when comparing the MAE indicators of FLSSNet, FLSSNet-Ld, and FLSSNet-Lo models, we observed that the ablation of MDC and OSSM both led to a significant increase in MAE. In addition, the comparison of BDE indicators shows that the effect of OSSM ablation on BDE is more significant. We speculate that MDC is mainly used as a module for optimizing long-range dependencies in convolutional networks, and its performance is limited by the size of the receptive field. However, the feature extraction ability of OSSM is insufficient compared to CNN, so the MAE of FLSSNet-Lo is higher than that of FLSSNet-Ld. Therefore, we chose to extract local long-range dependencies from MDC and further capture global long-range dependencies through OSSM. In addition, there was no significant difference in the four evaluation indicators between FLSSNet-Cd and FLSSNet-Co. As revealed by the analysis of FLSSNet-CL, FLSSNet-Cd and FLSSNet-Co lack the ability to fully model global dependency features, resulting in limited performance of MDC and OSSM.

Table 6. Quantitative results of the synergy between MDC and OSSM.

Methods	P2T	TFC	DCM	OSSM	MAE ↓	E_ξ ↑	S_α ↑	F_β ↑	BDE ↓
FLSSNet-Ld	✓	✓		✓	0.15	0.928	0.788	0.748	5.6426
FLSSNet-Lo	✓	✓	✓		0.14	0.912	0.819	0.721	8.8738
FLSSNet-Cd				✓	0.15	0.923	0.795	0.739	6.0830
FLSSNet-Co			✓		0.15	0.931	0.795	0.741	6.5504
FLSSNet	✓	✓	✓	✓	0.04	0.973	0.924	0.872	1.631

(b) Qualitative analysis: As shown in the first picture of Figure 14, due to the lack of spatial perception ability in FLSSNet-Cd and FLSSNet-Co, the performance of MDC and OSSM is limited, resulting in poor overall prediction performance. Compared with FLSSNet-Lo, FLSSNet-Ld performs better in local perception and detail capture, as shown by the propeller edge in the second image and the valve notch in the third image in Figure 14. OSSM has shortcomings in processing local perception, making it difficult to accurately capture tiny pixel details. However, MDC exhibits limitations in capturing distant pixel correlations, making it difficult to effectively integrate global information. In summary, we believe that the LRCAM effectively compensates for the CNN's deficiency in expressing long-range correlations during multimodal feature fusion. The local long-range dependency modeling of MDC and the global long-range correlation modeling of OSSM have a synergistic effect.

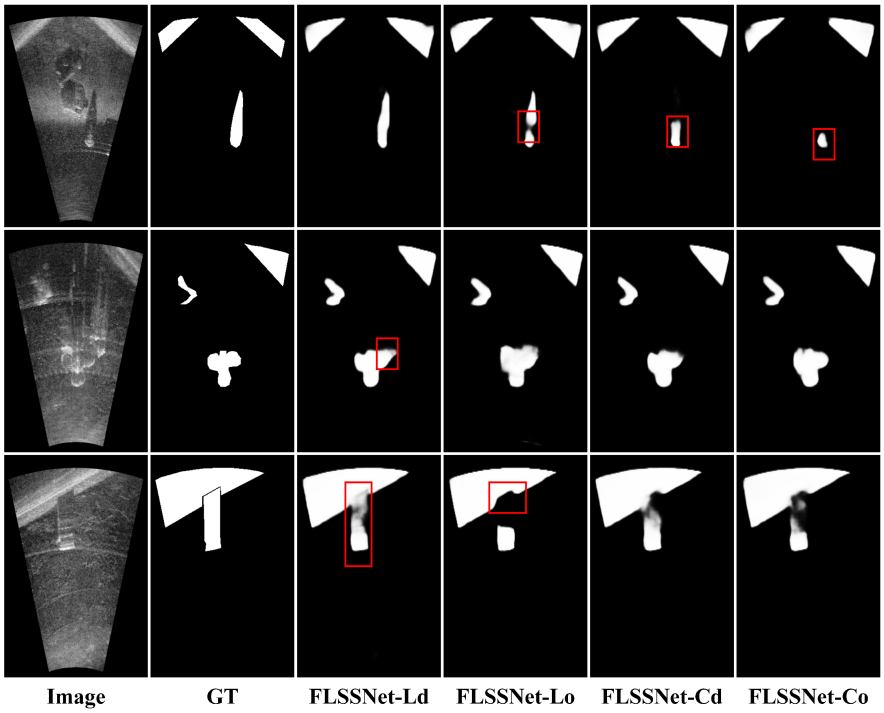


Figure 14. Visualization of ablation experiments using MDC and OSSM in CNN–Transformer hybrid architecture and pure CNN architecture.

(4) The effectiveness of the recursive contour extraction module: To further evaluate the effectiveness of the recursive contour extraction module (RCEM) in boundary detection, we quantitatively compared the changes in the boundary detection error (BDE) index of the model for different numbers of RCEM modules. From Table 7, it can be observed that when N is 2, the BDE index is in its optimal state. In addition, we further explored the contribution of RCEM in extracting local details and edge information. When N = 2, we present the visualization results of $X^{l}_{(n-1)}$ and $X^{c}_{(n-1)}$ of the second layer RCEM at different levels of RCR, represented by $S1L, S2L, \ldots, S5L$ and $S1C, S2C, \ldots, S5C$, respectively, where S represents the level number of RCR. As shown in the first image of Figure 15, deep features are gradually extracted and enhanced with local and edge spatial information through RCEM, which is consistent with the visualization results of the side output of the RCR module. Comparing the shallow features S1L and S1C (as shown in the second, third, and fourth images of Figure 15, we found that the S1C feature is superior to the S1L feature in capturing edge details. Meanwhile, the third and fourth figures in Figure 15 demonstrate the better extraction of local information by the deep feature S3L in the early stages. Therefore, we ultimately integrated local and edge features in the RCEM module for subsequent feature transfer and optimization.

Table 7. Quantitative results of the impact of different numbers of RCE on the BDE metric.

Methods	N	MAE ↓	E_{ξ} ↑	S_{α} ↑	F_{β} ↑	BDE ↓
FLSSNet-R-1	1	0.057	0.799	0.830	0.631	3.8041
FLSSNet	2	0.04	0.973	0.924	0.872	1.631
FLSSNet-R-3	3	0.049	0.820	0.812	0.833	1.7383
FLSSNet-R-4	4	0.043	0.897	0.867	0.845	1.7020

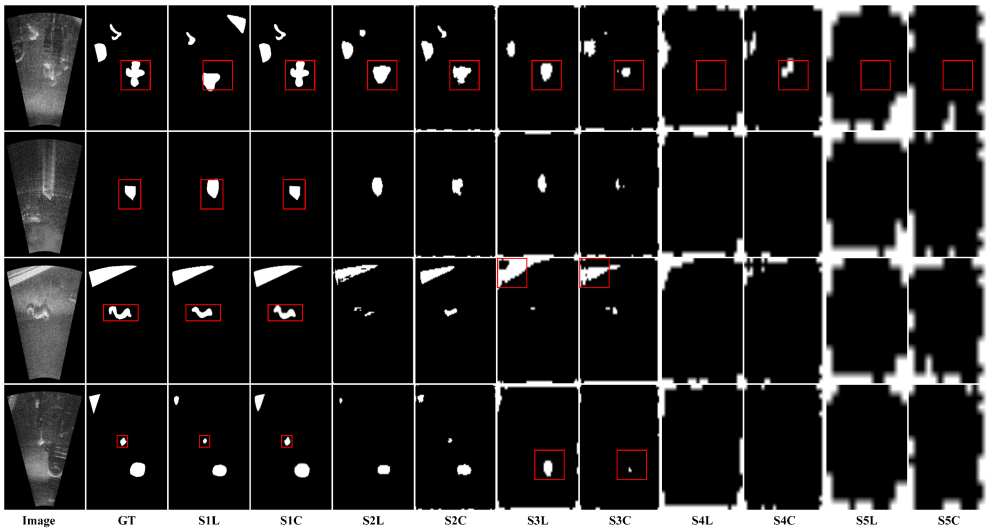


Figure 15. In different levels of RCR, the visualization results of $X^l_{(n-1)}$ and $X^c_{(n-1)}$ in the second layer of RCEM are presented. Here, $(s1, s2, \ldots, s5)$ represent the hierarchical sequence of RCEM, while L and C, respectively, denote the specific visualization results of $X^l_{(n-1)}$ and $X^c_{(n-1)}$.

5. Conclusions

Salient object detection (SOD) in forward-looking sonar (FLS) images has significant application value in tasks such as underwater environmental perception, exploration, and underwater rescue. In this paper, we propose an innovative hybrid network based on the CNN–Transformer–Mamba architecture, called FLSSNet, for SOD tasks in FLS images. FLSSNet introduces a two-stage strategy combining multimodal feature fusion and a recursive mechanism. Specifically, the core of FLSSNet lies in effectively integrating the local contextual information extracted by the CNN with the global spatial structure information captured by the Transformer. In the feature fusion stage, to address potential misalignment issues between different features, FLSSNet introduces the Transformer feature conversion module (TFC). This module leverages multi-scale convolution and self-attention mechanisms to effectively capture the synergy and correlation between CNN and Transformer features, ensuring the precision of feature fusion. Traditional encoder–decoder architectures mostly employ skip connections to directly link features from different levels. However, due to CNN's bias towards local features, it struggles to effectively capture long-range dependencies between target pixels. To address this issue, FLSSNet introduces the long-range correlation augmentation (LRCA) module, which strengthens the long-range correlations between target pixels through the synergistic use of convolution, sequential selective scanning, and attention mechanisms, effectively suppressing noise interference. Furthermore, to enhance the model's ability to refine edge contours, FLSSNet incorporates a recursive contour refinement (RCR) module. This module progressively extracts features from the decoder's output through multi-layer recursive loops, strengthening edge detection. To evaluate the advantages and feasibility of the proposed method, we compared FLSSNet with 25 state-of-the-art SOD methods and conducted ablation experiments to validate the effectiveness of the introduced modules. We analyzed the performance gains brought by different modules to the overall model, resulting in variant models FLSSNet-T, FLSSNet-L, and FLSSNet-R. Additionally, we examined the performance of these modules in both hybrid architectures and pure CNN architectures, leading to variant models FLSSNet-C, FLSSNet-CL, and FLSSNet-CR. We also compared the impact of different numbers of recursive contour extraction modules on the overall model. The experimental results demonstrated that FLSSNet outperformed in multiple key metrics, including mean

absolute error (MAE), F-measure (F_β), E-measure (E_ξ), structural similarity measure (S_α), and boundary displacement error (BDE).

Author Contributions: Investigation, J.L. (Jiayuan Li); Resources, S.R.; Data curation, Z.L.; Writing—original draft, J.L. (Juan Lei); Writing—review & editing, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Science Foundation of China under Grant 62171368, in part by the Science, Technology and Innovation of Shenzhen Municipality under Grant KJZD20230923115505011 and JCYJ20241202124931042.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Igbinenikaro, O.P.; Adekoya, O.O.; Etukudoh, E.A. Emerging underwater survey technologies: A review and future outlook. *Open Access Res. J. Sci. Technol.* **2024**, *10*, 071–084. [CrossRef]
2. Sughasini, K.; Rishitha, S.; Sreelaxmi, R.; PS, S.P.; Kunanidhi, G.C.; As, K. Underwater Search and Rescue Using Drone. In Proceedings of the 2024 International Conference on Computing and Data Science (ICCDS), Chennai, India, 26–27 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–4.
3. Kim, A.; Eustice, R.M. Real-time visual SLAM for autonomous underwater hull inspection using visual saliency. *IEEE Trans. Robot.* **2013**, *29*, 719–733. [CrossRef]
4. Yang, H.; Zhou, T.; Jiang, H.; Yu, X.; Xu, S. A Lightweight Underwater Target Detection Network for Forward-looking Sonar Images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2525113. [CrossRef]
5. Zhang, X.; Yang, P.; Cao, D. Vision Synthetic aperture image enhancement with near-coinciding Nonuniform sampling case. *Comput. Electr. Eng.* **2024**, *120*, 109818. [CrossRef]
6. Abu, A.; Diamant, R. Enhanced fuzzy-based local information algorithm for sonar image segmentation. *IEEE Trans. Image Process.* **2019**, *29*, 445–460. [CrossRef]
7. Mignotte, M.; Collet, C.; Perez, P.; Bouthemy, P. Sonar image segmentation using an unsupervised hierarchical MRF model. *IEEE Trans. Image Process.* **2000**, *9*, 1216–1231. [CrossRef] [PubMed]
8. Karoui, I.; Fablet, R.; Boucher, J.M.; Augustin, J.M. Seabed segmentation using optimized statistics of sonar textures. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 1621–1631. [CrossRef]
9. Tian, Y.; Lan, L.; Guo, H. A review on the wavelet methods for sonar image segmentation. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420936091. [CrossRef]
10. Ge, Y.; Zhang, Q.; Xiang, T.Z.; Zhang, C.; Bi, H. Tcnet: Co-salient object detection via parallel interaction of transformers and cnns. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 2600–2615. [CrossRef]
11. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
12. Jia, N.; Sun, Y.; Liu, X. TFGNet: Traffic salient object detection using a feature deep interaction and guidance fusion. *IEEE Trans. Intell. Transp. Syst.* **2023**, *25*, 3020–3030. [CrossRef]
13. Liu, Y.; Zhou, L.; Wu, G.; Xu, S.; Han, J. Tcgnet: Type-correlation guidance for salient object detection. *IEEE Trans. Intell. Transp. Syst.* **2023**, *25*, 6633–6644. [CrossRef]
14. Xia, C.; Sun, Y.; Li, K.C.; Ge, B.; Zhang, H.; Jiang, B.; Zhang, J. Rcnet: Related context-driven network with hierarchical attention for salient object detection. *Expert Syst. Appl.* **2024**, *237*, 121441. [CrossRef]
15. Chen, Z.; Xu, Q.; Cong, R.; Huang, Q. Global context-aware progressive aggregation network for salient object detection. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10599–10606.
16. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
17. Zhang, X.; Su, Y.; Gao, L.; Bruzzone, L.; Gu, X.; Tian, Q. A lightweight transformer network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–17. [CrossRef]
18. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.

19. Jain, J.; Singh, A.; Orlov, N.; Huang, Z.; Li, J.; Walton, S.; Shi, H. Semask: Semantically masked transformers for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 752–761.
20. Yuan, Y.; Gao, P.; Tan, X. M3Net: Multilevel, Mixed and Multistage Attention Network for Salient Object Detection. *arXiv* **2023**, arXiv:2309.08365.
21. Chen, J.; Zhang, H.; Gong, M.; Gao, Z. Collaborative compensative transformer network for salient object detection. *Pattern Recognit.* **2024**, *154*, 110600. [CrossRef]
22. Liu, N.; Zhang, N.; Wan, K.; Shao, L.; Han, J. Visual saliency transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4722–4732.
23. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 558–567.
24. Zhang, X.; Yang, P.; Dai, X. Vision Focusing multireceiver SAS data based on the fourth-order legendre expansion. *Circuits Syst. Signal Process.* **2019**,*38*, 2607–2629. [CrossRef]
25. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
26. Yuan, J.; Zhu, A.; Xu, Q.; Wattanachote, K.; Gong, Y. CTIF-Net: A CNN-transformer iterative fusion network for salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 3795–3805. [CrossRef]
27. Zeng, C.; Kwong, S.; Ip, H. Dual swin-transformer based mutual interactive network for RGB-D salient object detection. *Neurocomputing* **2023**, *559*, 126779. [CrossRef]
28. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
29. Zhang, Q.; Lin, J.; Tao, Y.; Li, W.; Shi, Y. Salient object detection via color and texture cues. *Neurocomputing* **2017**, *243*, 35–48. [CrossRef]
30. Liu, G.H.; Yang, J.Y. Exploiting color volume and color difference for salient region detection. *IEEE Trans. Image Process.* **2018**, *28*, 6–16. [CrossRef]
31. Cheng, M.M.; Mitra, N.J.; Huang, X.; Torr, P.H.; Hu, S.M. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 569–582. [CrossRef]
32. Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; Li, S. Salient object detection: A discriminative regional feature integration approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2083–2090.
33. Ji, Y.; Zhang, H.; Zhang, Z.; Liu, M. CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances. *Inf. Sci.* **2021**, *546*, 835–857. [CrossRef]
34. Han, J.; Zhang, D.; Cheng, G.; Liu, N.; Xu, D. Advanced deep-learning techniques for salient and category-specific object detection: A survey. *IEEE Signal Process. Mag.* **2018**, *35*, 84–100. [CrossRef]
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2016; pp. 770–778.
36. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
37. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
38. Wang, W.; Shen, J.; Cheng, M.M.; Shao, L. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5968–5977.
39. Fang, C.; Tian, H.; Zhang, D.; Zhang, Q.; Han, J.; Han, J. Densely nested top-down flows for salient object detection. *Sci. China Inf. Sci.* **2022**, *65*, 182103. [CrossRef]
40. Zhu, G.; Li, J.; Guo, Y. PriorNet: Two Deep Prior Cues for Salient Object Detection. *IEEE Trans. Multimed.* **2023**, *26*, 5523–5535. [CrossRef]
41. Zhu, J.; Qin, X.; Elsaddik, A. Dc-net: Divide-and-conquer for salient object detection. *Pattern Recognit.* **2024**, *157*, 110903. [CrossRef]
42. Li, J.; Su, J.; Xia, C.; Ma, M.; Tian, Y. Salient object detection with purificatory mechanism and structural similarity loss. *IEEE Trans. Image Process.* **2021**, *30*, 6855–6868. [CrossRef] [PubMed]
43. Tu, Z.; Ma, Y.; Li, C.; Tang, J.; Luo, B. Edge-guided non-local fully convolutional network for salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 582–593. [CrossRef]

44. Kousik, N.; Natarajan, Y.; Raja, R.A.; Kallam, S.; Patan, R.; Gandomi, A.H. Improved salient object detection using hybrid Convolution Recurrent Neural Network. *Expert Syst. Appl.* **2021**, *166*, 114064. [CrossRef]
45. Go, J.; Ryu, J. Spatial bias for attention-free non-local neural networks. *Expert Syst. Appl.* **2024**, *238*, 122053. [CrossRef]
46. Wu, Z.; Su, L.; Huang, Q. Stacked cross refinement network for edge-aware salient object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 7264–7273.
47. Zhang, L.; Wu, J.; Wang, T.; Borji, A.; Wei, G.; Lu, H. A multistage refinement network for salient object detection. *IEEE Trans. Image Process.* **2020**, *29*, 3534–3545. [CrossRef] [PubMed]
48. Wan, B.; Zhou, X.; Zheng, B.; Yin, H.; Zhu, Z.; Wang, H.; Sun, Y.; Zhang, J.; Yan, C. LFRNet: Localizing, focus, and refinement network for salient object detection of surface defects. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–12. [CrossRef]
49. Diko, A.; Avola, D.; Cascio, M.; Cinque, L. ReViT: Enhancing vision transformers feature diversity with attention residual connections. *Pattern Recognit.* **2024**, *156*, 110853. [CrossRef]
50. Gao, L.; Liu, B.; Fu, P.; Xu, M. Adaptive spatial tokenization transformer for salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602915. [CrossRef]
51. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
52. Yan, R.; Yan, L.; Geng, G.; Cao, Y.; Zhou, P.; Meng, Y. ASNet: Adaptive Semantic Network Based on Transformer-CNN for Salient Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5608716. [CrossRef]
53. Lou, J.; Lin, H.; Marshall, D.; Saupe, D.; Liu, H. TranSalNet: Towards perceptually relevant visual saliency prediction. *Neurocomputing* **2022**, *494*, 455–467. [CrossRef]
54. Wang, Q.; Liu, Y.; Xiong, Z.; Yuan, Y. Hybrid feature aligned network for salient object detection in optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
55. Hatamizadeh, A.; Kautz, J. MambaVision: A Hybrid Mamba-Transformer Vision Backbone. *arXiv* **2024**, arXiv:2407.08083.
56. Yue, Y.; Li, Z. Medmamba: Vision mamba for medical image classification. *arXiv* **2024**, arXiv:2403.03849.
57. Zhao, S.; Chen, H.; Zhang, X.; Xiao, P.; Bai, L.; Ouyang, W. Rs-mamba for large remote sensing image dense prediction. *arXiv* **2024**, arXiv:2404.02668. [CrossRef]
58. Liu, X.; Zhang, C.; Zhang, L. Vision mamba: A comprehensive survey and taxonomy. *arXiv* **2024**, arXiv:2405.04404.
59. Zhang, X.; Tan, C.; Ying, W. Vision An imaging algorithm for multireceiver synthetic aperture sonar. *Remote Sens.* **2019**, *11*, 672 [CrossRef]
60. Wu, Y.H.; Liu, Y.; Zhan, X.; Cheng, M.M. P2T: Pyramid pooling transformer for scene understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 12760–12771. [CrossRef]
61. Zhang, X.; Yang, P.; Feng, X.; Sun, H. Vision Efficient imaging method for multireceiver SAS. *IET Radar Sonar Navig.* **2022**, *16*, 1470–1483 [CrossRef]
62. Zhang, X.; Yang, P.; Cao, D. Vision An efficient method for the simulation of multireceiver SAS raw signal. *Multimed. Tools Appl.* **2024**, *83*, 37351–37368 [CrossRef]
63. Sun, T.; Di, Z.; Che, P.; Liu, C.; Wang, Y. Leveraging crowdsourced GPS data for road extraction from aerial imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7509–7518.
64. Ke, Y.Y.; Tsubono, T. Recursive contour-saliency blending network for accurate salient object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 2940–2950.
65. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. Basnet: Boundary-aware salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7479–7489.
66. Singh, D.; Valdenegro-Toro, M. The marine debris dataset for forward-looking sonar semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3741–3749.
67. Fan, D.P.; Cheng, M.M.; Liu, Y.; Li, T.; Borji, A. Structure-measure: A new way to evaluate foreground maps. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4548–4557.
68. Zhang, P.; Wang, D.; Lu, H.; Wang, H.; Ruan, X. Amulet: Aggregating multi-level convolutional features for salient object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 202–211.
69. Wu, Z.; Su, L.; Huang, Q. Cascaded partial decoder for fast and accurate salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3907–3916.
70. Zhao, J.; Zhao, Y.; Li, J.; Chen, X. Is depth really necessary for salient object detection? In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1745–1754.

71. Zhao, J.X.; Liu, J.J.; Fan, D.P.; Cao, Y.; Yang, J.; Cheng, M.M. EGNet: Edge guidance network for salient object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 8779–8788.
72. Wei, J.; Wang, S.; Huang, Q. F^3Net: Fusion, feedback and focus for salient object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12321–12328.
73. Zhao, X.; Pang, Y.; Zhang, L.; Lu, H.; Zhang, L. Suppress and balance: A simple gated network for salient object detection. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part II 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 35–51.
74. Xie, C.; Xia, C.; Ma, M.; Zhao, Z.; Chen, X.; Li, J. Pyramid grafting network for one-stage high resolution saliency detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11717–11726.
75. Liu, N.; Han, J.; Yang, M.H. PiCANet: Pixel-wise contextual attention learning for accurate saliency detection. *IEEE Trans. Image Process.* **2020**, *29*, 6438–6451. [CrossRef]
76. Liu, J.J.; Hou, Q.; Cheng, M.M.; Feng, J.; Jiang, J. A simple pooling-based design for real-time salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3917–3926.
77. Deng, Z.; Hu, X.; Zhu, L.; Xu, X.; Qin, J.; Han, G.; Heng, P.A. R3net: Recurrent residual refinement network for saliency detection. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; AAAI Press: Menlo Park, CA, USA, 2018; Volume 684690.
78. Chen, S.; Tan, X.; Wang, B.; Hu, X. Reverse attention for salient object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 234–250.
79. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
80. Wu, H.; Huang, P.; Zhang, M.; Tang, W.; Yu, X. CMTFNet: CNN and multiscale transformer fusion network for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 2004612. [CrossRef]
81. Shen, K.; Zhou, X.; Liu, Z. MINet: Multiscale Interactive Network for Real-Time Salient Object Detection of Strip Steel Surface Defects. *IEEE Trans. Ind. Inform.* **2024**, *20*, 7842–7852. [CrossRef]
82. Xiong, X.; Wu, Z.; Tan, S.; Li, W.; Tang, F.; Chen, Y.; Li, S.; Ma, J.; Li, G. SAM2-UNet: Segment Anything 2 Makes Strong Encoder for Natural and Medical Image Segmentation. *arXiv* **2024**, arXiv:2408.08870.
83. Gao, S.; Zhang, P.; Yan, T.; Lu, H. Multi-Scale and Detail-Enhanced Segment Anything Model for Salient Object Detection. *arXiv* **2024**, arXiv:2408.04326.
84. Wang, X.; Cheng, W.; Feng, Y.; Song, R. TSCNet: Topological structure coupling network for change detection of heterogeneous remote sensing images. *Remote Sens.* **2023**, *15*, 621. [CrossRef]
85. Zhang, X.; Yang, P.; Sun, M. Experiment results of a novel sub-bottom profiler using synthetic aperture technique. *Curr. Sci.* **2022**, *122*, 461–464. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

Contrastive Feature Disentanglement via Physical Priors for Underwater Image Enhancement

Fei Li [1], Li Wan [2], Jiangbin Zheng [3], Lu Wang [1,*] and Yue Xi [4]

[1] School of Computer Science and Technology, Xidian University, Xi'an 710071, China; lifei@xidian.edu.cn
[2] Hangzhou Institute of Technology, Xidian University, Hangzhou 311200, China; wanli@stu.xidian.edu.cn
[3] School of Software and Microelectronics, Northwestern Polytechnical University, Xi'an 710071, China; zhengjb@nwpu.edu.cn
[4] Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China; xiyue@xidian.edu.cn
* Correspondence: wanglu@xidian.edu.cn

Abstract: Underwater image enhancement (UIE) serves as a fundamental preprocessing step in ocean remote sensing applications, encompassing marine life detection, archaeological surveying, and subsea resource exploration. However, UIE encounters substantial technical challenges due to the intricate physics of underwater light propagation and the inherent homogeneity of aquatic environments. Images captured underwater are significantly degraded through wavelength-dependent absorption and scattering processes, resulting in color distortion, contrast degradation, and illumination irregularities. To address these challenges, we propose a contrastive feature disentanglement network (CFD-Net) that systematically addresses underwater image degradation. Our framework employs a multi-stream decomposition architecture with three specialized decoders to disentangle the latent feature space into components associated with degradation and those representing high-quality features. We incorporate hierarchical contrastive learning mechanisms to establish clear relationships between standard and degraded feature spaces, emphasizing intra-layer similarity and inter-layer exclusivity. Through the synergistic utilization of internal feature consistency and cross-component distinctiveness, our framework achieves robust feature extraction without explicit supervision. Compared to existing methods, our approach achieves a 12% higher UIQM score on the EUVP dataset and outperforms other state-of-the-art techniques on various evaluation metrics such as UCIQE, MUSIQ, and NIQE, both quantitatively and qualitatively.

Keywords: contrastive feature disentanglement; physical priors; underwater image enhancement

Academic Editors: Haiyong Xu, Jingyue Lu, Lei Zhang, Marc Pinto, Farman Ali and Xuebo Zhang

Received: 16 December 2024
Revised: 3 February 2025
Accepted: 20 February 2025
Published: 22 February 2025

Citation: Li, F.; Wan, L.; Zheng, J.; Wang, L.; Xi, Y. Contrastive Feature Disentanglement via Physical Priors for Underwater Image Enhancement. *Remote Sens.* **2025**, *17*, 759. https://doi.org/10.3390/rs17050759

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Underwater optical imaging has evolved into an indispensable modality in contemporary ocean remote sensing, including marine biology research, coral reef monitoring, underwater archaeology, and autonomous underwater vehicle (AUV) navigation [1]. However, the quality of underwater optical images often suffers from degradation due to various degradation factors, such as light scattering, absorption, and color distortion [2,3]. Such degradation mechanisms manifest in compromised visibility, attenuated contrast, and distorted color fidelity, which impede critical tasks like object detection, recognition, and tracking [4,5].

To address these challenges, underwater image enhancement (UIE) techniques aim to restore image clarity and improve visibility. Recent studies [6] highlight the significance

of UIE as a preprocessing step for enhancing the performance of vision-based systems in underwater environments. Despite significant progress, existing methods often fail to adequately address the inherent complexities of underwater imaging, particularly when applied to real-world scenarios with diverse environmental conditions.

Compared to in-air image enhancement tasks, such as image denoising [7], dehazing [8], and low-light enhancement [9,10], UIE presents distinctive challenges [2,11,12]. These challenges stem from the unique properties of the underwater environment, including the following: (1) light scattering and absorption: light scattering reduces image clarity, while absorption causes wavelength-dependent color distortions, leading to the predominance of blue and green hues in underwater images; (2) scene homogeneity: the uniformity of underwater scenes, with limited texture and natural contrast, makes it difficult for enhancement algorithms to distinguish between degraded and authentic features; (3) data scarcity: collecting large-scale paired datasets of degraded and clean underwater images is inherently challenging due to the dynamic and unpredictable nature of underwater environments. Most existing datasets rely on synthetic image pairs, which fail to fully capture real-world complexities, worsening the domain shift issue between synthetic and real images.

These inherent challenges have motivated the development of various UIE methodologies, which can be systematically categorized into three main approaches: traditional model-based [13–15], deep learning-based [7–10,16–19], and physics-guided learning-based methods [11,20–22].

Traditional model-based approaches utilize specific mathematical models to simulate the underwater imaging process, relying on prior knowledge for parameter estimation, such as Dark Channel Prior (DCP) [13] and Retinex [14,15]. However, these approaches are often constrained by their inability to adapt to the varied environmental conditions encountered in real-world underwater scenarios.

Deep learning-based methods have significantly advanced image restoration tasks [7–10]. Such data-driven approaches excel in learning latent features from large datasets and modeling the relationship between degraded and enhanced images. Recent UIE-specific methods [16–19] have demonstrated promising results, regardless of whether they use paired or unpaired training datasets. However, these methods encounter significant challenges: (1) the domain gap between synthetic and real-world underwater images limits generalization capabilities; (2) the scarcity of paired datasets makes pixel-to-pixel mapping difficult during training; and (3) the black-box nature of deep learning models reduces interpretability, making it harder to trust their outputs in critical applications.

To overcome the limitations associated with paired samples, semisupervised and unsupervised methods have been explored. Generative Adversarial Networks (GANs) [1] and contrastive learning (CL) [18,19] have demonstrated promise by enabling the use of unpaired datasets. However, these approaches often experience instability during training and require physical priors to facilitate cross-domain feature transfer.

Unlike traditional and deep learning-based approaches, physics-guided learning-based methods embed physical priors into neural network architectures, improving interpretability and enhancing generalization. These methods leverage physical information to guide network design [11,20,21] and establish optimal solution boundaries [22]. While promising, these approaches often lack mechanisms to fully leverage unpaired data or to handle the inherent variability of underwater environments.

In this paper, we propose a novel hybrid framework that combines supervised and unsupervised learning paradigms along with physical priors in a unified architecture to overcome the limitations of existing methods. The primary challenge in UIE arises from the difficulty in handling the variability of underwater environments and the lack of paired

data for supervised training. While many existing methods either rely on synthetic datasets or do not effectively address real-world variability, our framework uniquely integrates both supervised and unsupervised learning, making it capable of effectively leveraging unpaired data from real-world underwater images.

Our CFD-Net introduces a multi-stream decomposition approach that encodes degraded underwater images into a latent embedding space, separating components, including clean background, transmission maps, and ambient light, into distinct branches. This decomposition helps preserve the physical characteristics of the scene, enabling better alignment with the physical degradation process commonly encountered in underwater environments. By maintaining the realism of each generative component, our method ensures the estimated components are accurate and more reliable for subsequent enhancement. Specifically, we incorporate hierarchical contrastive learning mechanisms within a physics-guided architecture, enabling effective cross-domain feature learning to address the domain gap between synthetic and real-world underwater images. This allows our model to not only learn from synthetic data but also adapt well to the challenging variations of real-world underwater environments. Furthermore, we propose a novel cross-space contrastive loss function, which enhances feature separability within individual spaces and exclusivity between distinct spaces, improving the quality of decomposition and leading to superior image enhancement results.

This is the first approach that employs contrastive learning and physical feature decomposition for UIE, offering a unique perspective on underwater image restoration. By effectively addressing both the challenges of unpaired data and the inherent variability of underwater imaging, our method outperforms existing state-of-the-art techniques in both quantitative and qualitative evaluations. The primary contributions of this work are summarized as follows:

- We propose a novel unsupervised UIE method leveraging contrastive feature decomposition, offering a new perspective for underwater image enhancement.
- We introduce a unique cross-space and content contrastive loss, facilitating the simultaneous exploration of intra-similarity within latent spaces and inter-exclusiveness between feature spaces.
- Comprehensive experiments are conducted to evaluate the proposed method, demonstrating its outstanding performance both quantitatively and qualitatively.

The remainder of this paper is organized as follows: Section 2 discusses related work in UIE. Section 3 details the proposed methodology. Section 4 presents the experimental results and comparisons with state-of-the-art methods. Finally, Section 5 concludes the paper and summarizes key findings.

2. Related Work

2.1. Traditional Methods

Underwater image enhancement predominantly relies on image processing and computer vision techniques. These non-physical methods, distinct from physical or deep learning models, function at the pixel level to enhance visual quality, such as histogram equalization [23], white balancing [24], and Retinex [14]. Ancuti et al. [25] proposed a multi-scale fusion method that combines white balancing with traditional contrast histogram equalization, enhancing the quality of the original image. Hitam et al. [26] introduced a hybrid Contrast Limited Adaptive Histogram Equalization (CLAHE) technique, employing the Euclidean norm to merge outputs from RGB and HSV color models, thereby improving contrast and reducing noise. Additionally, Ren et al. [15] developed the Low-Rank Regularized Retinex Model (LR3M), which combines a low-rank prior with Retinex decom-

position to reduce noise in reflection images. While these methods are computationally efficient and easy to implement, they rely heavily on handcrafted priors and often fail to address the physical properties of underwater environments, limiting their effectiveness in complex scenarios.

2.2. Physical Methods

Inspired by the image degradation process, many low-quality restoration methods have been developed [13,27–30], each reconstructing the degradation process by estimating key parameters like light and transmission. Drawing inspiration from the Dark Channel Prior (DCP) used in dehazing tasks [27], Drews et al. [28] introduced the Underwater DCP (UDCP), which consists of transmission estimation methods tailored for underwater scenarios, primarily focusing on the blue and green channels to account for the rapid attenuation of red light underwater. Similarly, Galdran et al. [29] introduced the Red Channel Prior, which acknowledges the rapid attenuation of red wavelengths in water, thus providing another perspective on underwater image enhancement. Extending this concept, Peng et al. [13] introduced a technique incorporating depth-dependent color transformations for estimating environmental light. This approach enables adaptive color correction within the image formation framework, enhancing image clarity. Lastly, Akkaynak et al. [30] modified the conventional imaging model to address the limitations of standard models designed for atmospheric conditions. Liang et al. [31] introduced the enhancement method CDCR to restore color, detail, and contrast of degraded images. Despite their interpretability and reliance on well-defined mathematical models, physical methods often fail to generalize across diverse underwater environments. They rely on accurate parameter estimation, which is challenging in real-world scenarios with varying light conditions, turbidity levels, and depth variations.

2.3. Learning-Based Methods

Deep learning approaches have significantly advanced underwater image enhancement by overcoming the limitations inherent in traditional methods. Early methods, primarily based on Convolutional Neural Networks (CNNs), concentrated on extracting low-level feature representations for tasks such as color correction and contrast restoration. For example, Wang et al. [32] developed UIE-Net, a CNN-based approach aiming at color correction and dehazing. Similarly, Li et al. [33] introduced UWCNN, which integrates underwater scene priors to synthesize underwater images and reconstruct enhanced outputs without requiring explicit physical parameter estimation.

However, these methods often require large-scale paired datasets of degraded and clean images, which are challenging to obtain in underwater scenarios due to environmental variability and data scarcity. To address this issue, semisupervised and unsupervised learning frameworks have gained prominence in UIE tasks [11,34–36]. Notably, Li et al. [34] introduced WaterGAN to generate color-corrected underwater images, while Li et al. [35] adopted the CycleGAN [36] to facilitate unpaired feature domain transfer between degraded and enhanced images, reducing reliance on paired datasets. Chen et al. [11] proposed a generator with embedded physical priors for synthesizing training data and improving real-world underwater image enhancement.

More recently, contrastive learning (CL) has shown significant promise for self-supervised feature representation learning [37–40], particularly in image enhancement and restoration tasks [18,41,42]. For example, Wu et al. [41] introduced a contrast regularization technique leveraging CL for dehazing tasks. Liu et al. [42] applied a twin adversarial contrastive learning framework for underwater image enhancement. Huang et al. [18]

proposed the Semi-UIR model, which integrates unlabeled data using CL to improve underwater image restoration.

Despite these advancements, existing learning-based methods continue to face critical challenges: (1) generative models often fail to fully capture the relationship between factor-related feature distributions and the physical degradation process, leading to suboptimal image restoration; (2) current methods primarily focus on mapping degraded images to clean counterparts but lack mechanisms to enforce independence among decomposed components, such as backgrounds and transmission maps; (3) the domain gap between synthetic and real-world underwater images remains a major barrier to generalization.

To address these limitations, we propose a novel contrastive multi-stream decomposition architecture that leverages physical priors and contrastive learning strategies. By embedding physical knowledge into latent feature spaces and utilizing cross-space contrastive loss functions, our method enhances generalization, improves interpretability, and effectively bridges the domain gap. This approach uniquely integrates the decomposition of degraded components (e.g., clean background, transmission maps, and ambient light layers) with cross-domain feature learning, providing a more robust solution for underwater image enhancement.

3. Proposed Method

To systematically address the inherent challenges of underwater image enhancement, we propose a physics-guided multi-stream decomposition framework providing theoretical foundations. The proposed framework comprises three key components: (1) a physics-guided image formation model that establishes the theoretical foundation and mathematical formulations; (2) a multi-stream decomposition architecture that disentangles various degradation factors; and (3) a hierarchical contrastive learning mechanism that ensures the distinctiveness of features.

In this section, we first introduce the physics-guided image formation model (Section 3.1), then detail our network architecture (Section 3.2), and, finally, present the learning objectives (Section 3.3).

3.1. Physics-Guided Image Formation Model

The underwater image formation process is governed by wavelength-dependent optical principles that characterize light propagation in aquatic media. For each color channel $c \in \{R, G, B\}$, the formation model is:

$$I^c(x) = J^c(x)t^c(x) + (1 - t^c(x))A^c \tag{1}$$

where $I^c(x)$ denotes the observed degraded image, $J^c(x)$ represents the scene radiance, $t^c(x)$ characterizes the transmission map, and A^c indicates the ambient light. The transmission follows the Beer–Lambert law:

$$t^c(x) = \exp(-\beta^c d(x)) \tag{2}$$

where β^c represents the wavelength-dependent attenuation coefficient for channel c and $d(x)$ denotes the scene depth at position x. The attenuation coefficients satisfy $\beta^R > \beta^G > \beta^B$, which explains the characteristic blue-green color cast in underwater images.

To facilitate efficient network optimization while preserving physical consistency, we derive a simplified formulation based on two key observations from underwater optics. First, the transmission maps across RGB channels exhibit strong correlation due to the continuous nature of the attenuation spectrum:

$$T(x) = \mathbb{E}_c[t^c(x)] = \frac{1}{3} \sum_{c \in \{R,G,B\}} t^c(x) \tag{3}$$

where $T(x)$ represents the unified transmission map that captures the average attenuation effect across all channels. Second, the ambient light variation primarily affects intensity rather than spectral distribution:

$$\mathbf{A} = \mathbb{E}_c[\mathbf{A}^c] + \Delta\mathbf{A} \tag{4}$$

where $\mathbf{A}$ denotes the mean ambient light intensity and $\Delta\mathbf{A}$ represents the channel-specific intensity variation. These physical insights lead to our simplified formation model:

$$\mathbf{I} = \mathbf{J} \cdot \mathbf{T} + (1 - \mathbf{T}) \cdot \mathbf{A} \tag{5}$$

This physically grounded model naturally suggests a feature decomposition strategy where the degraded underwater image can be disentangled into three components: clean background $\mathbf{J}$ representing features free from degradation, transmission map $\mathbf{T}$ capturing depth-dependent attenuation, and ambient light $\mathbf{A}$ encoding global illumination. Such decomposition provides clear guidance for contrastive feature learning—features from different components should be mutually exclusive in the latent space, while features within each component should maintain consistency across varying degradation levels.

Based on this formulation, we design a multi-stream architecture with three specialized decoders: the background decoder ($\mathbf{D_b}$) estimates clean image $\mathbf{J}$, the transmission decoder ($\mathbf{D_t}$) computes transmission map $\mathbf{T}$, and the ambient light decoder ($\mathbf{D_a}$) predicts ambient light $\mathbf{A}$. This physically constrained decomposition framework not only ensures the physical validity of the enhanced results but also provides natural supervision for contrastive feature learning, which will be detailed in the following sections.

3.2. Multi-Stream Decomposition Architecture

Given an unpaired underwater image set $X_U = \{\mathbf{U}\}_{i=1}^{N1}$ and its corresponding high-quality reference set $X_H = \{\mathbf{H}\}_{i=1}^{N2}$, where N_1 and N_2 represent the sample sizes, we design a multi-stream architecture guided by the physical image formation model. The framework aims to learn disentangled representations that capture distinct physical properties: background features representing scene content, transmission features encoding depth-dependent attenuation, and ambient light features characterizing global illumination patterns. As illustrated in Figure 1, the framework consists of a shared encoder $\mathbf{E}$ and three specialized decoders $\{\mathbf{D_b}, \mathbf{D_t}, \mathbf{D_a}\}$.

The feature embedding network : The feature embedding network $\mathbf{E}$ extracts multi-scale features guided by the physics of underwater image formation. Given an input image $\mathbf{U}$, the hierarchical feature extraction process is formulated as:

$$F = \{f_1, f_2, f_3, f_4\} = \mathbf{E}(\mathbf{U}) \tag{6}$$

where f_1 (1/2 scale) preserves fine-grained details for background reconstruction, $\{f_2, f_3\}$ (1/4, 1/8 scales) capture medium-range context for transmission estimation, and f_4 (1/16 scale) encodes global information for ambient light prediction.

Specialized decoders: the framework incorporates three physically motivated decoders, each designed to capture specific aspects of the underwater formation model:

(1) **Background decoder $\mathbf{D_b}$**: following the multiplicative nature of background–transmission interaction in $\mathbf{J} \cdot \mathbf{T}$, the decoder utilizes features at all scales:

$$\mathbf{B} = \mathbf{D_b}(F) = \mathbf{D_b}(\{f_1, f_2, f_3, f_4\}) \tag{7}$$

(2) **Transmission decoder $\mathbf{D_t}$**: guided by the spatial continuity of transmission map $T(x) = \exp(-\beta d(x))$, the decoder focuses on medium- and high-level features:

$$\mathbf{T} = \mathbf{D_t}(F) = \mathbf{D_t}(\{f_2, f_3, f_4\}) \tag{8}$$

(3) **Ambient light decoder $\mathbf{D_a}$**: following the global nature of ambient light $\mathbf{A} = \mathbb{E}_c[\mathbf{A}^c]$, the decoder exclusively uses global semantic features:

$$\mathbf{A} = \mathbf{D_a}(F) = \mathbf{D_a}(\{f_4\}) \tag{9}$$

where $\mathbf{A}$ represents the estimated ambient light, derived from the deepest feature representations.

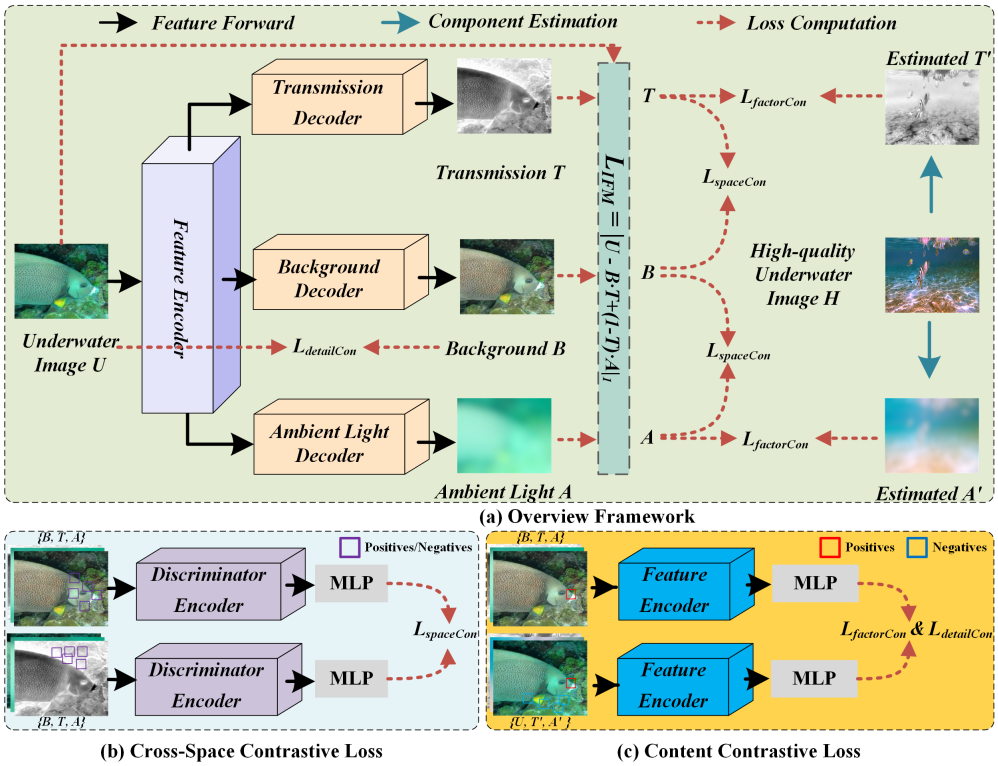


Figure 1. The proposed CFD-Net framework. (**a**) An overview of our framework consisting of three parallel decoders: transmission decoder $\mathbf{D_t}$ for estimating $\mathbf{T}$, background decoder $\mathbf{D_b}$ for reconstructing $\mathbf{B}$, and ambient light decoder $\mathbf{D_a}$ for predicting $\mathbf{A}$. (**b**) Cross-space contrastive loss $\mathcal{L}_{spaceCon}$ for feature alignment and (**c**) content contrastive loss $\mathcal{L}_{detailCon}$ and factor contrastive loss $\mathcal{L}_{factorCon}$ for preserving details.

The network components leverage proven architectures while incorporating task-specific modifications. The transmission decoder $\mathbf{D_t}$ and ambient light decoder $\mathbf{D_a}$ adopt U-Net-like structures [43] for effective parameter estimation. The background decoder $\mathbf{D_b}$ employs a modified ResNet architecture in [44] enhanced with dual-attention mechanisms:

$$\mathbf{D_b}(f) = \mathrm{Conv}_{1\times1}(\sum_{i=1}^{4} \alpha_i \cdot \mathrm{ResBlock}_b(f_i)) \tag{10}$$

where α_i represents learnable attention weights and $\mathrm{ResBlock}_b$ incorporates both spatial and channel attention for comprehensive feature refinement. This physically guided architecture ensures that each component maintains its specific role while enabling efficient feature extraction and enhancement.

3.3. Training Objectives

The training objectives are designed to tackle underwater image enhancement challenges by incorporating physical priors derived from the image formation model (IFM). The IFM provides a mathematical foundation for decomposing an underwater image **U** into its constituent components: clean background **B**, transmission map **T**, and ambient light **A**, as formulated in Equation (5). The UIE task is thus formulated as an ill-posed inverse problem of recovering the clean background **B** from the degraded observation **U** while simultaneously estimating the transmission map **T** and ambient light **A**.

The total loss function combines reconstruction fidelity, physical consistency, and feature learning:

$$\mathcal{L}_{\text{total}} = \lambda_{CL}\mathcal{L}_{\text{CL}} + \lambda_{IFM}\mathcal{L}_{\text{IFM}} + \lambda_{adv}\mathcal{L}_{\text{adv}} \tag{11}$$

where λ_{CL}, λ_{IFM}, and λ_{adv} are hyperparameters that balance the contributions of adversarial learning, physical consistency, and contrastive feature learning, respectively.

3.3.1. Hierarchical Contrastive Learning Function

To effectively disentangle underwater image components while preserving their physical relationships, we propose a hierarchical contrastive learning framework that operates at multiple semantic levels. This framework leverages self-supervised learning to optimize feature representations and guide the decomposition process through carefully designed contrastive objectives. The comprehensive contrastive learning objective integrates three complementary components:

$$\mathcal{L}_{\text{CL}} = \mathcal{L}_{\text{spaceCon}}(\mathbf{B}, \mathbf{T}, \mathbf{A}) + \mathcal{L}_{\text{factorCon}}(\mathbf{T}, \mathbf{T}') + \mathcal{L}_{\text{factorCon}}(\mathbf{A}, \mathbf{A}') + \mathcal{L}_{\text{detailCon}}(\mathbf{B}, \mathbf{U}) \tag{12}$$

where $\mathcal{L}_{\text{spaceCon}}$ enforces inter-component distinctiveness by regulating relationships between background, transmission, and ambient light features, $\mathcal{L}_{\text{factorCon}}$ maintains physical consistency by aligning estimated components with their reference counterparts, and $\mathcal{L}_{\text{detailCon}}$ ensures structural fidelity by preserving fine-grained image details during enhancement.

The framework implements these specialized contrastive losses to jointly optimize three fundamental aspects of the decomposition. First, intra-component coherence requires features within each component space (background, transmission, ambient light) to maintain high correlation, ensuring consistent physical interpretations through positive sample pairs drawn from the same component space. Second, inter-component separation demands features from different physical components to remain well separated in the latent space, preventing information leakage and ensuring clean decomposition through negative sample pairs across component spaces. Third, structural preservation guarantees that the enhanced outputs faithfully preserve the essential structural information present in the original underwater image while maintaining physical validity through detail-preserving contrastive constraints.

This hierarchical design enables our framework to learn physically meaningful and well-separated representations for each component while maintaining the intrinsic relationships that characterize underwater image formation. The following sections detail the specific formulation and implementation of each contrastive objective.

Degraded factor contrastive learning: To systematically model degradation-related features, the framework implements a physics-guided simulation of the underwater degradation process using unpaired image pairs. Consider a reference underwater image I_h^c, where its physical degradation process is characterized by:

$$I_h^c(x) = J_h^c(x)T_h^c(x)) + (1 - T_h^c(x))A_h^c \tag{13}$$

where $J_h^c(x)$ represents the scene radiance, $T_h^c(x)$ denotes the transmission distribution, and A_h^c characterizes the ambient illumination.

Through the introduction of a degradation coefficient $\alpha \in \{0, 1\}$, the degraded image formation I_l^c is mathematically formulated as:

$$\begin{aligned} I_l^c(x) &= J_h^c(x)(\alpha T_h^c(x)) + (1 - \alpha T_h^c(x))A_h^c \\ &= J_h^c(x)T_l^c(x) + (1 - T_l^c(x))A_l^c \end{aligned} \tag{14}$$

As established in Equations (13) and (14), images $I_l^c(x)$ and $I_h^c(x)$ share an identical scene radiance $J_h^c(x)$. The degraded variant $I_l^c(x)$ exhibits enhanced attenuation characteristics due to the constraint $\alpha T_h^c(x) < T_h^c(x)$. Consequently, the estimated transmission map $T_p^c(x) \in \{T\}$ and ambient light $A_p^c(x) \in \{A\}$ should converge to their respective reference values $T_h^c(x)$ and $A_h^c(x)$.

To achieve this objective, the framework defines contrastive objectives for the transmission map **T** and ambient light **A** estimation:

$$P_t(T) = \mathcal{L}_{\text{factorCon}}\left(T_h^c(x), T_p^c(x)\right) \tag{15}$$

$$P_a(A) = \mathcal{L}_{\text{factorCon}}\left(A_h^c(x), A_p^c(x)\right) \tag{16}$$

The framework implements a patch-based contrastive learning strategy where spatially corresponding patches within each decomposed component constitute positive pairs, while non-corresponding patches serve as negative samples. For a degraded input image **U**, the framework extracts feature representations **T** and **A**, with their high-quality counterparts denoted as $\mathbf{T}'$ and $\mathbf{A}'$. Feature embeddings are computed through specialized Residual encoder E_{res} and multi-layer perceptron MLP, yielding representations $v_{A_i} = MLP(E_{res}(P_{A_i}))$, $v_{A_i'} = MLP(E_{res}(P_{A_i'}))$, $v_{T_i} = MLP(E_{res}(P_{T_i}))$, and $v_{T_i'} = MLP(E_{res}(P_{T_i'}))$.

The factor-aware contrastive loss is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{factorCon}} = &\sum_{i=1}^{N} \frac{\exp\left(v_{A_i'} \cdot v_{A_i}/\tau\right)}{\exp\left(v_{A_i'} \cdot v_{A_i}/\tau\right) + \sum_{j=1}^{N} \exp\left(v_{A_j'} \cdot v_{A_i}/\tau\right)} + \\ &\sum_{i=1}^{N} \frac{\exp\left(v_{T_i'} \cdot v_{T_i}/\tau\right)}{\exp\left(v_{T_i'} \cdot v_{T_i}/\tau\right) + \sum_{j=1}^{N} \exp\left(v_{T_j'} \cdot v_{T_i}/\tau\right)} \end{aligned} \tag{17}$$

where N denotes the number of negative samples and τ represents the temperature coefficient. This formulation ensures that corresponding patches (e.g., P_{A_i} and P_{T_i}) exhibit stronger feature correlations with their high-quality counterparts $P_{A_i'}$ and $P_{T_i'}$ compared to randomly sampled negative patches, thereby promoting consistent feature learning for transmission and ambient light estimation.

Detail preservation contrastive learning: The structural and perceptual consistency between the enhanced image **B** and the observed image **U** is systematically enforced through a detail-preserving contrastive mechanism. Spatially corresponding patches from **B** and **U** are designated as positive pairs to maintain local structural fidelity, while non-corresponding patches serve as negative samples. The feature extraction process utilizes the encoder E_{res}, yielding patch-level representations $v_{U_i} = MLP(E_{res}(P_{U_i}))$ and $v_{B_i} = MLP(E_{res}(P_{B_i}))$. The detail preservation contrastive loss is formulated as:

$$\mathcal{L}_{\text{detailCon}} = \sum_{i=1}^{N} \frac{\exp\left(v_{U_i} \cdot v_{B_i}/\tau\right)}{\exp\left(v_{U_i} \cdot v_{B_i}/\tau\right) + \sum_{j=1}^{N} \exp\left(v_{U_j} \cdot v_{B_i}/\tau\right)} \tag{18}$$

Inter-component contrastive learning: The framework integrates cross-space contrastive learning to articulate relationships between decomposed components, thereby facilitating more refined decomposition. The clean image **B**, transmission map **T**, and ambient light **A** exhibit distinct physical characteristics: **T** encodes scene depth and attenuation information, **A** captures global illumination patterns, and **B** preserves intrinsic scene structure and texture.

To model these inter-component relationships, we propose a cross-space contrastive loss that regulates the integration among components. The loss function is formulated as:

$$\mathcal{L}_{\text{spaceCon}} = -\frac{1}{N_X}\sum_{k=1}^{N_X}\sum_{i=1}^{N_X}\frac{\exp\left(f_{X_i}\cdot f_{X_k}/\tau\right)}{\sum_{j=1}^{N_Y}\exp\left(f_{X_i}\cdot f_{Y_j}/\tau\right)} - \frac{1}{N_Y}\sum_{m=1}^{N_Y}\sum_{j=1}^{N_Y}\frac{\exp\left(f_{Y_j}\cdot f_{Y_m}/\tau\right)}{\sum_{i=1}^{N_X}\exp\left(f_{Y_j}\cdot f_{X_i}/\tau\right)} \quad (19)$$

where $X, Y \in \{B, T, A\}$ and N_X and N_Y denote the cardinalities of positive and negative sample sets, respectively. The feature vectors f_{X_i} and f_{Y_i} are extracted using discriminator encoder E_{dis}, which captures component-specific characteristics. The temperature parameter τ modulates the sensitivity of the similarity metric, following the principles established in [37].

This cross-space contrastive mechanism maximizes intra-component feature similarity while promoting inter-component feature distinctiveness, thereby ensuring effective separation of the physical components in the learned feature space. The selection of positive and negative samples is guided by non-local patch similarity to ensure meaningful contrastive relationships.

The complete contrastive learning objective integrates component-specific constraints:

$$\mathcal{L}_{CL} = \mathcal{L}_{factorCon} + \mathcal{L}_{spaceCon} + \mathcal{L}_{detailCon} \quad (20)$$

where each term enforces specific aspects of the physical decomposition model while maintaining feature distinctiveness through carefully selected positive and negative samples.

Sampling strategy: The effectiveness of contrastive learning relies on clear distinctions between positive and negative samples while maintaining intra-class similarity. Unlike conventional CL approaches that employ data augmentation [37], we propose a physics-guided sampling strategy based on local patch similarity:

$$Distance(p_i, p_{i_R}) = \left\| p_i - p_{i_\Omega} \right\|^2 \quad (21)$$

where p_i denotes the query patch and p_{i_Ω} represents candidate patches within support set Ω. Patches with the lowest distance values form positive pairs, while those with the highest values constitute negative pairs.

For detail preservation ($\mathcal{L}_{detailCon}$), positive pairs are selected from spatially corresponding regions between enhanced background **B** and input image **U**, which maintain consistent structural details. The factor consistency objective ($\mathcal{L}_{factorCon}$) forms positive pairs between estimated components $(\mathbf{T}, \mathbf{A})$ and their high-quality references $(\mathbf{T}', \mathbf{A}')$, considering depth-dependent attenuation for transmission and illumination consistency for ambient light. For component distinctiveness ($\mathcal{L}_{spaceCon}$), positive pairs are selected within each component space, while negative pairs are sampled across different components to maximize feature distinctiveness.

This physics-guided sampling strategy ensures both reliable feature learning through intra-component consistency and effective decomposition through inter-component distinctiveness, while maintaining computational efficiency through local patch-based selection.

3.3.2. Information Formulation Supervised Function

The decomposition framework leverages physical priors to guide the separation of an underwater image $\mathbf{U}$ into its constituent components: background $\mathbf{B}$, transmission map $\mathbf{T}$, and ambient light $\mathbf{A}$. These physics-based constraints ensure both the stability and the physical validity of the decomposition process. The estimated components $(\mathbf{T}, \mathbf{A})$ provide gradient supervision through an L1-based reconstruction loss:

$$\mathcal{L}_{IFM} = |\mathbf{U} - (\mathbf{B} \cdot \mathbf{T} + (1 - \mathbf{T}) \cdot \mathbf{A})|_1 \tag{22}$$

The L1 norm is specifically chosen over alternative metrics (e.g., L2) for its superior characteristics in preserving fine-grained details and robustness to outliers—properties crucial for high-fidelity image restoration. This constraint enforces the decomposed components to adhere to the physical degradation process defined by the IFM, guaranteeing physically consistent outputs from the decoders $(\mathbf{D_b}, \mathbf{D_t}, \mathbf{D_a})$.

3.3.3. Adversarial Learning Function

The adversarial learning framework optimizes decoder $\mathbf{D_b}$ to synthesize clean background images $\mathbf{B}$ that accurately emulate the characteristics of natural underwater scenes. This is achieved through an adversarial discriminator $\mathbf{Dis}$ that learns to differentiate between authentic and synthesized backgrounds. The corresponding adversarial loss is formulated as:

$$\mathcal{L}_{adv}(\mathbf{D_b}, \mathbf{Dis}) = \mathbb{E}_{\mathbf{B} \sim P_{bg}(\mathbf{B})}[\log \mathbf{Dis}(\mathbf{B})] + \mathbb{E}_{\mathbf{U} \sim P_{uw}(\mathbf{U})}[\log(1 - \mathbf{Dis}(\mathbf{D_b}(\mathbf{U})))] \tag{23}$$

where $P_{bg}(\mathbf{B})$ characterizes the probability distribution of authentic background images in the target domain and $P_{uw}(\mathbf{U})$ represents the distribution of underwater observations in the source domain. Through iterative optimization, the decoder $\mathbf{D_b}$ learns to minimize this objective, thereby generating perceptually realistic and physically plausible background images.

For transmission and ambient light components, we extend this adversarial learning strategy with specialized formulations $\mathcal{L}_{adv}(\mathbf{D_t}, \mathbf{Dis})$ and $\mathcal{L}_{adv}(\mathbf{D_a}, \mathbf{Dis})$, respectively. Specifically, the transmission decoder $\mathbf{D_t}$ employs local attention mechanisms to capture spatially varying degradation patterns, as the transmission map exhibits pixel-wise variations in projection rates. Conversely, the ambient light decoder $\mathbf{D_a}$ adopts a global representation approach, reflecting the scene-level nature of ambient illumination. For initialization, the transmission map $\mathbf{T}$ utilizes a second-order Laplacian filter to extract edge-based degradation cues, while the ambient light $\mathbf{A}$ leverages the dark channel prior.

Hence, the whole adversarial learning framework is formulated as:

$$\mathcal{L}_{adv} = \mathcal{L}_{adv}(\mathbf{D_b}, \mathbf{Dis}) + \mathcal{L}_{adv}(\mathbf{D_t}, \mathbf{Dis}) + \mathcal{L}_{adv}(\mathbf{D_a}, \mathbf{Dis}) \tag{24}$$

4. Experiments

4.1. Implementation Details

Network structure: The feature extraction backbone for $\mathcal{L}_{crossSpace}$ is specifically engineered to optimize inter-component feature discrimination through a hierarchical representation learning scheme. The framework implements a classification-based architecture similar to PatchGAN [45], incorporating a discriminator–generator structure (illustrated in Figure 1b). This design enables $\mathcal{L}_{crossSpace}$ to effectively capture discriminative semantic features necessary for component classification.

For $\mathcal{L}_{detailCon}$ and $\mathcal{L}_{factorCon}$ objectives, the feature extraction mechanism emphasizes local patch-level consistency between decomposed components and their corresponding

observations. The framework employs ResNet [46] as the backbone feature extractor, followed by a two-layer multilayer perceptron (MLP) with 256 hidden units for feature embedding (detailed in Figure 1c). The sampling parameters are empirically set as $N = 256$ for negative samples, $N_X = 8$ for positive samples, and $N_Y = 256$ for cross-space negative samples.

Training details: Our method employs a three-phase training protocol to ensure stable convergence and accurate component decomposition: **(1) background initialization:** the background decoder is optimized using $\mathcal{L}_{detailCon}$, $\mathcal{L}_{adv}$, and $\mathcal{L}_{IFM}$ for 10 epochs while keeping other components fixed; **(2) degradation component learning:** the transmission and ambient light decoders are jointly trained for 20 epochs using $\mathcal{L}_{factorCon}$ and auxiliary losses, with the background decoder parameters frozen; **(3) joint optimization:** all components are collectively refined for 170 epochs using $\mathcal{L}_{spaceCon}$ and the complete loss function set, enabling comprehensive feature interaction. The optimization framework adopts the momentum contrast mechanism from MOCO [40], with momentum coefficient 0.99 and temperature parameter 0.77. The implementation utilizes the PyTorch 1.9.0 framework [47] on NVIDIA RTX 4090 GPUs. The Adam optimizer [48] is employed with initial learning rate 1×10^{-4} , which is decreased by a factor of 0.1 after 100 epochs. Training proceeds for 200 epochs with mini-batch size 4 and input resolution 256×256. The loss weights are configured as $\lambda_{adv} = 1$, $\lambda_{IFM} = 0.5$, and $\lambda_{CL} = 0.1$.

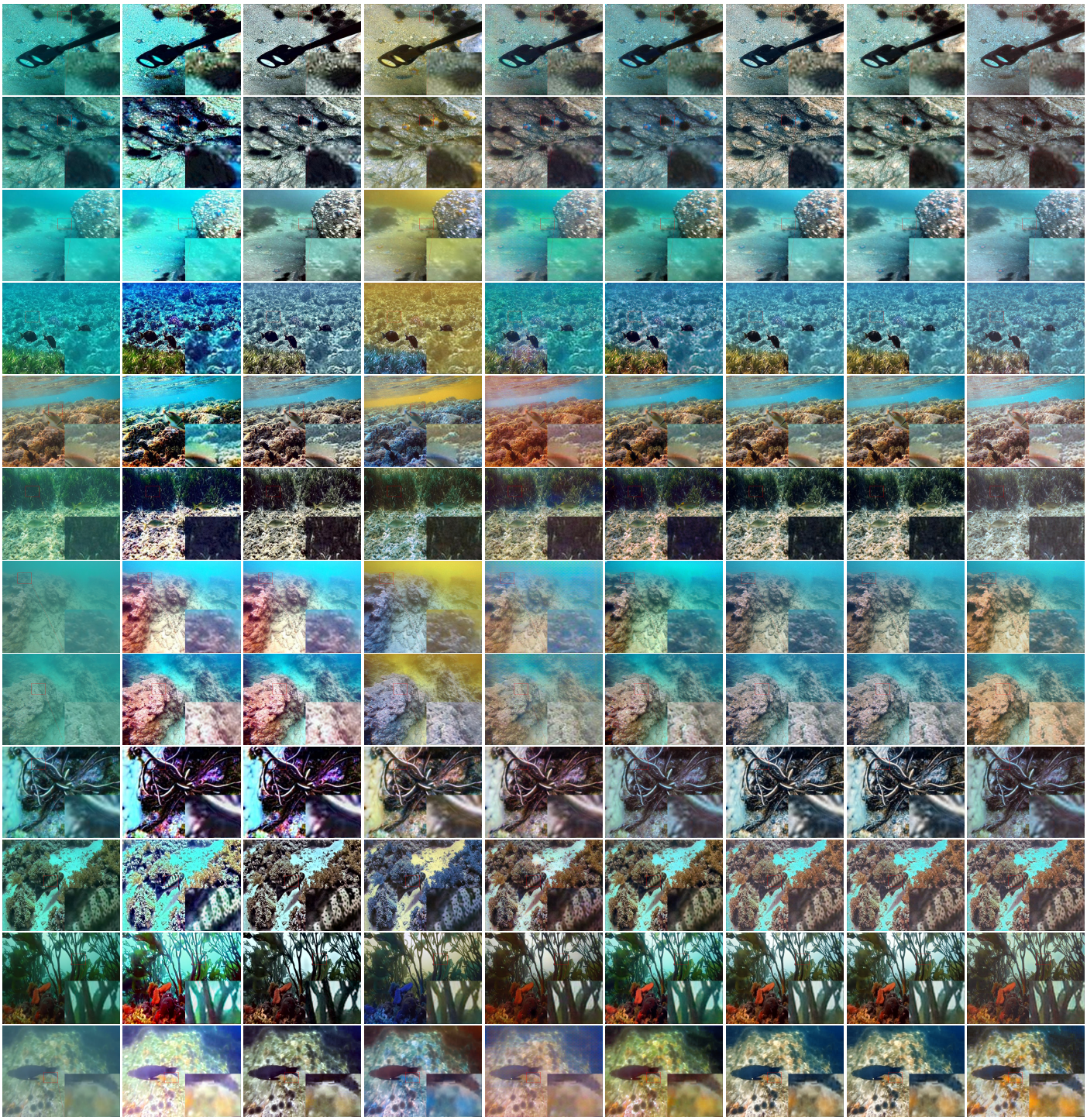
4.2. Datasets

The experimental dataset comprises 2000 unpaired unlabeled samples and 2000 paired images with referenced clean labels, ensuring comprehensive coverage of diverse underwater scenarios. The unlabeled images are randomly sampled from the EUVP [1]. Additionally, 2000 labeled images are selected from the LSUI dataset [49], which ensures diversity in real underwater scenes, water types, lighting conditions, and target categories. Pseudo labels (i.e., estimated T' and A') are generated for training by processing all labeled and unlabeled high-quality underwater images using widely adopted underwater transmission and ambient light estimation methods [50]; the details can be seen on the right side of Figure 1a. The testing set consists of four commonly used datasets, UIEB [51], EUVP [1], RUIE [52], and SUIM [17], frequently employed in underwater image enhancement (UIE) research.

4.3. Quantitative and Quantitative Comparison

The performance evaluation of the proposed framework encompasses both quantitative metrics and qualitative visual evaluation, benchmarking against state-of-the-art underwater image enhancement methods. The comparative analysis includes two traditional methods (GDCP [13] and MMLE [53]) and five learning-based approaches (WaterNet [51], FUINE [1], CWR [19], SEMUIR [18], and HUPE [54]). For fair comparison, all learning-based methods are retrained using the proposed training dataset.

Visual quality assessment: The qualitative evaluation examines enhancement performance across four diverse datasets, encompassing various underwater scenes and degradation patterns. As illustrated in Figure 2, existing methods often exhibit inconsistent enhancement results, resulting in over-smoothing artifacts and color distortions at both global and local scales. For instance, WaterNet [51] introduces artifacts along high-frequency edges, while FUINE [1] and CWR [19] demonstrate limitations in color fidelity and detail preservation under challenging conditions.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**)

Figure 2. Visual comparison of underwater image enhancement results across multiple datasets. From top to bottom: RUIE (blue-tinted underwater scenes), SUIM (shallow underwater images), EVUP (varied visibility conditions), and UIEB (diverse underwater environments). Note improved visibility details of underwater objects and more natural color rendition in our results. (**a**) Original. (**b**) GDCP [13]. (**c**) MMLE [53]. (**d**) Water [51]. (**e**) FUINE [1]. (**f**) CWR [19]. (**g**) SEMI [18]. (**h**) HUPE [54]. (**i**) CFD-Net.

The challenge of maintaining color fidelity while correcting hue distortions is particularly significant in unpaired learning frameworks. SEMUIR [18] addresses this through using estimated illumination maps as auxiliary inputs, showing improved color correction particularly for foreground objects (e.g., red and blue fish in the 11th and 12th rows of

Figure 2). The proposed framework achieves superior color restoration by integrating physical underwater image formation principles into the network architecture, thereby resulting in more natural and vibrant color distributions compared to methods relying solely on unpaired training data.

Quantitative performance analysis: The quantitative evaluation employs four widely used non-reference metrics, each targeting different dimensions of underwater image degradation. UIQM [55] evaluates three key aspects, colorfulness, sharpness, and contrast, addressing common underwater image issues. UCIQE [56] quantifies color casts and low contrast, which are typical challenges in underwater imaging. MUSIQ [57] uses a multi-scale transformer architecture to assess structural fidelity and texture preservation across frequency bands. And NIQE [58] models natural scene statistics (NSSs) in the wavelet domain to detect artificial artifacts, with lower scores indicating higher naturalness. Although these metrics are heuristic and domain-specific, they cover important facets of underwater image quality: UIQM/UCIQE focus on the unique challenges of color distortions and turbidity, while MUSIQ/NIQE contribute to texture preservation and naturalness. As evidenced in Table 1, our method achieves optimal balance across these orthogonal dimensions.

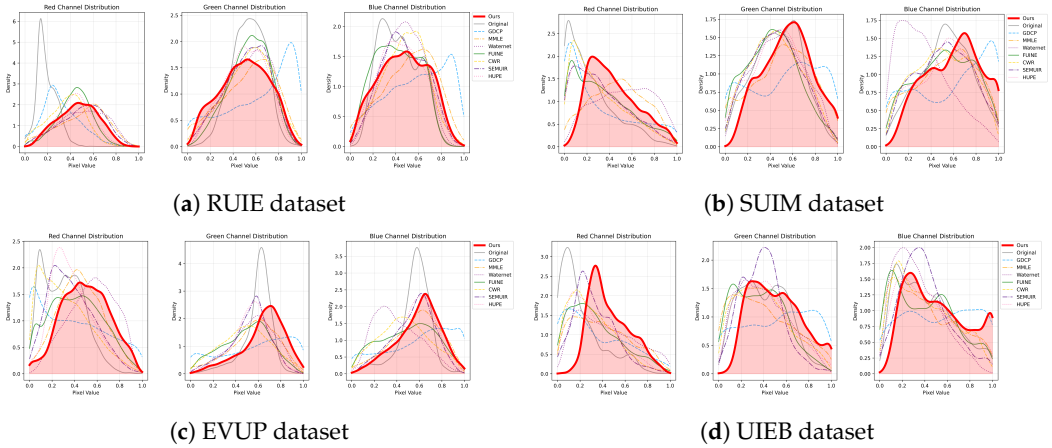
Our CFD-Net consistently outperforms existing methods in UIQM and NIQE metrics across all datasets while maintaining competitive performance in UCIQE and MUSIQ. While GDCP [13] demonstrates superior performance in UCIQE and SEMUIR [18] achieves high scores in MUSIQ, our method exhibits balanced performance across all metrics, effectively improving both contrast enhancement and semantic feature preservation. This comprehensive enhancement capability is attributed to the synergistic integration of decomposed generative branches and multi-scale contrastive learning mechanisms.

Table 1. Comprehensive evaluation of compared methods (quality metrics and computational efficiency) on benchmarks in terms of UIQM [55], UCIQE [56], MUSIQ [57], and NIQE [58] across RUIE [52], SUIM [17], EVUP [1], and UIEB [51] datasets. The ↑ indicates higher values are better, while ↓ indicates lower values are better. Best results are in **bold** and second-best results are underlined.

Method	UIQM ↑				UCIQE ↑				MUSIQ ↑				NIQE ↓				Efficiency	
	RUIE	SUIM	EVUP	UIEB	RUIE	SUIM	EVUP	UIEB	RUIE	SUIM	EVUP	UIEB	RUIE	SUIM	EVUP	UIEB	FLOPs (G)	Time (ms)
Original	2.294	2.027	2.254	1.677	0.523	0.597	0.543	0.533	33.874	60.855	43.678	41.697	5.062	3.957	4.976	6.905	-	-
GDCP [13]	2.738	1.970	2.337	1.899	**0.608**	**0.678**	**0.638**	**0.624**	32.671	58.897	40.886	**51.002**	4.738	3.885	4.697	5.190	-	173.8
MMLE [53]	2.871	2.137	2.337	1.953	0.567	0.612	**0.638**	0.580	**36.510**	**62.704**	40.886	40.345	4.859	4.045	4.697	4.845	-	91.7
WaterNet [51]	3.168	2.644	3.077	2.317	0.568	0.607	0.597	0.581	30.289	60.280	42.402	40.006	4.755	4.007	4.420	5.702	571.8	40.6
FUINE [1]	3.145	2.557	2.944	**2.867**	0.538	0.610	0.572	0.552	28.726	60.158	38.110	46.827	4.906	3.611	5.106	5.299	81.91	**2.9**
CWR [19]	3.154	2.847	3.008	2.459	0.583	0.637	0.618	0.607	25.310	58.915	37.854	30.131	4.730	4.058	4.744	5.338	338.9	20.8
SEMUIR [18]	3.063	2.502	2.957	2.164	0.554	0.636	0.599	0.570	32.446	62.272	**47.882**	42.460	4.633	3.439	4.352	5.697	105.6	43.4
HUPE [54]	3.000	2.481	2.779	2.198	0.550	0.637	0.602	0.582	29.766	54.57	38.817	35.559	4.598	4.148	5.136	7.771	87.5	50.2
Ours	**3.227**	**3.117**	**3.116**	2.793	0.558	0.620	0.600	0.569	34.398	61.348	47.175	39.412	**3.922**	**2.863**	**3.639**	**3.938**	147.5	19.3

FLOPs: float operations; time: 4090 GPU latency; "-": non-deep methods.

Figure 3a–d present the channel-wise color distribution analysis of compared methods across different datasets. The results demonstrate that the proposed method consistently achieves more balanced and natural color distributions, particularly evident in the red channel where underwater degradation is most severe. Additionally, Figure 4 provides statistical analysis of color characteristics across all methods in the EVUP dataset, utilizing mean, standard deviation, median, and color range metrics to comprehensively evaluate enhancement quality. Our method achieves balanced mean values (μ = 0.497, 0.642, 0.590) with stable standard deviations (σ = 0.207, 0.195, 0.196) across RGB channels, showing a notable 66.1% improvement in red channel mean intensity compared to original images (μ = 0.299, 0.592, 0.555) while maintaining more consistent color distributions than competing methods such as GDCP (σ = 0.296, 0.301, 0.292). These quantitative results validate the effectiveness of our method in addressing the characteristic color distortion in underwater imagery while preserving natural color appearances.



(**a**) RUIE dataset (**b**) SUIM dataset

(**c**) EVUP dataset (**d**) UIEB dataset

Figure 3. Channel-wise color distribution comparison of different underwater image enhancement methods on multiple datasets. Density curves illustrate distribution of pixel values in red (left), green (middle), and blue (right) channels. Different line styles and colors represent different enhancement methods.

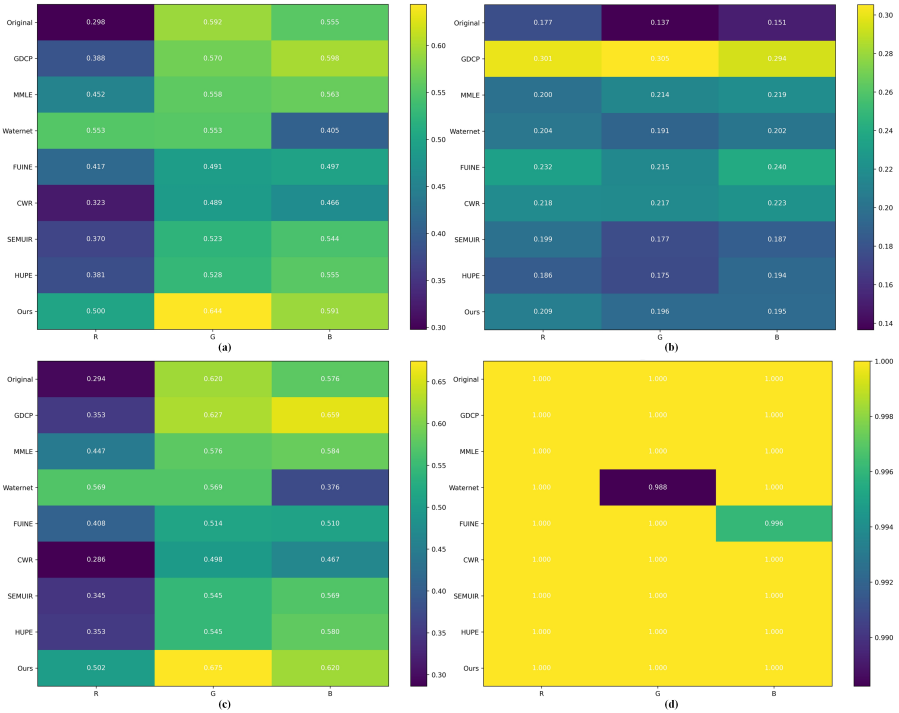


Figure 4. Statistical analysis of color characteristics for different enhancement methods. Heatmaps display (**a**) mean values, (**b**) standard deviation, (**c**) median values, and (**d**) color range for each RGB channel. Values are normalized to range [0, 1].

Comprehensive performance evaluation: To evaluate computational efficiency, we compare the FLOPs and inference time of various methods. All experiments are performed on an NVIDIA RTX 4090 GPU using 512 × 512 input images and a batch size of 1. Inference time is averaged over 100 forward passes to ensure consistency.

As shown in Table 1, our method achieves an inference time of 19.3 ms, making it the 2nd fastest among all methods. While FUINE achieves the fastest time (2.9 ms), it sacrifices quality, particularly in NIQE (5.106 vs. 3.922). Compared to CWR (20.8 ms) and WaterNet (40.6 ms), our model offers superior speed without compromising enhancement performance. Regarding computational cost, our method operates at 147.5 G FLOPs, significantly lower than WaterNet (571.8 G) and CWR (338.9 G). Although our approach has a higher FLOP count than FUINE (81.91 G) and HUPE (87.5 G), it strikes a favorable

balance between efficiency and quality. This makes it particularly suitable for real-time underwater vision applications, especially in resource-constrained environments.

4.4. Ablation Study

To comprehensively evaluate the contributions of different components, we conduct ablation experiments on the UIEB dataset using four complementary quality metrics. We categorize the losses into three groups: (1) physics-consistency losses (L_{IFM}), (2) feature contrastive losses ($L_{detailCon}$, $L_{spaceCon}$, $L_{factorCon}$), and (3) adversarial loss (L_{adv}). The relative performance drop, denoted as Δ, is computed as follows:

$$\Delta = \frac{M_{\text{full}} - M_{\text{ablated}}}{M_{\text{full}}} \times 100\% \tag{25}$$

where M_{full} denotes the metric value of the full model and M_{ablated} represents the value when specific components are removed.

As shown in Table 2, the physics-only baseline achieves moderate performance (UIQM = 1.667) and provides moderate enhancement quality, confirming the necessity of physical modeling. However, removing contrastive learning drastically reduces performance (e.g., UIQM $\downarrow$ 40.3%), indicating that feature-level constraints are critical for enhancement quality. Among the contrastive losses, $L_{detailCon}$ contributes the most to texture enhancement, improving UIQM by 50.3% (1.667 $\rightarrow$ 2.506), $L_{spaceCon}$ mainly enhances color balance, increasing UCIQE by 7.2%, and $L_{factorCon}$ improves perceptual quality and denoising, reducing NIQE from 4.107 to 3.980. The adversarial loss alone (L_{adv}) achieves reasonable enhancement, but its full effect emerges when combined with physical and contrastive constraints. Notably, the full model achieves the best balance between all metrics, demonstrating the complementary interactions between loss categories. These results confirm that feature contrastive learning is the most crucial factor for underwater image enhancement, while adversarial learning provides additional refinement, particularly in perceptual quality (MUSIQ $\uparrow$ 41.412 vs. 40.213).

Table 2. Component-wise ablation analysis on UIEB dataset (updated). The $\uparrow$ indicates higher values are better, while $\downarrow$ indicates lower values are better. Best results are in **bold**.

Category	Components	UIQM ↑	Δ UIQM	UCIQE ↑	Δ UCIQE	MUSIQ ↑	Δ MUSIQ	NIQE ↓	Δ NIQE
Physical only	L_{IFM}	1.667	−40.3%	0.572	−1.2%	39.476	−4.7%	4.107	−9.9%
Feature contrastive	$+L_{detailCon}$	2.506	−10.3%	0.522	−8.3%	38.912	−6.0%	4.246	−13.6%
	$+L_{spaceCon}$	2.301	−17.6%	0.551	−3.2%	39.102	−5.6%	4.102	−9.7%
	$+L_{factorCon}$	2.410	−13.7%	0.563	−1.1%	39.875	−3.7%	3.980	−6.5%
Adversarial	$+L_{adv}$	2.155	−22.8%	0.548	−3.7%	40.213	−2.9%	3.890	−4.1%
Full	All	**2.793**	-	**0.569**	-	**41.412**	-	**3.738**	-

Δ: Relative drop versus full model.

Figure 5 illustrates the impact of individual components on visual results. Initially, the L_{IFM} function restores the contour information but results in some color and contrast loss. The addition of $L_{detailCon}$ enhances fine-grained details, notably the printed text. When $L_{spaceCon}$ and $L_{factorCon}$ are introduced, color vibrancy improves, but some noise artifacts are also present. The adversarial loss (L_{adv}) enhances color fidelity and texture preservation, especially in low-light scenarios. The integration of all components yields the best overall performance in terms of clarity, detail retention, and contrast.

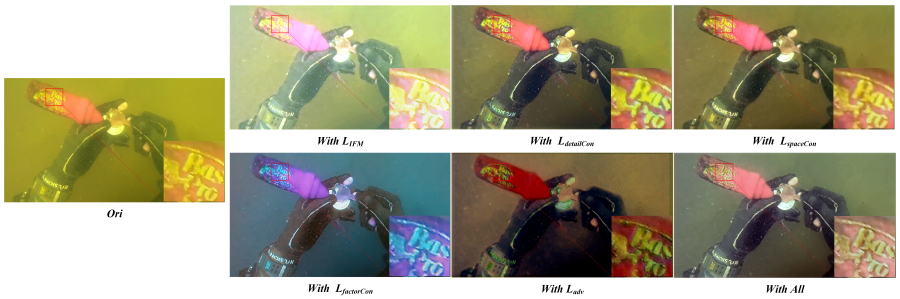


Figure 5. Visual comparison of ablation study results. From left to right: input image, results from ablation with IFM (L_{IFM}), with detail preservation ($L_{detailCon}$), with cross-space constraint ($L_{spaceCon}$), with factor constraint ($L_{factorCon}$), with adversarial loss (L_{adv}), and with full model.

4.5. Evaluation on Other Applications

To demonstrate the practical utility of the proposed framework, we evaluate its performance on downstream high-level computer vision tasks. Specifically, we assess the impact of image enhancement on underwater semantic segmentation performance using the SUIM dataset [17]. The evaluation compares the segmentation accuracy using original underwater images, images enhanced by competing methods, and those processed by our framework.

The segmentation results are illustrated in Figure 6, with quantitative evaluation metrics presented in Table 3. The proposed framework achieves superior performance across all evaluation metrics, attaining a PA of 0.830, mPA of 0.319, and mIoU of 0.284. These results represent significant improvements compared to both traditional methods (GDCP [13], MMLE [53]) and learning-based approaches (WaterNet [51], FUINE [1], CWR [19], SEMUIR [18], and HUPE [54]). The enhanced performance on semantic segmentation demonstrates the capability of our framework to recover discriminative features that benefit high-level computer vision tasks.

Table 3. Pixel Accuracy (PA), Mean Pixel Accuracy (mPA), and Mean Intersection over Union (mIoU) comparison on SUIM dataset. The ↑ indicates higher values are better. Best results are in **bold** and second-best results are underlined.

	GDCP	MMLE	WaterNet	FUNIE	CWR	SEMUIR	HUPE	Ours
PA ↑	0.812	0.805	0.655	0.806	0.800	0.816	<u>0.826</u>	**0.830**
MPA ↑	0.301	0.304	0.261	0.297	0.300	0.304	<u>0.312</u>	**0.319**
mIoU ↑	0.267	0.268	0.200	0.264	0.268	<u>0.273</u>	**0.284**	**0.284**

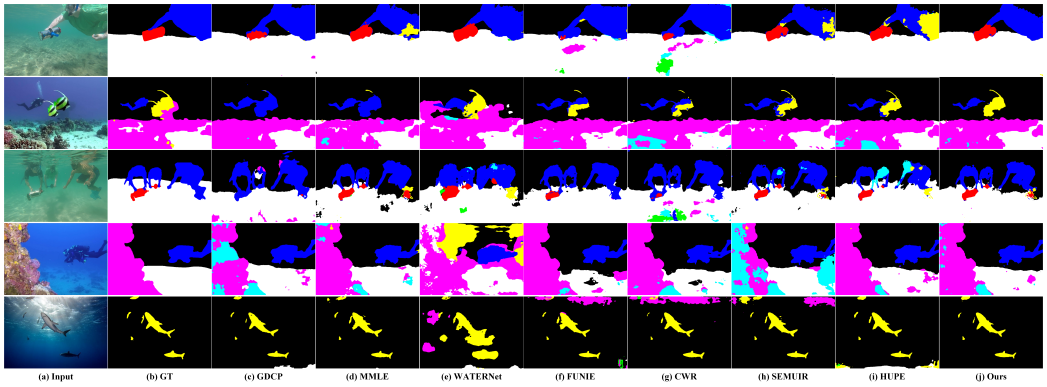


Figure 6. Evaluation of underwater semantic segmentation on SUIM dataset. The 1st and 2nd columns denote the input underwater images and their corresponding label masks. The remaining columns show the segmentation results of different methods.

5. Conclusions

Our CFD-Net integrates decomposed feature representation with physical priors, which combines underwater image enhancement techniques with unsupervised learning algorithms, effectively addressing the challenges posed by degradation processes and refined feature representations. Through extensive experiments, we demonstrate that CFD-Net consistently outperforms state-of-the-art methods in both image enhancement and high-level downstream tasks, such as semantic segmentation. The integration of physical priors and innovative learning mechanisms allows our framework to recover vibrant, natural colors and significantly improve segmentation accuracy. Future work will explore the extension of this approach to multi-frame sequences and videos, where temporal information and multi-view perspectives could provide additional cues for enhanced underwater image restoration.

Author Contributions: Conceptualization, F.L.; Data curation, L.W. (Li Wan); Methodology, F.L.; Supervision, J.Z.; Validation, L.W. (Li Wan); Writing—original draft, F.L.; Writing—review and editing, L.W. (Lu Wang) and Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was sponsored by the National Natural Science Foundation of China under Grant (62301397, 62402370), in part by the Aero Science Foundation of China under Grant 20240058081001 and the Fundamental Research Funds for the Central Universities under Grant XJSJ23034.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Islam, M.J.; Xia, Y.; Sattar, J. Fast underwater image enhancement for improved visual perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [CrossRef]
2. Song, Y.; Nakath, D.; She, M.; Köser, K. Optical imaging and image restoration techniques for deep ocean mapping: A comprehensive survey. *PFG-Photogramm. Remote Sens. Geoinf. Sci.* **2022**, *90*, 243–267. [CrossRef]
3. Grimaldi, M.; Nakath, D.; She, M.; Köser, K. Investigation of the Challenges of Underwater-Visual-Monocular-SLAM. *arXiv* **2023**, arXiv:2306.08738. [CrossRef]
4. Long, H.; Shen, L.; Wang, Z.; Chen, J. Underwater Forward-Looking Sonar Images Target Detection via Speckle Reduction and Scene Prior. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13. [CrossRef]
5. Cao, X.; Ren, L.; Sun, C. Dynamic target tracking control of autonomous underwater vehicle based on trajectory prediction. *IEEE Trans. Cybern.* **2022**, *53*, 1968–1981. [CrossRef]
6. Wang, Y.; Guo, J.; He, W.; Gao, H.; Yue, H.; Zhang, Z.; Li, C. Is Underwater Image Enhancement All Object Detectors Need? *IEEE J. Ocean. Eng.* **2024**, *49*, 606–621. [CrossRef]
7. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13619–13627.
8. Song, Y.; He, Z.; Qian, H.; Du, X. Vision transformers for single image dehazing. *IEEE Trans. Image Process.* **2023**, *32*, 1927–1941. [CrossRef]
9. Wu, Y.; Pan, C.; Wang, G.; Yang, Y.; Wei, J.; Li, C.; Shen, H.T. Learning Semantic-Aware Knowledge Guidance for Low-Light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1662–1671.
10. Zheng, S.; Gupta, G. Semantic-guided zero-shot learning for low-light image/video enhancement. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 581–590.
11. Chen, L.; Jiang, Z.; Tong, L.; Liu, Z.; Zhao, A.; Zhang, Q.; Dong, J.; Zhou, H. Perceptual underwater image enhancement with deep learning and physical priors. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 3078–3092. [CrossRef]
12. Li, C.; Anwar, S.; Hou, J.; Cong, R.; Guo, C.; Ren, W. Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE Trans. Image Process.* **2021**, *30*, 4985–5000. [CrossRef]
13. Peng, Y.T.; Cao, K.; Cosman, P.C. Generalization of the dark channel prior for single image restoration. *IEEE Trans. Image Process.* **2018**, *27*, 2856–2868. [CrossRef]

14. Rahman, Z.u.; Jobson, D.J.; Woodell, G.A. Multi-scale retinex for color image enhancement. In Proceedings of the 3rd IEEE International Conference on Image Processing, Lausanne, Switzerland, 16–19 September 1996; Volume 3, pp. 1003–1006.
15. Ren, X.; Yang, W.; Cheng, W.H.; Liu, J. LR3M: Robust low-light enhancement via low-rank regularized retinex model. *IEEE Trans. Image Process.* **2020**, *29*, 5862–5876. [CrossRef]
16. Fu, C.; Liu, R.; Fan, X.; Chen, P.; Fu, H.; Yuan, W.; Zhu, M.; Luo, Z. Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing* **2023**, *517*, 243–256. [CrossRef]
17. Islam, M.J.; Edge, C.; Xiao, Y.; Luo, P.; Mehtaz, M.; Morse, C.; Enan, S.S.; Sattar, J. Semantic segmentation of underwater imagery: Dataset and benchmark. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; pp. 1769–1776.
18. Huang, S.; Wang, K.; Liu, H.; Chen, J.; Li, Y. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18145–18155.
19. Han, J.; Shoeiby, M.; Malthus, T.; Botha, E.; Anstee, J.; Anwar, S.; Wei, R.; Armin, M.A.; Li, H.; Petersson, L. Underwater image restoration via contrastive learning and a real-world dataset. *Remote Sens.* **2022**, *14*, 4297. [CrossRef]
20. Liu, X.; Gao, Z.; Chen, B.M. IPMGAN: Integrating physical model and generative adversarial network for underwater image enhancement. *Neurocomputing* **2021**, *453*, 538–551. [CrossRef]
21. Qi, H.; Dong, X. Physics-aware semi-supervised underwater image enhancement. *arXiv* **2023**, arXiv:2307.11470.
22. Zhou, Y.; Yan, K.; Li, X. Underwater image enhancement via physical-feedback adversarial transfer learning. *IEEE J. Ocean. Eng.* **2021**, *47*, 76–87. [CrossRef]
23. Liu, Y.C.; Chan, W.H.; Chen, Y.Q. Automatic white balance for digital still camera. *IEEE Trans. Consum. Electron.* **1995**, *41*, 460–466.
24. Pizer, S.M. Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group. In Proceedings of the First Conference on Visualization in Biomedical Computing, Atlanta, GA, USA, 22–25 May 1990; Volume 337, p. 2.
25. Ancuti, C.; Ancuti, C.O.; Haber, T.; Bekaert, P. Enhancing underwater images and videos by fusion. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 81–88.
26. Hitam, M.S.; Awalludin, E.A.; Yussof, W.N.J.H.W.; Bachok, Z. Mixture contrast limited adaptive histogram equalization for underwater image enhancement. In Proceedings of the 2013 International Conference on Computer Applications Technology (ICCAT), Sousse, Tunisia, 20–22 January 2013; pp. 1–5.
27. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353.
28. Drews, P.; Nascimento, E.; Moraes, F.; Botelho, S.; Campos, M. Transmission estimation in underwater single images. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 2–8 December 2013; pp. 825–830.
29. Galdran, A.; Pardo, D.; Picón, A.; Alvarez-Gila, A. Automatic red-channel underwater image restoration. *J. Vis. Commun. Image Represent.* **2015**, *26*, 132–145. [CrossRef]
30. Akkaynak, D.; Treibitz, T. Sea-thru: A method for removing water from underwater images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1682–1691.
31. Liang, Z.; Zhang, W.; Ruan, R.; Zhuang, P.; Xie, X.; Li, C. Underwater image quality improvement via color, detail, and contrast restoration. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 1726–1742. [CrossRef]
32. Wang, Y.; Zhang, J.; Cao, Y.; Wang, Z. A deep CNN method for underwater image enhancement. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 1382–1386.
33. Li, C.; Anwar, S.; Porikli, F. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognit.* **2020**, *98*, 107038. [CrossRef]
34. Li, J.; Skinner, K.A.; Eustice, R.M.; Johnson-Roberson, M. WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robot. Autom. Lett.* **2017**, *3*, 387–394. [CrossRef]
35. Li, C.; Guo, J.; Guo, C. Emerging from water: Underwater image color correction based on weakly supervised color transfer. *IEEE Signal Process. Lett.* **2018**, *25*, 323–327. [CrossRef]
36. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
37. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
38. Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; Hinton, G.E. Big self-supervised models are strong semi-supervised learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22243–22255.
39. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.

40. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
41. Wu, H.; Qu, Y.; Lin, S.; Zhou, J.; Qiao, R.; Zhang, Z.; Xie, Y.; Ma, L. Contrastive learning for compact single image dehazing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10551–10560.
42. Liu, R.; Jiang, Z.; Yang, S.; Fan, X. Twin adversarial contrastive learning for underwater image enhancement and beyond. *IEEE Trans. Image Process.* **2022**, *31*, 4922–4936. [CrossRef]
43. Yang, Y.; Wang, C.; Guo, X.; Tao, D. Robust Unpaired Image Dehazing via Density and Depth Decomposition. *Int. J. Comput. Vis.* **2024**, *132*, 1557–1577. [CrossRef]
44. Chang, Y.; Guo, Y.; Ye, Y.; Yu, C.; Zhu, L.; Zhao, X.; Yan, L.; Tian, Y. Unsupervised Deraining: Where Asymmetric Contrastive Learning Meets Self-similarity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2638–2657. [CrossRef]
45. Demir, U.; Unal, G. Patch-based image inpainting with generative adversarial networks. *arXiv* **2018**, arXiv:1803.07422.
46. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part II 14; Springer: Cham, Switzerland, 2016; pp. 694–711.
47. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
48. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
49. Peng, L.; Zhu, C.; Bian, L. U-shape transformer for underwater image enhancement. *IEEE Trans. Image Process.* **2023**, *32*, 3066–3079. [CrossRef] [PubMed]
50. Ehsan, S.M.; Imran, M.; Ullah, A.; Elbasi, E. A single image dehazing technique using the dual transmission maps strategy and gradient-domain guided image filtering. *IEEE Access* **2021**, *9*, 89055–89063. [CrossRef]
51. Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; Tao, D. An underwater image enhancement benchmark dataset and beyond. *IEEE Trans. Image Process.* **2019**, *29*, 4376–4389. [CrossRef]
52. Liu, R.; Fan, X.; Zhu, M.; Hou, M.; Luo, Z. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4861–4875. [CrossRef]
53. Zhang, W.; Zhuang, P.; Sun, H.H.; Li, G.; Kwong, S.; Li, C. Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement. *IEEE Trans. Image Process.* **2022**, *31*, 3997–4010. [CrossRef]
54. Zhang, Z.; Jiang, Z.; Ma, L.; Liu, J.; Fan, X.; Liu, R. HUPE: Heuristic Underwater Perceptual Enhancement with Semantic Collaborative Learning. *Int. J. Comput. Vis.* **2025**, 1–19. [CrossRef]
55. Panetta, K.; Gao, C.; Agaian, S. Human-visual-system-inspired underwater image quality measures. *IEEE J. Ocean. Eng.* **2015**, *41*, 541–551. [CrossRef]
56. Yang, M.; Sowmya, A. An underwater color image quality evaluation metric. *IEEE Trans. Image Process.* **2015**, *24*, 6062–6071. [CrossRef]
57. Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; Yang, F. Musiq: Multi-scale image quality transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 5148–5157.
58. Mittal, A.; Soundararajan, R.; Bovik, A. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209–212. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

Narrowband Radar Micromotion Targets Recognition Strategy Based on Graph Fusion Network Constructed by Cross-Modal Attention Mechanism

Yuanjie Zhang [1], Ting Gao [2], Hongtu Xie [1,*], Haozong Liu [1], Mengfan Ge [1], Bin Xu [1], Nannan Zhu [3] and Zheng Lu [4]

[1] School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China; zhangyj357@mail2.sysu.edu.cn (Y.Z.)
[2] Air Early Warning System Application Laboratory, Air Force Early Warning Academy, Wuhan 430019, China
[3] School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China
[4] Institute of Remote Sensing Satellite, China Academy of Space Technology, Beijing 100094, China
* Correspondence: xiehongtu@mail.sysu.edu.cn

Abstract: In the domain of micromotion target recognition, target characteristics can be extracted through either broadband or narrowband radar echoes. However, due to technical limitations and cost constraints in acquiring broadband radar waveform data, recognition can often only be performed through narrowband radar waveforms. To fully utilize the information embedded within narrowband radar waveforms, it is necessary to conduct in-depth research on multi-dimensional features of micromotion targets, including radar cross-sections (RCSs), time frequency (TF) images, and cadence velocity diagrams (CVDs). To address the limitations of existing identification methodologies in achieving accurate recognition with narrowband echoes, this paper proposes a graph fusion network based on a cross-modal attention mechanism, termed GF-AM Net. The network first adopts convolutional neural networks (CNNs) to extract unimodal features from RCSs, TF images, and CVDs independently. Subsequently, a cross-modal attention mechanism integrates these extracted features into a graph structure, achieving multi-level interactions among unimodal, bimodal, and trimodal features. Finally, the fused features are input into a classification module to accomplish narrowband radar micromotion target identification. Experimental results demonstrate that the proposed methodology successfully captures potential correlations between modal features by incorporating cross-modal multi-level information interactions across different processing stages, exhibiting exceptional accuracy and robustness in narrowband radar micromotion target identification tasks.

Keywords: narrowband radar; micromotion target recognition; graph fusion network; cross-modal attention mechanism; multimodal feature fusion

Academic Editor: Massimiliano Pieraccini

Received: 12 December 2024
Revised: 18 January 2025
Accepted: 11 February 2025
Published: 13 February 2025

Citation: Zhang, Y.; Gao, T.; Xie, H.; Liu, H.; Ge, M.; Xu, B.; Zhu, N.; Lu, Z. Narrowband Radar Micromotion Targets Recognition Strategy Based on Graph Fusion Network Constructed by Cross-Modal Attention Mechanism. *Remote Sens.* **2025**, *17*, 641. https://doi.org/10.3390/rs17040641

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid development of radar technology, significant progress has been made in the application of radars for the target detection and recognition, both in classifying stationary targets [1–3] and detecting high-speed moving targets [4–10]. A radar system works by transmitting electromagnetic waves and receiving their reflected signals, enabling the effective detection and tracking of the target's motion state. Traditional radar target recognition methods typically rely on geometric features and radar scattering characteristics of the targets. However, as the movement patterns of targets become more complex,

especially with the emergence of micromotion targets, traditional methods face significant challenges in terms of the recognition accuracy and robustness.

In complex adversarial environments, primary targets and interfering targets frequently co-exist, significantly challenging the identification capabilities of defense systems. Given their similar physical properties and morphological characteristics, conventional shape-based detection methodologies demonstrate limited efficacy in discriminating between these entities. Consequently, there exists an urgent imperative to explore and utilize alternative feature information of targets to enhance the accuracy and reliability of identification processes [9–13].

In addition to the translational motion, primary targets and interfering targets exhibit the micromotion in the form of the spin, precession, and nutation motion during the spatial flight [14]. Since primary targets typically employ attitude control techniques to enhance hit accuracy, their micromotion characteristics diverge significantly from those of interfering targets. The diverse forms of the micromotion yield disparate modulation effects on radar echo signals, thereby enabling the effective distinction between primary targets and interfering targets through the analysis of their respective micromotion characteristics in radar echo signals.

The identification of micromotion targets can be broadly divided into two main categories [15]. The first category involves extracting physical and statistical features from radar echo signals, which can be subsequently used for classification through template matching or statistical classifiers. For example, researchers extract physical features such as the fundamental frequency, peak-to-peak difference, and 3dB Doppler bandwidth from the radar cross-section (RCS), as well as statistical features such as the mean, variance, and sequential center distance [9,16]. Moreover, features can be extracted through image processing methods, such as converting echo signals into time frequency (TF) images using the short-time Fourier transform (STFT). From these images, physical features such as the Doppler period, frequency, and bandwidth, as well as statistical features such as the entropy and third-order moments, can be obtained [17,18]. In addition to the RCS and TF features, other features, such as the cadence velocity diagram (CVD) [19,20] and high-resolution range profile (HRRP) [14,21–23] are also used for the micromotion target identification.

However, traditional classification methods, such as the support vector machine (SVM) and k-nearest neighbor (KNN) methods, that rely on physical and statistical features, demonstrate the limited performance in complex scenarios. To address these limitations, the second category of methods focuses on designing deep learning networks to automatically learn features from the data for the micromotion target recognition [24–35]. For instance, Chen et al. proposed the RCS Net to facilitate the feature learning from the RCS data [25,26], while Bai et al. developed a network specifically designed for recognizing the TF data to classify micromotion targets [27]. Similarly, Feng et al. utilized the HRRP data for the target recognition, achieving the micromotion target classification [28]. However, while TF data effectively capture the Doppler information of individual scattering centers, they lack the ability to provide the detailed distance information. In contrast, the HRRP excels in representing the target distance information but cannot intuitively express the Doppler characteristics. Consequently, relying on a single data modality often fails to comprehensively describe the micromotion target's shape, structure, and motion. Although single-modality micromotion target recognition methods can generally achieve accuracies above 90% under their respective experimental conditions, the inherent limitations of each modality restrict their potential for achieving higher recognition performance.

To address the limitations of single-modality approaches, researchers have increasingly focused on fusing the multimodal data to enhance target recognition performance [36–38]. Multimodal learning involves extracting and integrating features from multiple data sources,

and the multimodal fusion methods generally categorized into the early fusion, intermediate fusion, and late fusion methods. The early fusion method concatenates features from different modalities at the pre-model stage and processes them through a unified model. In the intermediate fusion, features from different modalities are extracted independently using specific networks and fused at an intermediate layer before further processing through a shared network. For example, the feature-level fusion by combining the HRRP, range instantaneous Doppler (RID), and joint time frequency (JTF) data has been presented in [33], which can achieve the improved target recognition performance. Similarly, the multidomain feature-level fusion network based on an attention mechanism named MdFFNet [34] leverages the complementary characteristics of the HRRP, RCS, and TF data, incorporating an attention mechanism in the feature-level fusion module to selectively enhance important domain features. The late fusion method processes the data from different modalities through independent networks and performs the fusion at the decision-making stage. For instance, Tian et al. fused the narrowband and broadband radar data at the decision level [32], and other studies have also combined broadband and narrowband radar echoes for the target recognition.These methods combining broadband and narrowband radar data have shown promising results in their respective experimental settings, with recognition accuracies typically reaching above 95%. Although the multimodal fusion of broadband and narrowband radar demonstrates significant performance advantages in target recognition, radar systems typically employ narrowband radar echoes due to practical engineering constraints. Narrowband radar echoes offer numerous advantages, including highly focused energy, strong anti-interference capabilities, and low signal processing complexity. Consequently, achieving efficient micromotion target recognition based on single narrowband echoes has become an important research direction in this field.

To fully excavate the information embedded in the multimodal data derived from narrowband radar echoes, this paper emphasizes the exploration of the role and interdependence of each modality. By interpreting the multimodal fusion as a hierarchical interaction process, this paper proposes a graph fusion network constructed by the cross-modal attention mechanism named GF-AM Net, which can achieve the high-precision micromotion target identification by fusing multimodal features. First, the convolutional neural network (CNN) is used to extract features from RCS, TF, and CVD modalities, effectively capturing their unique characteristics. Then, these features are progressively fused within a graph structure through a cross-modal attention mechanism, simulating multilevel interactions across the unimodal, bimodal, and trimodal configurations. The hierarchically fused features are subsequently processed through a classification module, facilitating the accurate recognition of the micromotion target based on narrowband radar data.

The main contributions of this paper are as follows: (1) A multi-domain feature fusion method is proposed for narrowband radar signals, which effectively integrates RCS, TF, and CVD characteristics extracted from narrowband radar echoes to achieve robust micromotion target identification. (2) A graph fusion network based on the cross-modal attention mechanism is designed. In the network architecture, a graph-like structure is utilized to hierarchically design single-modal, dual-modal, and tri-modal interactions, where modal interactions are treated as vertices. Different vertex information is fused through the cross-modal attention mechanism to better capture the inter-modal interactions. (3) Through simulation experiments, the effectiveness of the GF-AM Net in narrowband radar data is verified, which significantly improves the identification accuracy of micromotion targets.

2. Signal Modeling

2.1. Micromotion Signal Modeling

In challenging radar detection scenarios, primary targets and interfering targets will show different micromotion forms during the flight, through which the targets can be well classified and identified. The motion process of the micromotion model is usually manifested as a precession motion [33,34], as shown in Figure 1. If the target spins, the target will rotate around its symmetry axis with an angular velocity w_s, and the precession motion is to rotate around the symmetry axis ON at the same time as the target spins, with an angular velocity of w_c and a precession angle of θ, i.e., the angle between symmetry axes ON and Oz. Let the scattering center P on the spatial cone be in the coordinate system $oxyz$, with the initial coordinate is $p_0 = (x_p, y_p, z_p)^T$. At the moment t, in the precession model, the coordinates of scattering center P in the coordinate system $oxyz$ are given by

$$p_c(t) = R_c(t)R_s(t)p_0 \tag{1}$$

where $R_s(t)$ is the spin matrix, $R_c(t)$ is the precession matrix, and the angular velocity of the spin is w_s. From the Rodrigues equation, the rotational transformation matrix $R(t)$ can be expressed as follows:

$$R(t) = \vec{I} + \hat{e}\sin\omega t + \hat{e}^2(1 - \cos\omega t) \tag{2}$$

where $\hat{e}$ is skew symmetric matrix of the angular rotation axis $\vec{e} = [e_x, e_y, e_z]^T$. Thus, the spin matrix expression is given by

$$R_s(t) = \begin{bmatrix} \cos(w_s t) & -\sin(w_s t) & 0 \\ \sin(w_s t) & \cos(w_s t) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

Similarly, the precession matrix $R_c(t)$ can be expressed as follows:

$$R_c(t) = \begin{bmatrix} \cos(w_c t) & -\cos\theta\sin(w_c t) & \sin\theta\sin(w_c t) \\ \cos\theta\sin(w_c t) & 1-\cos^2\theta[1-\cos(w_c t)] & \sin\theta\cos\theta[1-\cos(w_c t)] \\ -\sin\theta\sin(w_c t) & \sin\theta\cos\theta[1-\cos(w_c t)] & 1-\sin^2\theta[1-\cos(w_c t)] \end{bmatrix} \tag{4}$$

where w_c is the precession angular velocity, and θ denotes the precession cone angle. These matrices provide a comprehensive description of the motion process and positional changes of the scattering center P in the precession model.

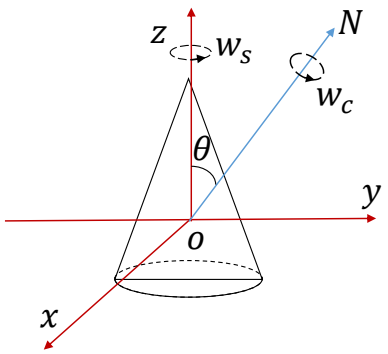


Figure 1. Diagram of the micromotion model.

2.2. Narrowband Radar Signal Modeling

Assuming that $S_{th}(\theta_s, f_s)$ represents the static electromagnetic scattering data of the micromotion target, which changes with the attitude angle θ_s and frequency f_s. The attitude angle change of the micromotion target in the flight process can be obtained by establishing the micromotion signal model, and the narrowband radar echo signals can be obtained by introducing this dynamically changing attitude angle into the electromagnetic scattering data of the narrowband radar target. Thus, the RCS data and narrowband radar echo signals can be obtained during the micromotion target's motion data $S_{th}(\theta_a(t), f_p)$, where f_p represents a frequency point within the signal bandwidth.

The narrowband radar echo can be processed using STFT to obtain the TF image, which illustrates the signal's distribution in both time and frequency. The critical step in constructing the TF image is selecting an appropriate window function. This function is applied to extract local signals at each time moment t of the narrowband radar signal. Then, a Fourier transform is performed on the local signals within each window to obtain the local spectrum. Finally, the local spectra are arranged sequentially to form the TF image of the signal. The TF data are constructed by organizing these spectra as follows:

$$T(t, f_d) = STFT(S_{th}(\theta_a(t), f_p)) \tag{5}$$

where f_d denotes the Doppler frequency. The TF image displays the variation in the frequency of the micromotion target over time during its flight, providing insight into the trend of the micromotion characteristics as the micromotion target moves.

To analyze the micromotion characteristics more comprehensively, the fast Fourier transform (FFT) of the obtained TF image along the time axis can be used to derive the CVD [30,31]. The CVD is a visual representation, which is used to analyze the frequency velocity characteristics of the micromotion target. It provides the valuable information about the repetition tempo and the maximum Doppler shift of each frequency component. Moreover, the CVD does not rely on the initial position of the moving target, making it more robust than the TF image. The relationship between narrowband radar echo, TF image, and CVD is illustrated in Figure 2. The formula for CVD is as follows:

$$C(f_c, f_d) = FFT(T(t, f_d)) \tag{6}$$

where f_c denotes the cadence frequency.

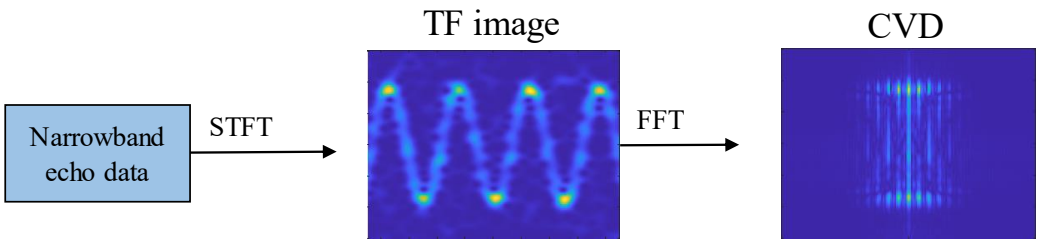


Figure 2. Process of converting the narrowband radar echoes into the TF image and CVD.

3. Proposed Method

In this section, a narrowband radar micromotion target identification method based on the GF-AM Net using the multimodal feature fusion is proposed. The detailed network structure of the GF-AM Net is designed to achieve the high-precision recognition of the micromotion target in the narrowband radar. The GF-AM Net operates in several key stages to efficiently process the multimodal data and enhance the micromotion target recognition accuracy. First, this network extracts unimodal features from three different modalities

(the RCS, TF, and CVD) using the CNNs. Each modality provides the distinct information about the micromotion target's micromotion characteristics, and CNNs are well-suited to effectively extract deep-level features within these modalities. The unimodal features derived from each modality are then fed into a graph fusion structure that incorporates a cross-modal attention mechanism. This structure is designed to simulate layer-by-layer interactions between different modalities, aiming to uncover potential inter-modal associations and facilitate progressive fusion from the unimodal to multimodal data. Within this fusion structure, the network initially generates bimodal features by combining two different modalities to explore the interrelationships between them. Subsequently, the three-modal features are further interacted to generate trimodal features, completing the deep fusion of the multimodal data. To optimize feature representation further, different weights are assigned to the unimodal, bimodal, and trimodal features based on their relative importance. This weighting mechanism enhances the recognition task by emphasizing the most informative features from each modality, allowing for a more comprehensive portrayal of the micromotion characteristics of micromotion targets. By incorporating these weighted features, the network improves both the accuracy and robustness of the recognition task. Finally, the weighted output is passed into a classification module, yielding the final recognition results and enabling high-precision recognition of micromotion targets in the narrowband radar. The overall structure of the GF-AM Net is illustrated in Figure 3.

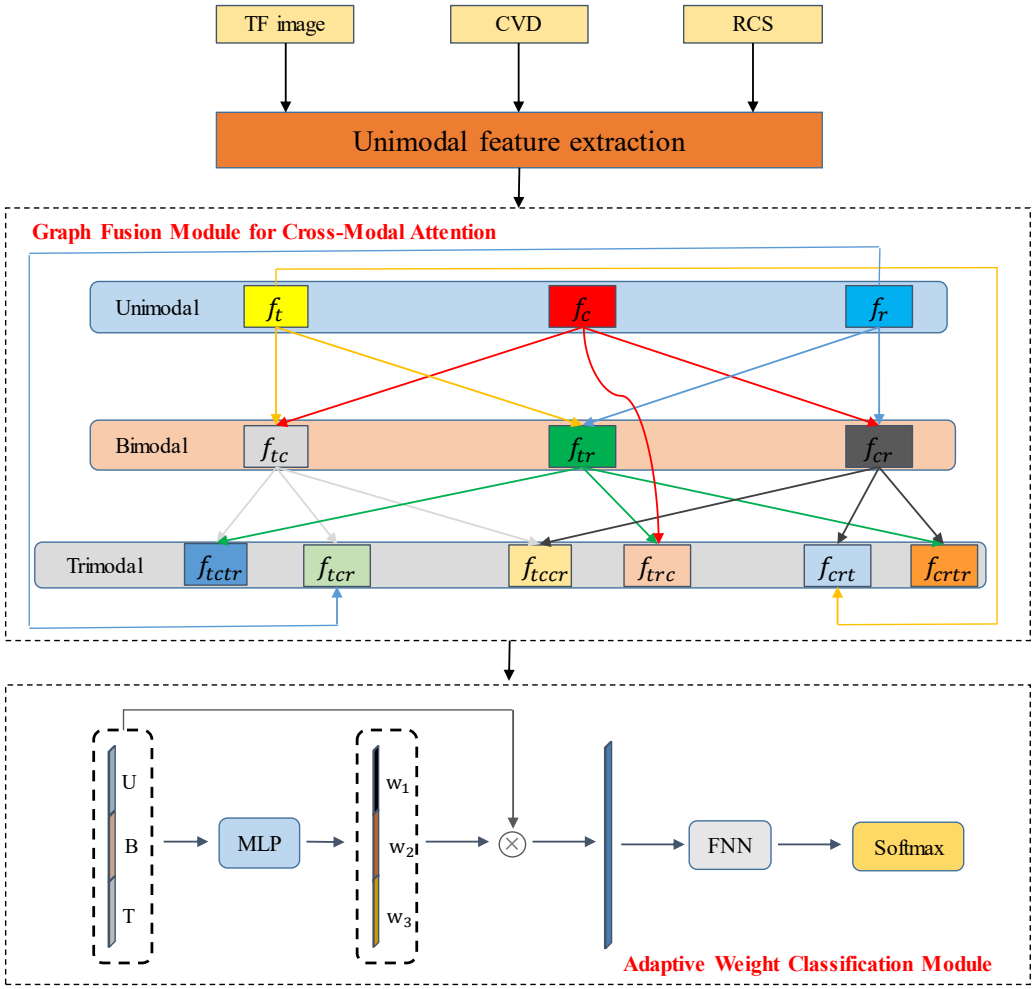


Figure 3. Overall structure of the GF-AM Net.

3.1. Unimodal Feature Extraction

In order to achieve the unimodal feature extraction from the different modal data, two distinct CNN structures have been employed to process one-dimensional (1D) time-series data (RCS) and two-dimensional (2D) image data (TF image and CVD). By leveraging the CNN's efficient information retention capabilities, such as the local connectivity, weight sharing, and pooling operations, the network can effectively extract unimodal features that capture the most relevant information from these modalities. The local connectivity in CNNs allows the network to focus on local patterns in the data, which is particularly useful for both the time-series RCS data and the spatially structured image data (TF image and CVD). Weight sharing reduces the number of parameters, enabling the model to generalize better, while pooling operations help to retain the most important features while reducing the spatial dimensions. This approach ensures that each modality's features are extracted efficiently, and then the obtained unimodal features can provide the valuable information for subsequent fusion and classification tasks. The structure of the unimodal feature extraction network is shown in Figure 4.

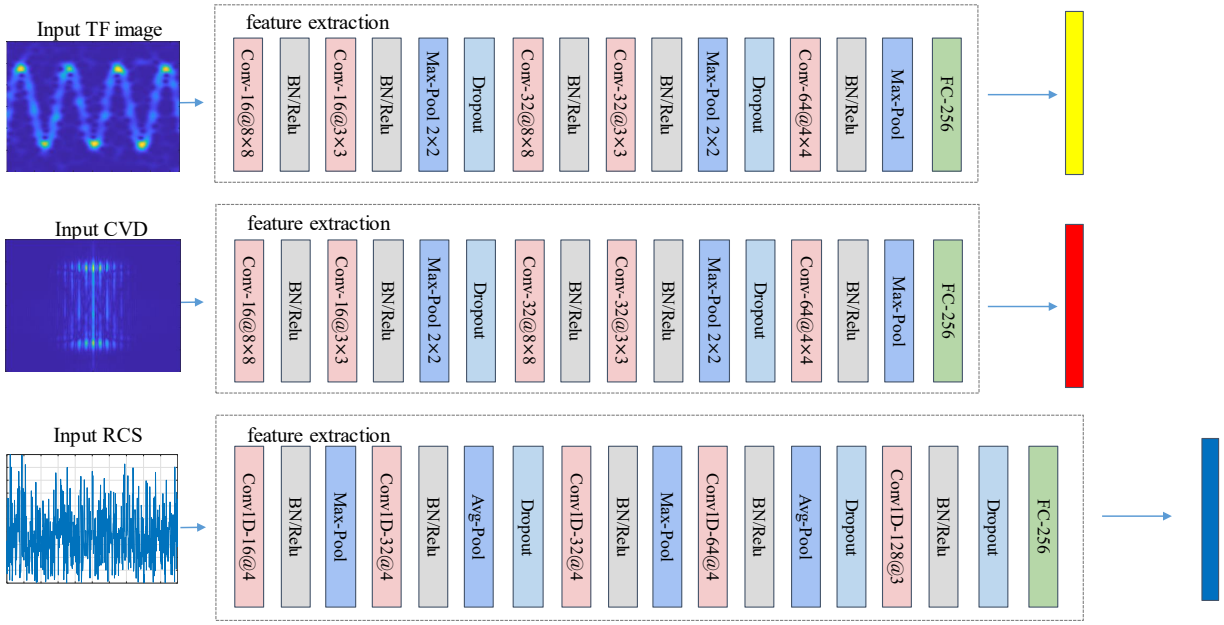


Figure 4. Structure diagram of the unimodal feature extraction network.

When handling the RCS data, we employ a 1D convolution (Conv1D) module when the feature extraction is employed, which is well-suited for the time-series data. The Conv1D module operates through the layer-by-layer convolution, effectively processing the long time-series data by expanding the receptive field. This allows the model to capture both local and global dependencies in the data. For instance, using the Conv1D-16@4, where "16" denotes the number of output channels and "4" represents the size of the convolution kernels, the network extracts complex temporal features. After the convolution operation, a pooling layer is applied to reduce the size of the feature map, thereby enhancing the feature compactness. A batch normalization (BN) layer is then used to accelerate the convergence process, followed by a dropout layer to improve the generalization and prevent the overfitting.

In contrast to the RCS time-series data, both TF image and CVD are the 2D image data in the frequency domain. To effectively extract complex spatial features from these images, a 2D convolution (Conv2D) module for the feature extraction was adopted. Conv2D is capable of recognizing spatial patterns within the images and constructing the deep, high-

level semantic feature representations layer by layer. Similar to the RCS processing based on the Conv1D module, the Conv2D module is followed by pooling layers, a BN layer, and a dropout layer to enhance the model performance and prevent the overfitting.

Ultimately, with the design of two CNNs, this paper extracts independent features from each of the three modalities—RCS, TF image, and CVD. These extracted features serve as the foundational elements for the subsequent cross-modal graph-based fusion. By separately processing the time-series data (RCS) and image data (TF image and CVD) through specialized CNN architectures, the model is able to capture the unique characteristics of each modality. These unimodal features, in turn, provide the necessary input for the graph fusion process, enabling effective integration of the different modalities to enhance the overall performance of the micromotion target recognition.

3.2. Graph Fusion Module for Cross-Modal Attention Mechanism Construction

In order to better interact with two modal features (i.e., vertices in the graph fusion module), the cross-modal attention mechanism has been used to the cross-modal attention mechanism was used to effectively implement the interaction of two vertices. This paper takes the example of generating $f_{t,c}$ through f_t and f_c, the structure of which is shown in Figure 5. The interaction between these two features is captured by the attention mechanism, facilitating the fusion of the complementary information from both modalities. This interaction enhances the representation of the fused features, which is critical for improving the overall recognition accuracy of the micromotion target.

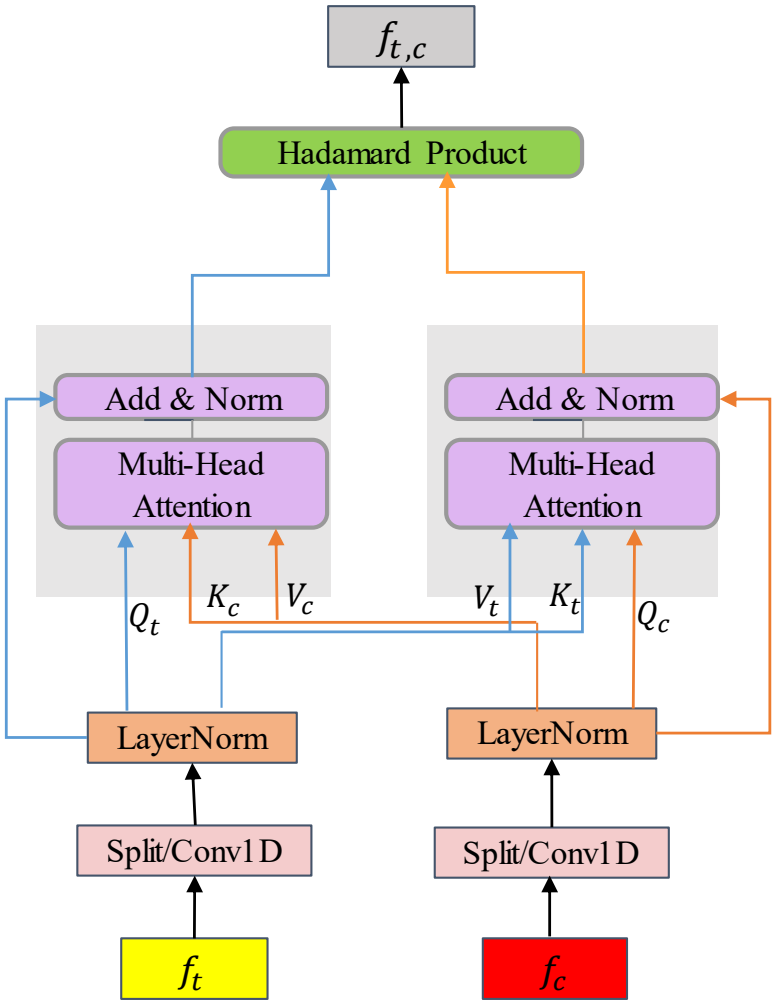


Figure 5. Cross-modal attention mechanism in the graph fusion module.

In Figure 5, f_t and f_c are first passed through the Split and Conv1D, where Split transforms the input features into inputs suitable for the multimodal attention mechanism, and Conv1D enables the elements of the input features to have interactions on their neighborhood elements, as follows:

$$\hat{f}_{\{r,t,c\}} = \text{Conv1D}(f'_{\{r,t,c\}}, k_{\{r,t,c\}}) \in \mathbb{R}^{N_{\{r,t,c\}} \times d} \tag{7}$$

where $k_{\{r,t,c\}}$ is the size of the convolution kernel, $N_{\{r,t,c\}} \times d$ is the feature obtained after splitting, N is the number of segments, and d is the length of each segment. In order to improve the stability and training efficiency of the neural network, stabilize the input distribution and enhance the generalization ability of the model, the Layernorm technique is used [39], which is given by

$$\text{LayerNorm}(f) = \frac{f - \mu}{\sigma}\gamma + \beta \tag{8}$$

where μ and σ are the mean and standard deviation of the input features of the layer. γ and β are learnable scaling and panning parameters that preserve the expressive power of the network.

After the normalization, the two features are subjected to a multi-head cross-modal attention mechanism [40]. The formula for the attention mechanism from the TF image modality to CVD modality is given by

$$\text{Attention}_i^t = \text{softmax}\left(\frac{(X_t W_{ti}^Q)(X_c W_{ci}^K)^T}{\sqrt{d_k}}\right)(X_c W_{ci}^V) \tag{9}$$

$$\text{MultiHead}(Q_t, K_c, V_c) = \text{Concat}(\text{Attention}_1^t, \text{Attention}_2^t, \ldots, \text{Attention}_h^t)W^O \tag{10}$$

The attention mechanism from the CVD to TF image is formulated as follows:

$$\text{Attention}_i^c = \text{softmax}\left(\frac{(X_c W_{ci}^Q)(X_t W_{ti}^K)^T}{\sqrt{d_k}}\right)(X_t W_{ti}^V) \tag{11}$$

$$\text{MultiHead}(Q_c, K_t, V_t) = \text{Concat}(\text{Attention}_1^c, \text{Attention}_2^c, \ldots, \text{Attention}_h^c)W^O \tag{12}$$

where $W_{ti}^Q, W_{ci}^K, W_{ci}^V, W_{ci}^Q, W_{ti}^K$ and W_{ti}^V are the learnable projection matrices with $i \in \{1, \ldots, h\}$ denoting the i-th attention head and h being the total number of attention heads, and X_t and X_c are the feature sequences of the two modalities, TF image and CVD. After computing the cross-modal attention, the residual concatenation of the attention outputs is performed, and then $f_{t,c}$ can be obtained by passing through a LayerNorm layer and performing a Hadamard product. Thus, it can have

$$\text{Output}_t = \text{LayerNorm}(\text{MultiHead}(Q_t, K_c, V_c) + X_t) \tag{13}$$

$$\text{Output}_c = \text{LayerNorm}(\text{MultiHead}(Q_c, K_t, V_t) + X_c) \tag{14}$$

$$f_{t,c} = \text{Output}_t \odot \text{Output}_c \tag{15}$$

To fully exploit the information from each modality and explore the interactions of bimodal and trimodal features, we utilize a three-layer graph fusion network module for the progressive fusion of unimodal, bimodal, and trimodal features [37]. In this structure, each layer of the graph corresponds to a different level of modality fusion: the first layer handles unimodal features, the second layer handles bimodal features, and the third layer captures trimodal features. By modeling the interaction between the modalities as nodes and edges

in the graph, the network learns to progressively refine the inter-modal relationships at each layer. This layer-by-layer interaction allows the model to capture fine-grained connections between the different features, facilitating a more comprehensive fusion and ultimately improving the overall performance of the target recognition task.

In the first layer (unimodal layer), the features of each modality are mainly self-information, denoted as f_r, f_t and f_c (data features from RCS, TF image, and CVD modalities, respectively). Since the importance of the features may vary from modality to modality, we apply the attention mechanism in this layer in order to weight the contribution of each modality. The attention mechanism is applied to weights to the modal features according to their importance, resulting in a weighted average unimodal feature output. After this processing, each modal feature not only contains its own information but also reflects its importance in the network, which is shown below:

$$\alpha_m = MAN(f_m) \tag{16}$$

$$U = \frac{1}{3}\sum_m \alpha_m \cdot f_m, m \in \{r, t, c\} \tag{17}$$

where α_m is the weight of each modality and MAN is the network from which the weights are extracted, consisting of sigmoid activation dense layers. U is the final unimodal vector.

The second layer, referred to as the bimodal dynamic learning layer, focuses on the interaction between two modalities. In this layer, the vertices representing the bimodal features are obtained through a two-by-two interaction of the unimodal information. Specifically, the cross-modal attention mechanism, as discussed earlier, is employed to facilitate the interaction between pairs of unimodal features. This mechanism enables the dynamic exchange of the information between different modalities, allowing the model to learn the interdependencies and complementary information between them. The detailed calculations for this interaction have been provided in the previous section, where the attention mechanism adaptively modulates the information flow between the modalities to refine the bimodal feature representation. Thus, it can have

$$f_{m_1,m_2} = CAM(f_{m_1}, f_{m_2}) \odot CAM(f_{m_2}, f_{m_1}) \tag{18}$$

$$f_{m_1}, f_{m_2} \in \{f_r, f_t, f_c\}, f_{m_1} \neq f_{m_2} \tag{19}$$

where CAM refers to the cross-modal attention mechanism, where the first parameter serves as the query, and the second parameter represents the key and value of the attention mechanism. The operation $\odot$ denotes the Hadamard product, which represents the element-wise multiplication between two values. Through this cross-modal attention mechanism, the model is able to extract features of interest from each modality for the other, facilitating the capture of the complementary information between modalities. This interaction enables the model to generate the more comprehensive, multi-dimensional feature representations, which enrich the overall understanding of the target. By tapping into the complementary characteristics of each modality, the model significantly enhances its ability to generalize and improve both the accuracy and robustness of the target recognition task.

f_{m_1,m_2} is the information vector of the vertices of the second layer. For the weights of the edges, since the similarity of the two vectors increases with the increase in the inner product of these two vectors, it means that, when the inner product of the two vectors is larger, their information overlaps and there is little or no complementary information. Given that the bimodal interactions exhibit varying degrees of significance, the edge weights connecting vertices between the first and second layers can be determined as follows:

$$S_{m_1,m_2} = \tilde{f}_{m_1}^T \tilde{f}_{m_2} \tag{20}$$

$$\hat{\alpha}_{m_1,m_2} = \frac{\alpha_{m_1} + \alpha_{m_2}}{S_{m_1,m_2} + 0.5} \tag{21}$$

$$\alpha_{m_1,m_2} = \frac{e^{\hat{\alpha}_{m_1,m_2}}}{\sum_{m_i \neq m_j} e^{\hat{\alpha}_{m_i,m_j}}} \tag{22}$$

where $\tilde{f}_{m_1}$ denotes the Softmax normalized vector of f_{m_1} (the normalization ensures that the computed similarity is between 0 and 1), S_{m_1,m_2} denotes the similarity of the two vectors, $\hat{\alpha}_{m_1,m_2}$ denotes the weights of the unnormalized edges, and α_{m_1,m_2} denotes the normalized weights. The weighted average of the bimodal vertex information of the second layer can be obtained as follows:

$$B = \sum \alpha_{m_1,m_2} \cdot f_{m_1,m_2} \tag{23}$$

In the bimodal fusion stage of the second layer, the cross-modal attention mechanism is employed to interact the vertices of the bimodal features two-by-two, generating the trimodal vertices of the third layer. This fusion process results in three distinct trimodal vertices, each representing the integration of the information from different modalities. Furthermore, to achieve a more comprehensive integration of modal features, the cross-layer fusion is also performed by combining the unimodal feature vertices from the first layer, those not involved in the corresponding bimodal vertex, into the second layer's bimodal vertices. This generates additional trimodal vertices, leading to a total of six trimodal vertices, each representing a different combination of modalities. This multi-layered fusion approach, incorporating both bimodal and trimodal interactions, allows the network to capture the more complex and nuanced micromotion information from diverse modal combinations. For the edges connecting the third layer, the same weight calculation method is applied as in the previous layers, ensuring the consistent and effective fusion. The weighted information from each trimodal vertex is then aggregated, resulting in the final trimodal information representation, denoted as T. This comprehensive fusion strategy not only strengthens the correlation between the modalities, but also equips the network with the detailed modal interaction information, which can enhance its ability to accurately capture the micromotion features of micromotion targets.

3.3. Adaptive Weight Classification Module

From the graph fusion module, it can obtain the information corresponding to the unimodal, bimodal, and trimodal features. To enable the model to flexibly assign weights to different layers of features and then adaptively adjust the contribution of each layer, the unimodal, bimodal, and trimodal features are concatenated, which is denoted as U, B, and T, respectively. These concatenated features are then passed through a multilayer perceptron (MLP) composed of three fully connected layers, where the potential interactions between the different features are captured. Finally, the weight coefficients are computed through a Sigmoid activation function, as shown in the Adaptive Weight Classification Module in Figure 3. The process is formalized as follows:

$$\omega = \sigma(W_3 \cdot \mathrm{ReLU}(W_2 \cdot \mathrm{ReLU}(W_1 \cdot x + b_1) + b_2) + b_3) \tag{24}$$

where x represents the input after concatenating U, B, and T. W_1, W_2, W_3, b_1, b_2, and b_3 are all learnable parameters in the fully connected layer and σ is the Sigmoid activation function.

ω will be divided into three weight coefficients $\omega = [\omega_1^T, \omega_2^T, \omega_3^T]^T$. The weighting coefficients are weighted with the input features of different modalities of the input, and finally the weighted features are obtained with the following expression:

$$F_{\text{fuse}} = \omega_1 U + \omega_2 B + \omega_3 T \tag{25}$$

where ω_1 is the weight of U, ω_2 is the weight of B, and ω_3 is the weight of T.

By using the weighted features, the contribution of each layer in the graph fusion network module can be dynamically adjusted, which can enhance the model's robustness and improve the feature representation. The weighted features are then passed into a feed-forward neural network (FNN), which consists of multiple fully connected layers, activation functions, batch normalization, and dropout layers. This structure enables the network to map the features to the final classification output. The FNN performs the hierarchical feature compression and expansion, allowing the network to extract the global information from the high-dimensional input features. Additionally, regularization techniques, such as the dropout and batch normalization, are employed to improve the generalization ability of the model, thereby reducing the risk of the overfitting. This process ensures that the rich feature set is preserved while enhancing the model's robustness to the noise, ultimately improving the final classification accuracy. Through this method, the network is able to make more accurate predictions while maintaining a high degree of the generalization to unseen data.

4. Experimental Results and Analysis

4.1. Dataset Generation

Since the measured data are difficult to obtain, the dimensions and micromotion parameters of the four types of micromotion targets are designed according to the literature [33,34], and their three-dimensional (3D) model structure is shown in Figure 6. The physical optics (PO) method is used to obtain the simulated narrowband radar electromagnetic (EM) data and RCS data. Since the designed four types of micromotion targets have rotational symmetry, the azimuth angle is fixed at 90°, and the pitch angle is varied from 0 to 180° with an interval of 0.1°, so that the narrowband radar echo data can be obtained, as shown in Figure 7.

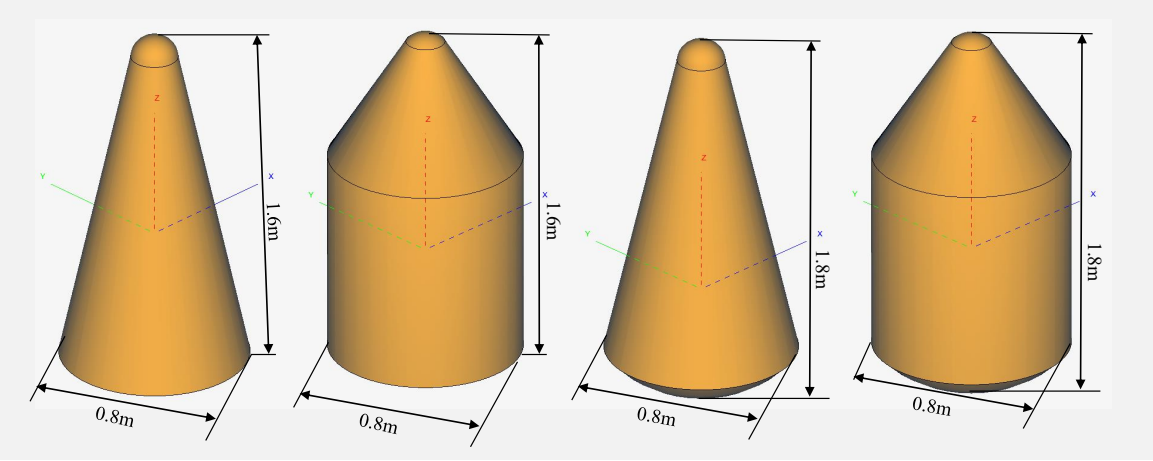


Figure 6. 3D structure of the four types of micromotion targets. From left to right, they are Target 1, Target 2, Target 3, and Target 4.

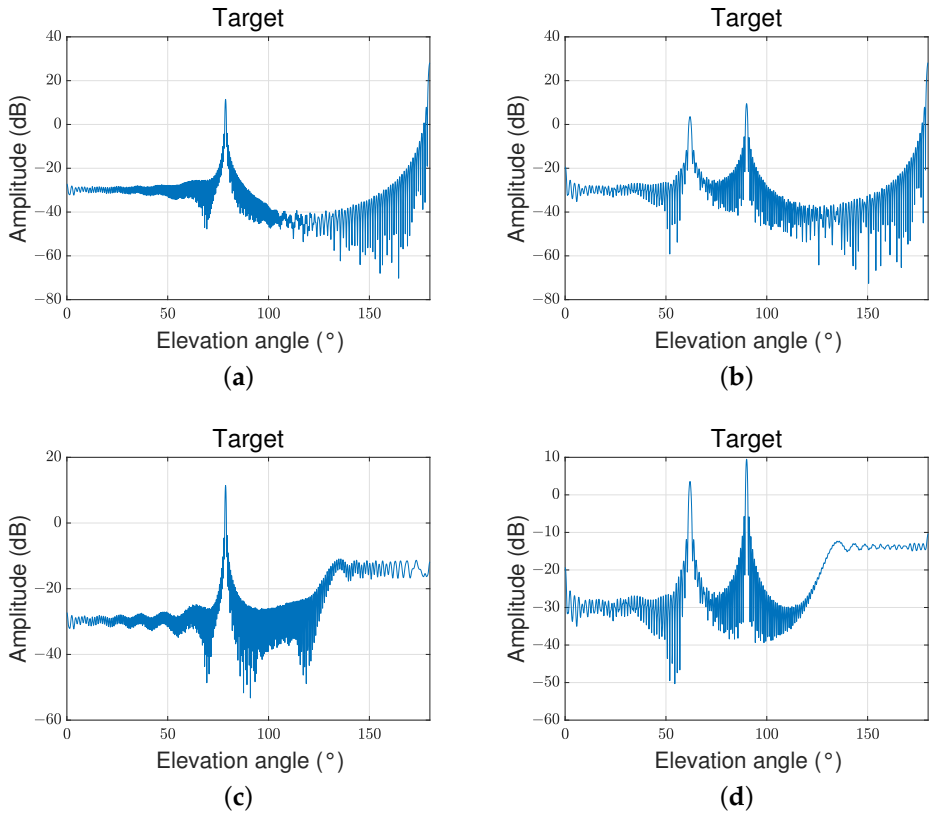


Figure 7. Static narrowband radar echo data for four types of micromotion targets. (**a**) Target 1; (**b**) Target 2; (**c**) Target 3; (**d**) Target 4.

In order to better simulate the real situation of the micromotion target, let us set the pulse repetition frequency (PRF) of the narrowband radar signal to be 600 Hz and the observation time to be 1 s. The initial pitch angle is from 31° to 40° with an interval of 1°, and the precession frequency and precession angle of each target are different, in which the precession frequency interval is 0.2 Hz and the precession angle interval is 0.15°, and the actual parameters are shown in Table 1. Using the micromotion parameters with the static narrowband radar echo data obtained from the simulation to generate 1210 sample data ($10 \times 11 \times 11$) for each category, there are 4840 sample data for four categories of micromotion targets. In order to better realize the recognition of micromotion targets in reality, this paper divides the dataset from 31 to 40° according to the angle, with the first fifty percent as the training set, and the second fifty percent as the test set. In order to simulate the effect of the noise on the echo, we added the Gaussian noise into the signal model, in four cases of 0 dB, 5 dB, 10 dB, and 15 dB. The dynamic narrowband radar echo data are processed accordingly to obtain the RCS time series, TF image, and CVD of the micromotion target. Partial plots of the three modalities of the micromotion target with a signal-to-noise ratio (SNR) of 5 dB are shown in Figures 8–10.

Table 1. Micromotion parameters of the micromotion targets.

Targets	Initial Elevation Angle	Precession Frequency	Precession Angel
Target 1	31:1:40°	2:0.2:4 Hz	4:0.15:5.5°
Target 2	31:1:40°	1.5:0.2:3.5 Hz	2:0.15:3.5°
Target 3	31:1:40°	1:0.2:3 Hz	3:0.15:4.5°
Target 4	31:1:40°	0.5:0.2:2.5 Hz	1:0.15:2.5°

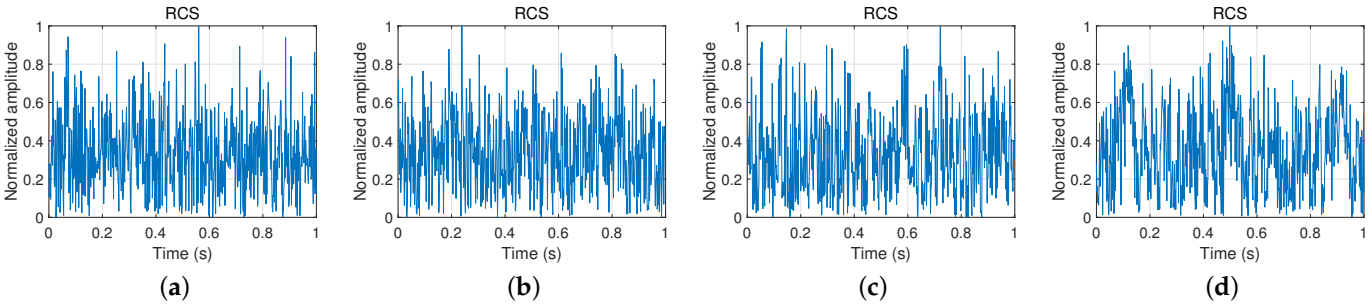


Figure 8. RCS for four types of micromotion targets. (**a**) Target 1; (**b**) Target 2; (**c**) Target 3; (**d**) Target 4.

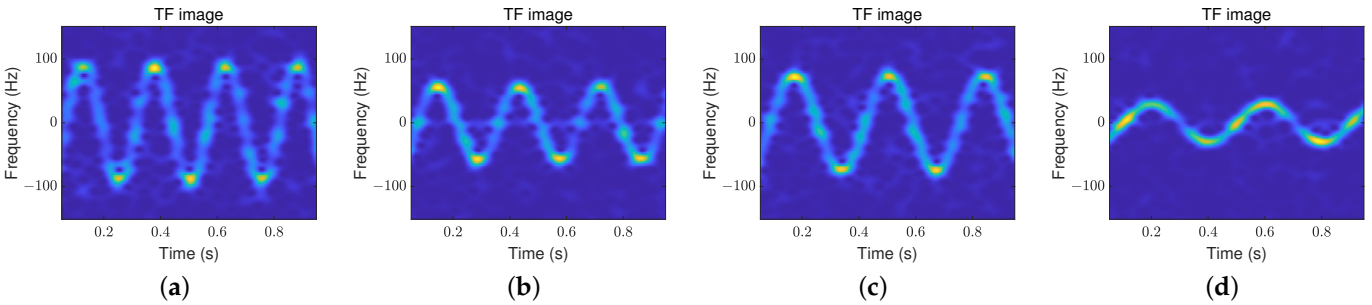


Figure 9. TF image for four types of micromotion targets. (**a**) Target 1; (**b**) Target 2; (**c**) Target 3; (**d**) Target 4.

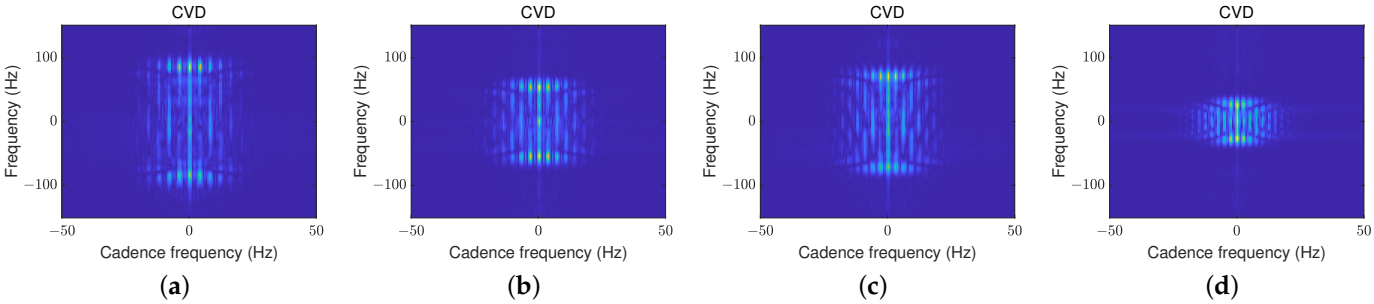


Figure 10. CVD for four types of micromotion targets. (**a**) Target 1; (**b**) Target 2; (**c**) Target 3; (**d**) Target 4.

4.2. Hyperparameter Settings and Comparison Methods

In this paper, two different networks are used for the feature extraction. Conv1D is used to extract features from RCS time series and Conv2D is used to extract unimodal features from TF image and CVD, the extracted features are fed into the graph fusion network for the final output. The learning rate of the whole network is set to 0.0005, the batch size is set to 64, the epoch is set to 100, and then Adam is used as the optimizer of the network to update the weights. The comparison methods are given as follows:

(1) RCS Net [25]. This model mainly consists of a one-dimensional CNN with the RCS time series as its input, which is utilized for the micromotion target recognition.

(2) MobileNet Vit [41]. This model is a combination of the CNN and vision in Transformer, which makes the model able to capture both local features and global information. Firstly, it uses the convolutional layer to extract the local information in the image, and then it uses the attention mechanism to focus on the long-distance dependency and global contextual information, which makes the model able to capture the semantic information of the image in a more comprehensive way. This model can capture the semantic information of the image more comprehensively, which can be well applied to CVD and TF images.

(3) Transfer learning-based parallel network (TLPS Net) [31]. This model takes the TF image and CVD as inputs and applies the migration learning on the ResNet-18 network to construct two parallel networks for the micromotion target recognition.
(4) Attention-augmented cross-modal feature fusion recognition network (ACM-FR Net) [33]. This model employs a feature fusion approach to integrate data from three modalities, i.e., the HRRP, JTF image, and RID. Initially, the features of each modality are extracted, after which the modal information is interactively fused through an adaptive cross-modal feature fusion mechanism. This process enables the identification of micromotion targets. The proposed method utilizes modalities derived from the broadband radar echo data. To facilitate a better comparison with the existing methods, corresponding modifications are made in this paper to ensure that the modalities used are consistent with those in the current model. Specifically, the network designed for the HRRP and JTF data is applied to the CVD and TF data, while the RCS data are processed using the network proposed in this paper.

4.3. Performance Experiments

To demonstrate the effectiveness of the proposed method, it was compared with four other methods, and the accuracy results are shown in Table 2. As can be seen, the proposed method outperforms all the other four methods, achieving an accuracy of 99.55%. This significant improvement can be attributed to the combination of the cross-modal attention with the graph fusion network module, which effectively fuses the information across different modalities, thereby enhancing the recognition accuracy of the micromotion target.

Table 2. Recognition accuracy by different methods.

Methods	Domain	Accuracy
RCS Net	RCS	0.7711
Mobilenetvit_1	TF	0.9236
Mobilenetvit_2	CVD	0.8789
TLPS Net	TF and CVD	0.9165
ACM-FR Net	RCS, TF and CVD	0.9905
GF-AM Net	RCS, TF and CVD	0.9955

To better understand the learning dynamics of different methods, we analyze their training behavior by examining the evolution of accuracy and loss metrics. The convergence curves of different methods during the training process are presented in Figure 11. The proposed GF-AM Net demonstrates stable training behavior throughout the training stage. The accuracy and loss curves indicate consistent and smooth convergence during the training process. Notably, while maintaining the stable training performance, the proposed method achieves the superior accuracy on the test set (as shown in Table 2), demonstrating the better generalization ability.

To further analyze the performance of the proposed method across different target classes, the confusion matrix for each method is plotted and shown in Figure 12. In the confusion matrix, the horizontal axis represents the predicted targets, the vertical axis represents the actual targets, and the values in the cells indicate the accuracy rate. From the matrix, it is evident that the accuracy difference between the smallest and largest class is only 1.5%, which highlights the balanced performance of the proposed method across different classes when compared to other methods.

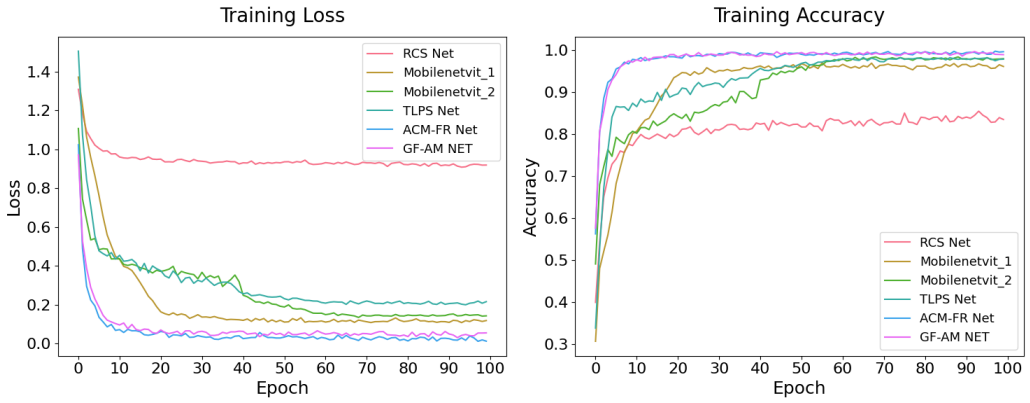


Figure 11. Comparison of training curves for different methods.

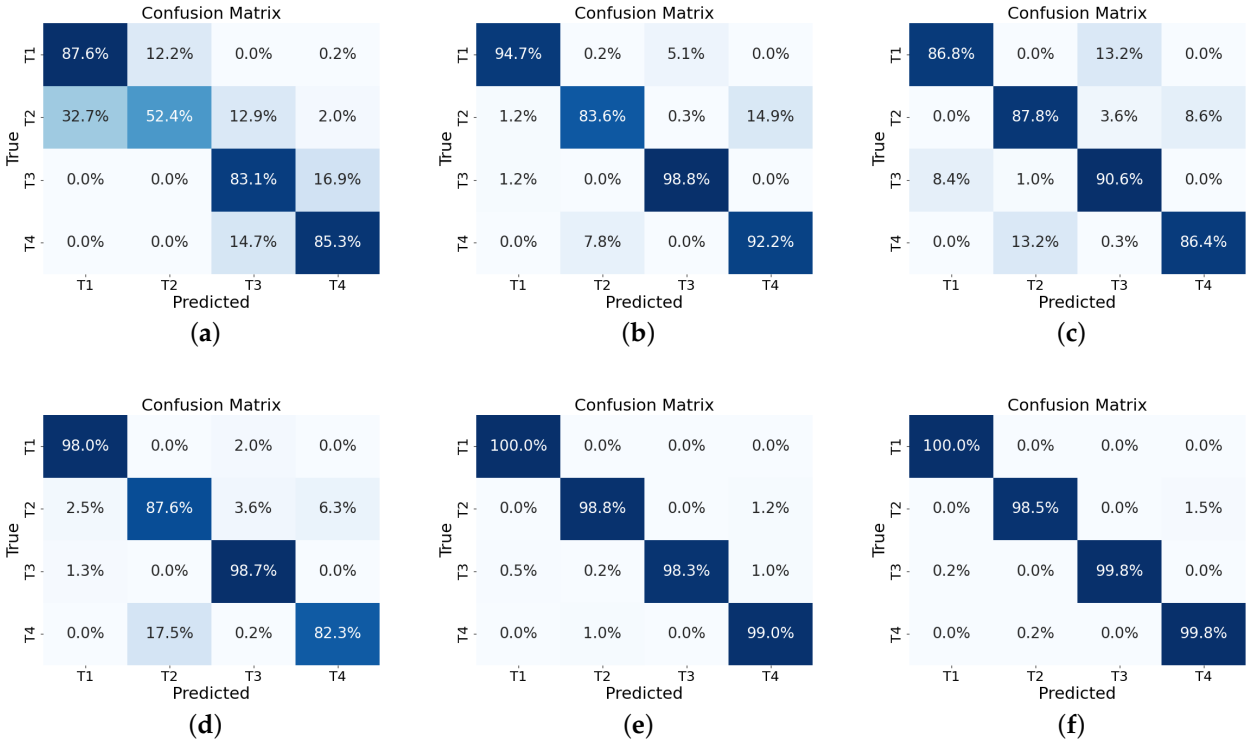


Figure 12. Confusion matrix of different methods. (**a**) RCS Net; (**b**) Mobilenetvit_1; (**c**) Mobilenetvit_2; (**d**) TLPS Net; (**e**) ACM-FR Net; (**f**) GF-AM Net.

To better highlight the superior performance of the proposed method compared to other methods under narrowband radar conditions, t-SNE is utilized to visualize the feature distributions of the proposed and all comparison methods, which is shown in Figure 13. The results can clearly highlight the ability of each method to the differentiate the data. It can be seen that the proposed method in this paper exhibits a more distinct separation between the four categories, with a higher degree of clustering among similar points, in contrast to the other methods. The proposed method indicates a more compact and well-defined clustering structure, suggesting that the GF-AM Net has effectively learned a richer set of features. As a result, the proposed method achieves a better distribution of features in the feature space, further confirming its superior ability to distinguish between micromotion targets.

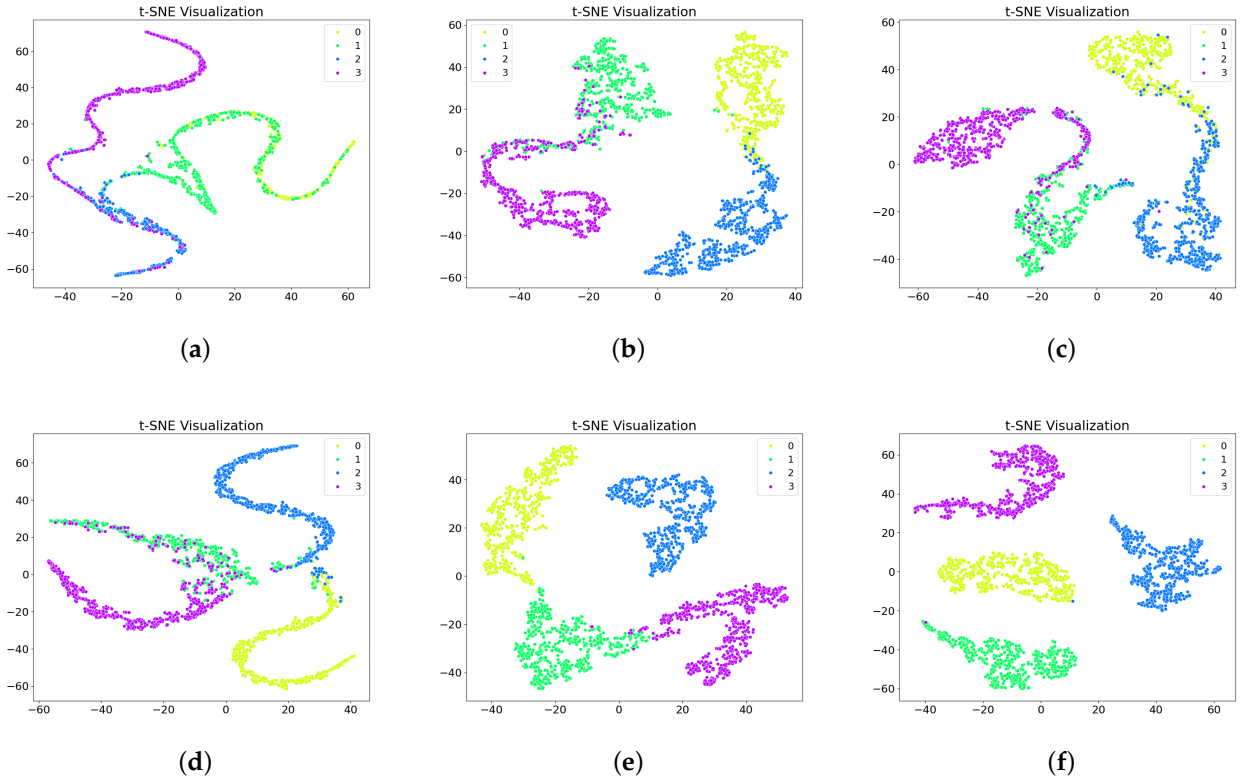


Figure 13. Visualization of the corresponding features extracted by the different networks. (**a**) RCS Net; (**b**) Mobilenetvit_1; (**c**) Mobilenetvit_2; (**d**) TLPS Net; (**e**) ACM-FR Net; (**f**) GF-AM Net.

4.4. Ablation Experiments

To more effectively validate the performance of the tri-modal fusion approach compared to unimodal and bimodal methods, this paper first performed the unimodal feature extraction, then fed the information from the first and second layers of the graph fusion module into the output classification layer. Specifically, the unimodal mode uses only the vertices' information from the first layer, while the bimodal mode leverages the fused features from the first and second layers. The results presented in Table 3 show that, under a 5 dB SNR condition, the tri-modal fusion method improves the recognition performance by at least 6.4% compared to the unimodal mode and at least 0.7% compared to the bimodal mode. This indicates that the information fusion across three modalities provides a more comprehensive feature representation than unimodal or bimodal features, resulting in improved robustness and better noise immunity, which is crucial for the accurate recognition of micromotion targets. These results demonstrate that the proposed network can efficiently extract representative features from all three modalities.

Table 3. Comparison of the recognition effect of the unimodal and multimodal.

Modality	Domain	Accuracy
Unimodal	RCS	0.7777
	TF	0.9310
	CVD	0.9198
Bimodal	RCS and TF	0.9884
	RCS and CVD	0.9860
	TF and CVD	0.9351
Trimodal	RCS, TF and CVD	0.9955

To comprehensively evaluate the proposed multidomain adaptive feature fusion algorithm, we conduct comparative experiments with different fusion approaches, including concatenation-based (Concat Net) and weighted summation-based (Weight Net) fusion techniques. The experimental results are shown in Table 4. The results demonstrate that the proposed feature fusion method achieves the superior performance compared to these alternative approaches. Through the integration of adaptive cross-modal attention mechanisms and graph fusion techniques, the proposed method enables the more precise and nuanced integration of the multimodal data, leading to enhanced target recognition performance. While each fusion approach offers its own merits in the feature integration, the proposed method excels by effectively modeling and leveraging the complex interdependencies between different modalities, which achieves the more robust and discriminative representations of the micromotion target.

Table 4. Comparison of the recognition accuracy of different fusion methods.

Method	Domain	Accuracy
Concat Net	RCS, TF and CVD	0.9636
Weight Net	RCS, TF and CVD	0.9888
GF-AM Net	RCS, TF and CVD	0.9955

To systematically evaluate the effectiveness of different methods in the ablation study, we conduct a detailed analysis of their training dynamics. Figure 14 illustrates the training curves obtained from the ablation experiments, demonstrating the evolution of accuracy and loss metrics throughout the training process. While several comparative methods exhibit comparable convergence patterns with the proposed approach, empirical results on the test set demonstrate that the proposed method achieves superior generalization capability, as shown in Tables 3 and 4.

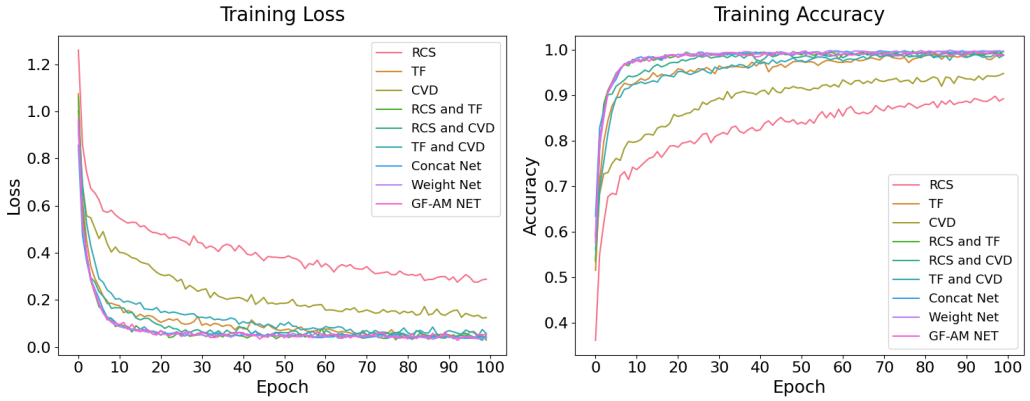


Figure 14. Training curve comparison in ablation experiments.

4.5. Different Networks at Different SNRs

In the above experiments, the Gaussian noise is introduced into the RCS, TF image, and CVD data to simulate real-world conditions. In this section, this paper further evaluates the robustness of the proposed method by testing its performance at different SNRs, ranging from 0 dB to 15 dB. The experimental results are summarized in Table 5. As expected, the recognition accuracy of all methods tends to decrease as the SNR decreases. However, the proposed method demonstrates remarkable robustness, achieving a recognition accuracy of 96.61% even at 0 dB SNR. This recognition is significantly higher than that of the other comparison methods, highlighting the strong anti-interference capabilities and overall robustness of the proposed approach in noisy environments. The results confirm that

the proposed method can effectively extract meaningful features and maintain the high recognition accuracy, even under extreme conditions where the SNR is as low as 0 dB. This underscores the proposed method's ability to handle noisy data and its suitability for real-world applications in challenging radar scenarios.

Table 5. Comparison of the recognition accuracy by different methods under different SNRs.

Methods	Domain	0 dB	5 dB	10 dB	15 dB
RCS Net	RCS	0.5847	0.7711	0.8731	0.9050
Mobilenetvit_1	TF	0.7959	0.9236	0.9314	0.9409
Mobilenetvit_2	CVD	0.7814	0.8789	0.9372	0.9421
TLPS Net	TF and CVD	0.8145	0.9165	0.9302	0.9335
ACM-FR Net	RCS, TF, and CVD	0.9070	0.9905	0.9917	0.9926
GF-AM Net	RCS, TF, and CVD	0.9661	0.9955	0.9975	0.9983

4.6. Discussion

To evaluate the computational efficiency of different models, we compare their theoretical computational cost in terms of GFLOPS (Giga Floating Point Operations Per Second). As shown in Table 6, the proposed GF-AM Net achieves a moderate computational cost of 0.52 GFLOPS, which is significantly lower than TLPS Net (4.74 GFLOPS). While RCS Net shows the lowest computational complexity of 0.01 GFLOPS, the proposed method maintains a better balance between model complexity and performance (as discussed in previous sections). In future work, we plan to further optimize the network architecture to reduce its computational complexity while maintaining its high performance.

Table 6. Comparison of GFLOPS computational cost across different methods.

Method	GFLOPS
RCS Net	0.01
Mobilenetvit_1	0.6
Mobilenetvit_2	0.6
TLPS Net	4.74
ACM-FR Net	0.29
GF-AM Net	0.52

5. Conclusions

The aim of this paper is to address the challenge of the micromotion target identification under narrowband radar conditions. To fully use the information contained in narrowband waveforms, the GF-AM Net for the micromotion target identification using the narrowband radar was proposed, which can make the comprehensive use of the information from different modalities. In this network, two types of CNNs are first employed to extract features from three modalities (RCS, TF image, and CVD), followed by the GF-AM Net. This mechanism simulates multi-level interactions between unimodal, bimodal, and trimodal features. The unimodal, bimodal, and trimodal features are then fused according to their relative importance, ultimately achieving the high-precision micromotion target recognition under narrowband radar conditions. The experimental results from electromagnetic simulations involving four types of micromotion targets demonstrate that the proposed GF-AM Net achieves a recognition accuracy of 99.55%, showcasing its high accuracy and robustness. This validates the effectiveness of GF-AM Net in extracting and integrating multimodal features to improve the micromotion target identification, even in challenging radar environments. The proposed method offers a significant potential for the

real-world applications, where the fusion of the multimodal information plays a crucial role in enhancing the performance of radar-based target recognition systems.

Author Contributions: Conceptualization, Y.Z., H.X. and T.G.; methodology, Y.Z. and H.X.; software, Y.Z. and H.X.; validation, Y.Z. and H.X.; formal analysis, H.X., N.Z. and Z.L.; investigation, Y.Z. and H.X.; resources, Y.Z., H.X. and H.L.; data curation, Y.Z., H.X. and M.G.; writing—original draft preparation, Y.Z. and H.X.; writing—review and editing, Y.Z., H.X., H.L., M.G. and B.X.; visualization, Y.Z. and H.X.; supervision, Y.Z., H.X. and T.G.; project administration, Y.Z., H.X. and Z.L.; funding acquisition, H.X., T.G. and N.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Shenzhen Science and Technology Program (Grant No. JCYJ20240813151238049 and Grant No. 202206193000001, 20220815171723002), by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515011588), by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (Grant No. 24qnpy202 and Grant No. 24qnpy164), by the China Postdoctoral Science Foundation (Grant No. 2024M753740), by the National Natural Science Foundation of China (Grant No. 62203465 and Grant No. 62201614), and the Science and Technology Planning Project of Key Laboratory of Advanced IntelliSense Technology, Guangdong Science and Technology Department (Grant No. 2023B1212060024). Hongtu Xie is the corresponding author.

Data Availability Statement: Data are contained within the article and can be made available by the authors.

Acknowledgments: The authors would like to thank the editors and reviewers for their very competent comments and helpful suggestions to improve this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xie, H.; Zhang, Y.; He, J.; Yi, S.; Zhang, L.; Zhu, N. Foliage-concealed target change detection scheme based on convolutional neural network in low-frequency ultra-wideband SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 19302–19316.
2. Xie, H.; Yi, S.; He, J.; Zhang, Y.; Lu, Z.; Zhu, N. Low-frequency ultra-wideband synthetic aperture radar foliage-concealed target change detection strategy based on image stacks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 19817–19830.
3. Xie, H.; He, J.; Lu, Z.; Hu, J. Two-level feature-fusion ship recognition strategy combining HOG features with dual-polarized data in SAR images. *Remote Sens.* **2023**, *15*, 4393. [CrossRef]
4. Tang, W.; Yu, L.; Wei, Y.; Tong, P. Radar target recognition of ballistic missile in complex scene. In Proceedings of the 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Chongqing, China, 11–13 December 2019; pp. 1–6.
5. Zhu, J.; Yin, T.; Guo, W.; Zhang, B.; Zhou, Z. An underwater target azimuth trajectory enhancement approach in BTR. *Appl. Acoust.* **2025**, *230*, 110373.
6. Zhu, J.; Xie, Z.; Jiang, N.; Song, Y.; Han, S.; Liu, W.; Huang, X. Delay-Doppler map shaping through oversampled complementary sets for high-speed target detection. *Remote Sens.* **2024**, *16*, 2898. [CrossRef]
7. Zhu, J.; Song, Y.; Jiang, N.; Xie, Z.; Fan, C.; Huang, X. Enhanced Doppler resolution and sidelobe suppression performance for Golay complementary waveforms. *Remote Sens.* **2023**, *15*, 2452. [CrossRef]
8. Xie, Z.; Wu, L.; Zhu, J.; Lops, M.; Huang, X.; Shankar, B. RIS-Aided radar for target detection: Clutter region analysis and joint active-passive design. *IEEE Trans. Signal Process.* **2024**, *72*, 1706–1723.
9. Choi, I.O.; Park, S.H.; Kim, M.; Kang, K.B.; Kim, K.T. Efficient discrimination of ballistic targets with microomotions. *IEEE Trans. Aerosp. Electron. Syst.* **2019**, *56*, 1243–1261.
10. Duan, J.; Xie, P.; Zhang, L.; Ma, Y. Space target dynamic identification by exploiting geometrical feature flow from ISAR image sequences. *IEEE Sens. J.* **2022**, *22*, 21877–21884.
11. Zhang, X.; Yang, P.; Cao, D. Synthetic aperture image enhancement with near-coinciding Nonuniform sampling case. *Comput. Electr. Eng.* **2024**, *120*, 109818.
12. Xie, Z.; Xu, Z.; Han, S.; Zhu, J.; Huang, X. Modulus constrained minimax radar code design against target interpulse fluctuation. *IEEE Trans. Veh. Technol.* **2023**, *72*, 13671–13676. [CrossRef]

13. Silberman, G.L. Parametric classification techniques for theater ballistic missile defense. *Johns Hopkins APL Tech. Dig.* **1998**, *19*, 323.
14. Chen, X.; Ye, C.; Wang, Y.; Zhang, Y.; Hu, Q. Unambiguous estimation of multidimensional parameters for space precession targets with wideband radar measurements. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5112716. [CrossRef]
15. Hanif, A.; Muaz, M.; Hasan, A.; Adeel, M. Micro-Doppler based target recognition with radars: A review. *IEEE Sens. J.* **2022**, *22*, 2948–2961. [CrossRef]
16. Lei, X.; Fu, X.; Wang, C.; Gao, M. Statistical feature selection of narrowband RCS sequence based on greedy algorithm. In Proceedings of the 2011 IEEE CIE International Conference on Radar, Chengdu, China, 24–27 October 2011; Volume 2, pp. 1664–1667.
17. Gao, H.; Xie, L.; Wen, S.; Kuang, Y. Micro-Doppler signature extraction from ballistic target with micromotions. *IEEE Trans. Aerosp. Electron. Syst.* **2010**, *46*, 1969–1982. [CrossRef]
18. Wang, Z.; Luo, Y.; Li, K.; Yuan, H.; Zhang, Q. Micro-Doppler parameters extraction of precession cone-shaped targets based on rotating antenna. *Remote Sens.* **2022**, *14*, 2549. [CrossRef]
19. Clemente, C.; Pallotta, L.; De Maio, A.; Soraghan, J.J.; Farina, A. A novel algorithm for radar classification based on Doppler characteristics exploiting orthogonal pseudo-Zernike polynomials. *IEEE Trans. Aerosp. Electron. Syst.* **2015**, *51*, 417–430. [CrossRef]
20. Persico, A.R.; Clemente, C.; Gaglione, D.; Ilioudis, C.V.; Cao, J.; Pallotta, L.; De Maio, A.; Proudler, I.; Soraghan, J.J. On model, algorithms, and experiment for micro-Doppler-based recognition of ballistic targets. *IEEE Trans. Aerosp. Electron. Syst.* **2017**, *53*, 1088–1108. [CrossRef]
21. Pan, X.Y.; Wang, W.; Feng, D.J.; Liu, Y.C.; Wang, G.Y. Signature extraction from rotating targets based on a fraction of HRRPs. *IEEE Trans. Antennas Propag.* **2014**, *63*, 585–592. [CrossRef]
22. Persico, A.R.; Ilioudis, C.V.; Clemente, C.; Soraghan, J.J. Novel classification algorithm for ballistic target based on HRRP frame. *IEEE Trans. Aerosp. Electron. Syst.* **2019**, *55*, 3168–3189. [CrossRef]
23. Zhu, N.; Hu, J.; Xu, S.; Wu, W.; Zhang, Y.; Chen, Z. Micromotion parameter extraction for ballistic missile with wideband radar using improved ensemble EMD method. *Remote Sens.* **2021**, *13*, 3545. [CrossRef]
24. Yang, D.; Wang, X.; Peng, Z.; Hu, L.; Li, J. Micro-Doppler separation based on U-Net and plot-curve association for ballistic target. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *59*, 827–836. [CrossRef]
25. Chen, J.; Xu, S.; Chen, Z. Convolutional neural network for classifying space target of the same shape by using RCS time series. *IET Radar Sonar Navig* **2018**, *12*, 1268–1275. [CrossRef]
26. Narasimhamurthy, R.; Khalaf, O.I. Deep learning network for classifying target of same shape using RCS time series. *Air Traffic Manag. Control* **2021**, *9*, 25.
27. Bai, X.; Wang, L.; Zhou, F.; Li, Y.; Hui, Y. Deep CNN for micromotion recognition of space targets. In Proceedings of the 2016 CIE International Conference on Radar (RADAR), Guangzhou, China, 10–13 October 2016; pp. 1–5.
28. Feng, B.; Chen, B.; Liu, H. Radar HRRP target recognition with deep networks. *Pattern Recognit.* **2017**, *61*, 379–393. [CrossRef]
29. Zhang, Y.P.; Zhang, Q.; Kang, L.; Luo, Y.; Zhang, L. End-to-end recognition of similar space cone–cylinder targets based on complex-valued coordinate attention networks. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5106214. [CrossRef]
30. Lee, J.I.; Kim, N.; Min, S.; Kim, J.; Jeong, D.K.; Seo, D.W. Space target classification improvement by generating microDoppler signatures considering incident angle. *Sensors* **2022**, *22*, 1653. [CrossRef] [PubMed]
31. Jung, K.; Lee, J.I.; Kim, N.; Oh, S.; Seo, D.W. Classification of space objects by using deep learning with micro-Doppler signature images. *Sensors* **2021**, *21*, 4365. [CrossRef]
32. Tian, X.; Bai, X.; Xue, R.; Qin, R.; Zhou, F. Fusion recognition of space targets with micromotion. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *58*, 3116–3125. [CrossRef]
33. Tian, X.; Bai, X.; Zhou, F. Recognition of micromotion space targets based on attention-augmented cross-modal feature fusion recognition network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5104909. [CrossRef]
34. Zhang, Y.; Xie, Y.; Kang, L.; Li, K.; Luo, Y.; Zhang, Q. Feature-Level fusion recognition of space targets with composite micromotion. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *60*, 934–951.
35. Kim, J.H.; Kwon, S.Y.; Kim, H.N. Spectral-Kurtosis and image-embedding approach for target classification in micro-Doppler signatures. *Electronics* **2024**, *13*, 376. [CrossRef]
36. Siriwardhana, S.; Kaluarachchi, T.; Billinghurst, M.; Nanayakkara, S. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access* **2020**, *8*, 176274–176285.
37. Mai, S.; Hu, H.; Xing, S. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 164–172.
38. Zhao, F.; Zhang, C.; Geng, B. Deep multimodal data fusion. *ACM Comput. Surv.* **2024**, *56*, 1–36.
39. Lei Ba, J.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.

40. Vaswani, A. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 6000–6010.
41. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

Three-Dimensional Non-Uniform Sampled Data Visualization from Multibeam Echosounder Systems for Underwater Imaging and Environmental Monitoring

Wenjing Cao *, Shiliang Fang, Chuanqi Zhu, Miao Feng, Yifan Zhou and Hongli Cao

Key Laboratory of Underwater Acoustic Signal Processing, Ministry of Education, Southeast University, Nanjing 210096, China; slfang@seu.edu.cn (S.F.); chuanqizhu@seu.edu.cn (C.Z.); miaofeng@seu.edu.cn (M.F.); yf_zhou@seu.edu.cn (Y.Z.); 103200023@seu.edu.cn (H.C.)

* Correspondence: wenjingcao@seu.edu.cn

Abstract: This paper proposes a method for visualizing three-dimensional non-uniformly sampled data from multibeam echosounder systems (MBESs), aimed at addressing the requirements of monitoring complex and dynamic underwater flow fields. To tackle the challenges associated with spatially non-uniform sampling, the proposed method employs linear interpolation along the radial direction and arc length weighted interpolation in the beam direction. This approach ensures consistent resolution of three-dimensional data across the same dimension. Additionally, an opacity transfer function is generated to enhance the visualization performance of the ray casting algorithm. This function leverages data values and gradient information, including the first and second directional derivatives, to suppress the rendering of background and non-interest regions while emphasizing target areas and boundary features. The simulation and experimental results demonstrate that, compared to conventional two-dimensional beam images and three-dimensional images, the proposed algorithm provides a more intuitive and accurate representation of three-dimensional data, offering significant support for the observation and analysis of spatial flow field characteristics.

Keywords: multibeam echosounder systems; three-dimensional data visualization; non-uniform sampling; opacity transfer function

Academic Editors: Xuebo Zhang, Haiyong Xu, Jingyue Lu, Lei Zhang, Marc Pinto and Farman Ali

Received: 4 December 2024
Revised: 11 January 2025
Accepted: 13 January 2025
Published: 15 January 2025

Citation: Cao, W.; Fang, S.; Zhu, C.; Feng, M.; Zhou, Y.; Cao, H. Three-Dimensional Non-Uniform Sampled Data Visualization from Multibeam Echosounder Systems for Underwater Imaging and Environmental Monitoring. *Remote Sens.* **2025**, *17*, 294. https://doi.org/10.3390/rs17020294

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recent advancements in underwater acoustic technology have made the detection of complex non-uniform flow fields a key focus for researchers. Imaging sonar plays a crucial role in measuring underwater flow fields and forms the basis for real-time, efficient assessments of these environments [1–4]. More specifically, multibeam echosounder systems (MBESs), which feature wide coverage and high resolution, are commonly used to assess complex underwater environments [5–7].

MBESs employ phased array and beamforming technology to enable real-time measurements of underwater three-dimensional space [8–10]. The transmitting and receiving transducers utilize curved and linear arrays, respectively, enabling the scanning of multiple transmission angles and ensuring wide coverage. Beamforming technology is employed to capture time-domain signals from various receiving beam directions. These signals are subsequently processed to generate sonar images, offering an intuitive and detailed representation of the characteristics of the observed area [11].

In underwater monitoring applications, MBESs generally present two-dimensional slice maps in horizontal and vertical, including time history maps and echo intensity maps at various radial distances within a single emission fan [12–14]. However, these two-dimensional images do not adequately illustrate the benefits of wide coverage and high precision associated with MBESs, nor do they fully capture the characteristics such as underwater turbulence [15–17]. To effectively monitor the water disturbance field and its dynamic changes, utilizing the acquired three-dimensional echo data to generate underwater images is essential [18]. Two-dimensional visualization of three-dimensional data facilitates a comprehensive and detailed representation of underwater images from multiple perspectives. This approach effectively captures and illustrates intricate features and dynamic variations within flow fields, enabling a deeper understanding of their complex structures and behaviors.

MBESs acquire intensity images at various radial distances corresponding to each receiving beam angle across different emission-angle beam sectors in polar coordinates. However, as the radial distance increases, the resolution between transmitting and receiving beam angle decreases, leading to reduced image quality and significant challenges in target identification at extended ranges [19]. Moreover, capturing two-dimensional slice images at consistent horizontal or vertical distances is inherently impractical in polar coordinates, further limiting multidimensional visualization and data analysis capabilities. To address these challenges, interpolation methods are often employed to facilitate data analysis [20–22]. These methods efficiently transform data from the polar coordinate system to the Cartesian coordinate system, ensuring consistent resolution within the same dimension while enhancing the accuracy and interpretability of sonar images. Such processing steps substantially expand the applicability of MBESs in complex, non-uniform flow field environments and provide a solid foundation for subsequent image rendering and advanced processing.

The two-dimensional visualization of three-dimensional data offers an intuitive approach for monitoring underwater flow fields. Volume rendering is a crucial technique to represent three-dimensional data images as it effectively captures the details and dynamic characteristics of underwater flow fields, making it particularly suitable for visualizing complex datasets [23–26]. Among volume rendering techniques, the ray-casting and ray-marching algorithms are widely utilized. In the context of ray tracing algorithm, Liu et al. derived the relationship between the ray tracing error and the observation elevation angle and established a stochastic model for acoustic positioning, effectively reducing the impact of low-elevation-angle observations on positioning accuracy [27]. The ray-casting algorithm in particular is preferred due to its superior image quality and efficient rendering speed [28–35]. By modeling light propagation within a three-dimensional space, this algorithm effectively captures and renders fine-grained features within the data, thereby enhancing the readability and realism of the visualized images. Moreover, the ability to adjust parameters such as perspective and opacity allows for the highlighting of regions of interest or detailed information, further improving the effectiveness of data interpretation and analysis.

The transfer function plays a critical role in mapping data values to opacity and color, significantly influencing the visualization outcome. Traditional transfer functions, which rely solely on data values, face inherent limitations when dealing with complex scenarios such as intricate boundaries, multi-material interfaces, and high-dimensional data characteristics [36–38]. To address these challenges, histogram-based transfer functions have been employed to delineate tissue properties, enhancing the precision of overlapping tissue classification [39,40]. Moreover, multidimensional transfer functions, incorporating first- and second-order derivatives of the gradient direction, improve material classification

accuracy and effectively manage complex boundaries [41,42]. A novel multidimensional transfer function was introduced for visualizing flow fields, leveraging flow field parameters such as velocity, gradient, curl, helicity, and divergence [43]. This approach enables the efficient extraction and tracking of critical flow field features. Furthermore, a semi-automatic transfer-function-creation method was developed, significantly streamlining the visualization process [44]. Additionally, the integration of ray-casting methods with GPU acceleration technology facilitates real-time extraction and visualization of dynamic and complex structures, offering enhanced capabilities for practical applications [45–48].

The original data collected by MBESs is processed using beamforming and radial layering techniques to obtain acoustic parameters, such as backscattering strength and radial flow velocity, for all measurement units within the transmission fan. During the radial layering of time-domain signals at different receiving angles, the layer thickness is determined by the pulse width of the transmitted signal. Compared to horizontal layering methods, this radial layering approach eliminates the need to consider the utilization efficiency of signals within a single layer or the overlap rate between layers.

Three-dimensional data obtained from MBESs need to be visualized in the form of images to facilitate the intuitive evaluation of water body disturbance characteristics. However, the visualization of MBES three-dimensional measurement data faces several challenges: (1) with increasing distance, the resolution between adjacent transmission and reception angles gradually decreases, leading to spatial resolution inconsistencies; (2) representation in polar coordinates struggles to accurately capture two-dimensional acoustic cross-sectional images at the same depth, horizontal distance, or vertical distance; and (3) developing a visualization method that not only comprehensively captures three-dimensional information but also effectively conveys critical data features in detail remains a significant challenge.

This paper presents a visualization method for three-dimensional non-uniform sampled data from MBESs. It aims to resolve the challenge of image representation of underwater turbulence in three-dimensional datasets. MBESs exhibit inconsistent distance resolution between the transmitting and receiving beam angles, and conventional two-dimensional images do not adequately represent three-dimensional spatial information. The proposed method effectively uses the detailed and comprehensive three-dimensional data obtained from the MBES to visualize three-dimensional information. Initially, the combination of beamforming algorithms with layered processing allows for the processing of the original echo data received by the MBES, which results in the acquisition of echo intensity for the spatial measurement block. By using the active sonar equation, it is possible to derive the backscattering strength, which represents the scattering characteristics of the measurement block. In addition, applying linear interpolation and arc length weighted interpolation algorithms addresses the problem of inconsistent distance resolution in the transmission and reception beams. Finally, we generate an opacity transfer function in the ray-casting algorithm to reveal the detailed image display and complex three-dimensional water data analysis. The contributions of the proposed method can be summarized as follows:

1. A method was proposed for converting three-dimensional original echo data into two-dimensional image visualization to address the limitations of conventional data display methods.

2. The integration of linear and arc length weighted interpolation enhances the clarity of sonar images and addresses the problem of inconsistent distance resolution.

3. The generated opacity transfer function of the ray-casting algorithm facilitates the visualization of complex water-body data, effectively separating the target from the background while revealing detailed information of the target region.

The remainder of this paper is organized as follows: Section 2 describes materials and methods, including the ray-casting method, the formulation of an opacity transfer function derived from data values, gradients, and second derivatives, the linear and arc length weighted interpolation algorithm, complexity analysis, preprocessing of MBESs, and algorithm description. Numerical simulations and test verification are covered in Section 3. Section 4 presents the discussion and conclusion of this paper.

2. Materials and Methods

2.1. Three-Dimensional Data Visualization Based on the Ray-Casting Algorithm

The ray-casting algorithm proposed by Levoy (a direct volume-rendering method utilizing image sequences) has been extensively employed in various domains, including medical imaging and computer vision [49,50]. The fundamental concept of this algorithm is the light-absorption model, articulated as follows:

$$I = \int_a^b q(s) e^{-\int_a^s \kappa(u)du} ds \tag{1}$$

where I represents the light intensity transmitted from voxel a to voxel b within the volume data, $q(s)$ denotes the contribution of voxel light intensity in the light's direction, and $\int \kappa$ is the attenuation function governing the intensity contribution along the light direction.

In the discrete case, (1) can be viewed as the compositing operation of color contributions from all sampled points along the ray direction. Therefore, the final pixel color value obtained by forward compositing along the direction of the viewing ray can be expressed as follows:

$$I = \sum_{i=1}^{n} C_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{2}$$

where C_i represents the color value and α_i denotes the opacity of the ith sampling point in the light beam direction. Figure 1 shows the discrete sampling in the direction of the ray for the absorption model. For the visualization process, rendering relies on color values, while opacity governs the visibility throughout the rendering procedure. There are two ways for this algorithm to synthesize colors (front to back and back to front), and the first one is introduced. Figure 2 shows the synthesis process. (2) can be expressed in a recursive form as follows:

$$C_i^{\Sigma} = C_{i-1}^{\Sigma} + \left(1 - \alpha_{i-1}^{\Sigma}\right) C_i \alpha_i \tag{3}$$

$$\alpha_i^{\Sigma} = \alpha_{i-1}^{\Sigma} + \left(1 - \alpha_{i-1}^{\Sigma}\right) \alpha_i \tag{4}$$

where C_i^{Σ} represents the weighted cumulative value of opacity from the first to the ith sampling point, and α_i^{Σ} denotes cumulative opacity, with initial values of $C_0^{\Sigma} = 0$ and $\alpha_0^{\Sigma} = 0$.

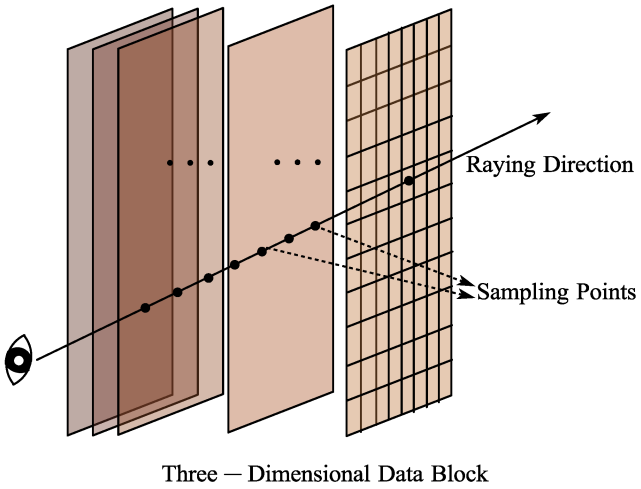


Figure 1. Ray projection path and sampling of three-dimensional data.

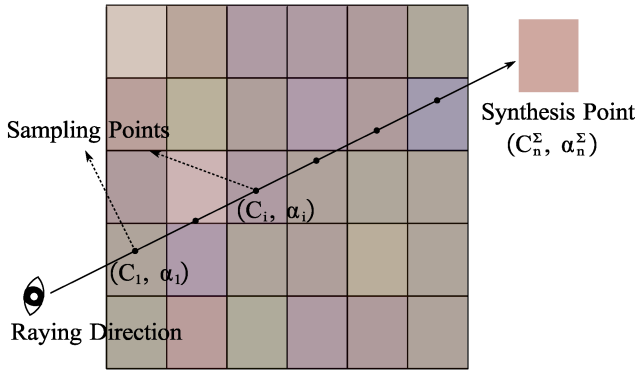


Figure 2. The synthesis process of color values on the raying direction. The colors of the blocks represent the colors of the corresponding pixel points.

2.2. An Opacity Transfer Function Generation Based on Data Values, Gradients, and Second Derivatives

Formulation of transfer functions is a significant area of investigation in three-dimensional data visualization with the ray-casting technique. The primary purpose of the transfer function is to convert scalar values and multidimensional data into color values and opacity. In practical applications, various transfer functions are often designed to enhance data visualization, thereby emphasizing the content of interest [41,51]. The first derivative of scalar data (the gradient) indicates the direction of maximum change of the data, while its magnitude represents the local rate of change of the scalar data. Here, $f^{'}(x)$ denotes the magnitude of the gradient of $f(x)$, with $f(x)$ signifying the scalar function of the data. Moreover, the second-order derivative, denoted as $f^{''}(x)$, can more precisely delineate intricate limits. Figure 3 illustrates the optimal boundary between homogenous materials. It shows a grayscale image of a homogeneous substance. We extract a line from the grayscale image and plot the curves of data value, gradient, and second derivative—Figure 3. Figure 3 reveals that the gradient amplitude in the uniform zone is comparatively small but significantly larger in the variable region; at the boundary, the gradient is high, and the second derivative is null.

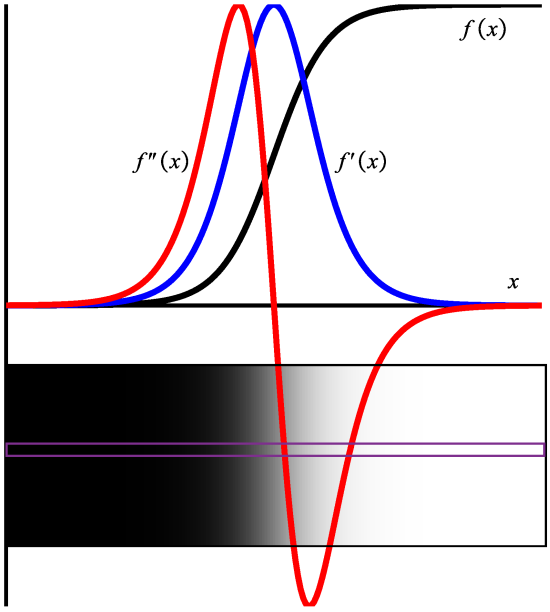


Figure 3. The correlation among data values $f(x)$, gradient magnitudes $f^{'}(x)$, and second-order derivatives $f^{''}(x)$ in relation to boundary.

The gradient derived by Jerrold E. [52] can be expressed as follows:

$$\mathbf{D}_{\mathbf{u}} f = \nabla f \cdot \mathbf{u} \tag{5}$$

$$\mathbf{D}_{\widehat{\nabla f}} f = \nabla f \cdot \widehat{\nabla f} = \nabla f \cdot \frac{\nabla f}{\|\nabla f\|} = \|\nabla f\| \tag{6}$$

where $\mathbf{D}_{\mathbf{u}} f$ and $\mathbf{D}_{\widehat{\nabla f}} f$ represent the first derivatives in direction $\mathbf{u}$ and gradient direction $\widehat{\nabla f}$, respectively, whereas ∇f denotes the first derivative of f.

The second derivative in the direction of the gradient can be expressed as:

$$\mathbf{D}^2_{\widehat{\nabla f}} f = \mathbf{D}_{\widehat{\nabla f}}(\|\nabla f\|) = \nabla(\|\nabla f\|) \cdot \widehat{\nabla f} = \frac{1}{\|\nabla f\|} \nabla(\|\nabla f\|) \cdot \nabla f \tag{7}$$

Alternatively, it can be derived using the Taylor expansion in the direction of gradient $\widehat{\nabla f}$ as follows:

$$\mathbf{D}^2_{\widehat{\nabla f}} f = \frac{1}{\|\nabla f\|^2} (\nabla f)^T \mathbf{H} f \nabla f \tag{8}$$

where $\mathbf{H} f$ is the Hessian matrix of f, representing the 3×3 matrix of f's second-order partial derivatives. Alternatively, the Laplace $\nabla^2 f$ approximation may be used to derive the following relation:

$$\mathbf{D}^2_{\widehat{\nabla f}} f \approx \nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \tag{9}$$

The formulations for these three second-order derivatives possess distinct advantages and limitations. The Hessian matrix approach is the most (numerically) precise. However, other methods offer better efficiency in practice while maintaining a specific degree of precision. The Laplacian compution is direct and efficient; however, it is susceptible to noise. The second derivative method for gradient direction represents a good choice regarding precision and processing complexity.

The mapping from data values to boundaries can be expressed using the first and second derivatives of scalar data [44]:

$$p(v) = -\sigma^2 \frac{h(v)}{g(v)} \tag{10}$$

where $g(v)$ represents the average first-order directional derivative of all $f(\mathbf{x}) = v$ at location $\mathbf{x}$, while $h(v)$ denotes the average second-order directional derivative across all v.

We can determine the range of various target areas upon identifying all boundaries. This article's research indicates that, during data visualization, it is essential to separately consider the opacity transfer functions of the background, target area, and boundary. To distinguish the target area from the background and accurately observe the intensity variations at the boundary and within the target area, the transfer function of the opacity must fulfill the following criteria: (1) The background opacity should be set to zero; (2) The boundary area thickness remains constant to enhance visual effects, while the boundary area opacity decreases as data values increase; (3) The opacity of the target area decreases as the data value increases.

Segmented multiple opacity functions can efficiently delineate the target area and emphasize intricate details via image representation. Assuming the data value range is set to D and partitioned into N non-overlapping subsets, referred to as $D_1, D_2, \cdots, D_N$, it can be expressed as:

$$\begin{cases} D = \bigcup_{i=1}^{N} D_i \\ D_i \cap D_j = \emptyset, (1 < i < j < N) \end{cases} \tag{11}$$

In the paper, the transfer function of opacity can be expressed as:

$$\alpha(\mathbf{x_i}) = \begin{cases} C_{T_1} \cdot \mathrm{e}^{(C(\mathbf{x_i}))} + C_{T_0}, if \;\; v > v_{boundary}, \; |h(f(\mathbf{x_i}))/g(f(\mathbf{x_i}))| \geq p_{ratio} \\ \alpha_v + C_B \cdot \alpha_v \cdot |g_v - g(f(\mathbf{x_i}))|/r \cdot |h(f(\mathbf{x_i}))/g(f(\mathbf{x_i}))|, if \;\; v > v_{boundary}, \; |h(f(\mathbf{x_i}))/g(f(\mathbf{x_i}))| < p_{ratio} \\ 0, else \end{cases} \tag{12}$$

where α_v and g_v represent the opacity and gradient amplitude of the chosen boundary region, C_T denotes a constant, $v_{boundary}$ signifies the boundary value between the target region and the background, p_{ratio} indicates the ratio of the established boundary, and r representes the thickness of the boundary.

2.3. Meshing Based on Linear and Arc Length Weighted Interpolation Algorithms

The distance resolution of the receiving and transmitting beam angles diminishes with increasing radial distance. The examination of two-dimensional sectional images at a consistent depth, horizontal distance, and vertical distance presents challenges. Consequently, it is essential to interpolate the three-dimensional data processed by MBESs [53–55]. A method combining linear and arc length weighted interpolation techniques is adopted, enabling effective meshing of three-dimensional strength data. This approach facilitates the transformation from the polar coordinate system to the Cartesian coordinate system while maintaining consistent resolution within the same dimension.

The schematic diagram of the transmitting sector in the Cartesian coordinate system is shown in Figure 4. Note that the origin of the rectangular and polar coordinate systems coincide. The transmitting sector is symmetric about the yOz plane; i.e., a beam with a receiving angle of 0° is projected on the y-axis. A measuring block is A on the transmitting sector whose geometric center is P_i and radial distance is r. At the same radial distance, the point on the 0° beam that corresponds to point P_i is P_0. The projection of point P_0 onto the z axis is P_z. The projection of arc $\overset{\frown}{P_0 P_i}$ onto the xOy plane is arc $\overset{\frown}{P_0' P_i'}$. Hence, the conversion

relationship between a polar coordinate system and a Cartesian coordinate system can be formulated as:

$$\begin{cases} x = h/\cos\beta \cdot \tan\theta \\ y = h \cdot \tan\beta \\ z = h \end{cases} \tag{13}$$

$$h = R \cdot \cos\theta \cdot \cos\beta \tag{14}$$

where β is the transmitting angle of the sector, θ is the receiving beam angle of measuring block A, h is the distance from point P_z to origin O, and R is the radial distance of measuring block A on the transmitting sector.

From (13) and (14), it can be derived that the coordinate value in Cartesian coordinates is

$$\begin{cases} x = R \cdot \sin\theta \\ y = R \cdot \cos\theta \cdot \sin\beta \\ z = R \cdot \cos\theta \cdot \cos\beta \end{cases} \tag{15}$$

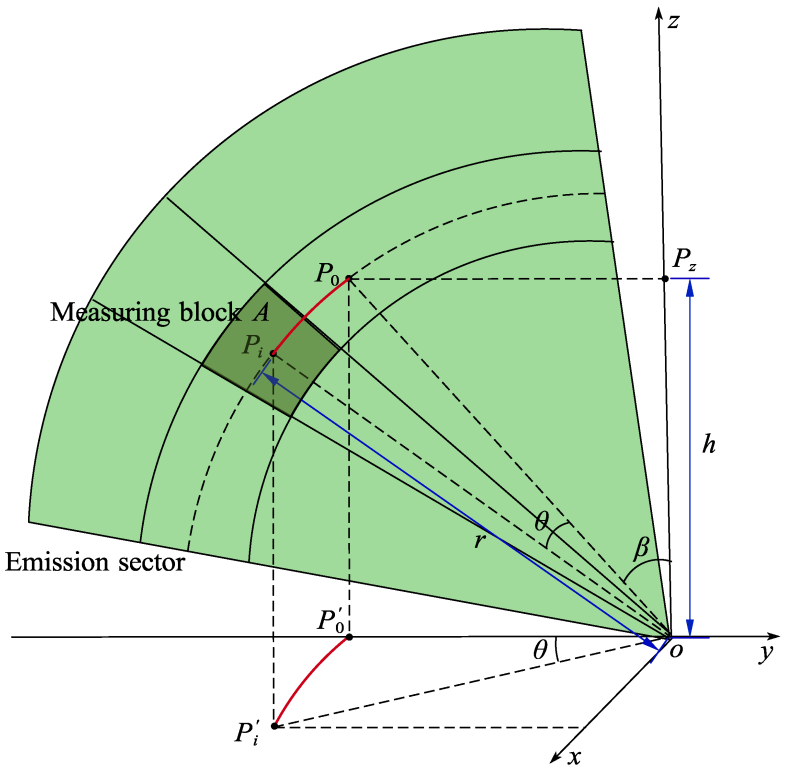


Figure 4. The transmitting sector in the Cartesian coordinate system. The green fan-shaped area represents the transmitting sector.

According to (15), the polar coordinate $P_i(r_{P_i}, \beta_{P_i}, \theta_{P_i})$ can be derived from the Cartesian coordinate $P_i(x, y, z)$. The transmitting angle is $\beta = \{\beta_1, \beta_2, \cdots, \beta_M\}$; within a certain transmitting angle sector, the receiving beamforming angle is $\theta = \{\theta_1, \theta_2, \cdots, \theta_N\}$; and the radial distance of all layers is $r = \{r_1, r_2, \cdots, r_L\}$. M, N, and L represent the number of transmission angles, the number of reception beam angles, and the number of layers into which the radial distance is divided, respectively. The nearest transmitting angle $\{\beta_l, \beta_{l+1}\}$, receiving beam angle $\{\theta_j, \theta_{j+1}\}$, and radial distance $\{r_k, r_{k+1}\}$ are found, respectively, at point $P_i(r_{P_i}, \beta_{P_i}, \theta_{P_i})$, as shown in Figures 5 and 6 . Figure 5 shows the emission sector with an angle of β_i. Thus, the following relation can be obtained:

$$\beta_{\Delta 1} = |\beta_{P_i} - \beta_l|, \quad \beta_{\Delta 2} = |\beta_{P_i} - \beta_{l+1}| \tag{16}$$

$$\theta_{\Delta 1} = \left|\theta_{P_i} - \theta_j\right|, \quad \theta_{\Delta 2} = \left|\theta_{P_i} - \theta_{j+1}\right| \tag{17}$$

$$r_{\Delta 1} = \left|r_{P_i} - r_k\right|, \quad \mathrm{r}_{\Delta 2} = \left|r_{P_i} - r_{k+1}\right| \tag{18}$$

The interpolation process of spatial three-dimensional numbers is as follows: First, the two emission fan surfaces separately are interpolated, and then the two interpolation results are interpolated again. The interpolation diagram for a single emission fan is shown in Figure 5. P_3 and P_6 are the corresponding points of point P_i on the sectors of the two emission angles $\{\beta_i, \beta_{i+1}\}$, respectively. As can be seen from Figure 6, P_3 can be obtained through the interpolation of the surrounding four points A, B, C, and D. From the two points B and C, the acoustic magnitude at P_1 can be obtained by using arc length fixed weight interpolation as follows:

$$\zeta_{P_1} = \frac{\theta_{\Delta 2}}{\theta_{\Delta 1} + \theta_{\Delta 2}}\zeta_B + \frac{\theta_{\Delta 1}}{\theta_{\Delta 1} + \theta_{\Delta 2}}\zeta_C \tag{19}$$

where ζ_B and ζ_C are the acoustic quantity values at points B and C, respectively.

Similarly, the acoustic value at P_2 is

$$\zeta_{P_2} = \frac{\theta_{\Delta 2}}{\theta_{\Delta 1} + \theta_{\Delta 2}}\zeta_A + \frac{\theta_{\Delta 1}}{\theta_{\Delta 1} + \theta_{\Delta 2}}\zeta_D \tag{20}$$

The acoustic value at point P_3 can be obtained through linear interpolation of P_1 and P_2:

$$\zeta_{P_3} = \frac{r_{\Delta 2}}{r_{\Delta 1} + r_{\Delta 2}}\zeta_{P_1} + \frac{r_{\Delta 1}}{r_{\Delta 1} + r_{\Delta 2}}\zeta_{P_2} \tag{21}$$

Referring to the process of P_3, acoustic quantity ζ_{P_6} at P_6 can be calculated.

The spatial positions of P_3 and P_6 are shown in Figure 6. The acoustic quantity expression at P_i can be obtained from the arc length fixed weight interpolation as follows:

$$\zeta_{P_i} = \frac{\beta_{\Delta 2}}{\beta_{\Delta 1} + \beta_{\Delta 2}}\zeta_{P_3} + \frac{\beta_{\Delta 1}}{\beta_{\Delta 1} + \beta_{\Delta 2}}\zeta_{P_6} \tag{22}$$

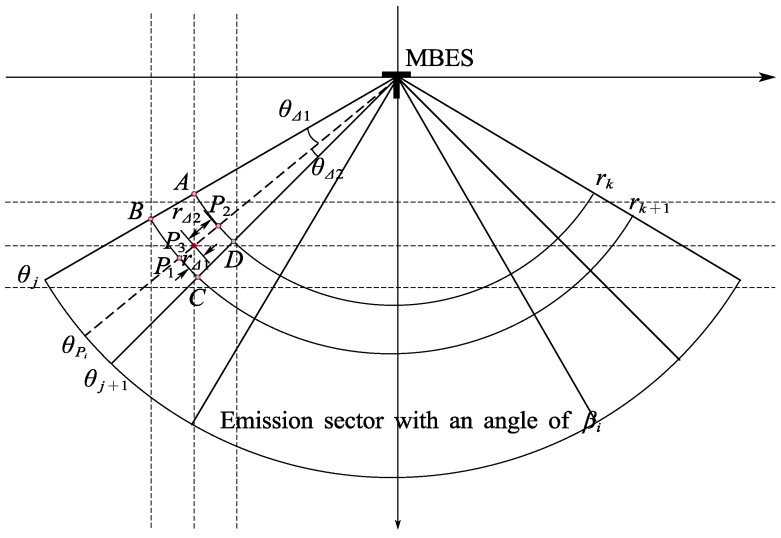


Figure 5. The interpolation diagram on a single transmitting sector. The fan-shaped area represents the transmittint sector with an angle of β_i.

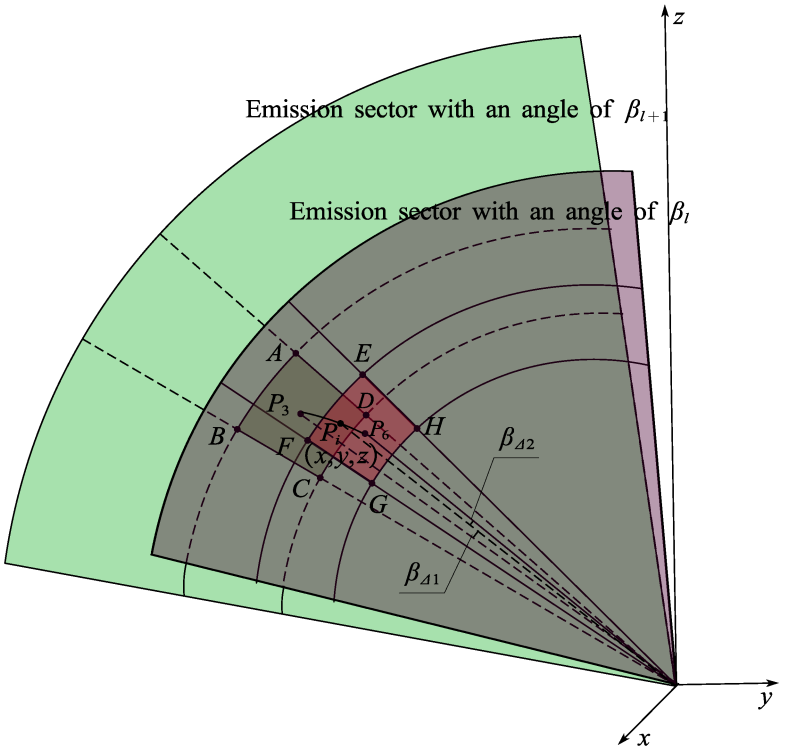


Figure 6. Interpolation between different transmitting sectors. The purple and green fan-shaped areas represent the transmitting sectors with angles β_i and β_{i+1}, respectively.

2.4. Complexity Analysis

For three-dimensional data, the first-order derivative $g(f(\mathbf{x_i}))$ and the second-order derivative $h(f(\mathbf{x_i}))$ along the gradient direction can be precomputed during the preprocessing stage. When the ray direction changes, the precomputed gradient values can be directly utilized to efficiently determine the opacity and color values at the sampling points, thereby accelerating the rendering process.

Assuming the resolution of the visualization image is $M \times N$, the number of rays required for rendering is $M \times N$. The number of sampling points along each ray is K, which is influenced by the size of the three-dimensional data and the sampling step size. As the data volume increases or the sampling step size decreases, K increases correspondingly. Assuming the computational complexity for each ray is $O(K)$, the overall computational complexity is $O(K \times M \times N)$.

Increasing the resolution and the number of sampling points can improve the quality of rendered images. However, this improvement comes at the cost of increased computational resource demands.

2.5. Preprocessing of MBESs

An MBES employs a multi-element transmitting and receiving array to enable wide coverage and high resolution. To minimize the sampling interval between data frames and acquire relatively precise spatial real-time data, the system exclusively gathers original data during the experiment without additional data processing. Consequently, in addition to beamforming, preprocessing is essential for the original data collected by an MBES.

2.5.1. Backscattering Strength

The backscattering cross-section per unit volume is denoted as σ, which reflects the scattering capability of all scatterers within a unit volume. The backscattering strength per unit volume can be derived as follows:

$$BS_V = 10\log_{10}(\sigma) \tag{23}$$

According to the active sonar equation, it can be written as

$$EL = SL - 2TL + BS_V + 10\log_{10}(V) \tag{24}$$

The formula for calculating the source level (SL) is given as follows:

$$SL = 20\log_{10}(V_r \cdot d) - M_s \tag{25}$$

where V_r is the root-mean-square (RMS) voltage received by the standard hydrophone, expressed in volts (V), d denotes the distance between the sound source and the hydrophone, measured in meters (m), and M_s is the sensitivity of the standard hydrophone, expressed in decibels (dB) with a reference value of 1 μPa · m. TL is the transmission loss related to the transducer range (r) and attenuation coefficient (α). Its calculation expression is

$$TL = 20\log_{10}(r) + \alpha r \tag{26}$$

Based on the reflection and projection of the scatterer, as discussed in [56], the backscattering strength per unit volume can be expressed as

$$BS_V = EL + 2TL - SL + \sum_{0}^{i-1}\delta(j)\cdot\left(20\log_{10}(1-|R_I(j)|)\right) - 10\log_{10}\left(\Phi_T^2\cdot\Phi_R^2\right) - 10\log_{10}V \tag{27}$$

where EL denotes the echo level, δ signifies the Dirac function, and $R_I(j)$ is the backscattering coefficient of the ith layer. V denotes the volume of the measurement block. Φ_T and Φ_R represent the directionality of the transmitting and receiving array, respectively.

2.5.2. Multilayered Methodology

The transmitting array of the MBES was configured as a curved array, which resulted in a transmitting sector characterized by wide coverage. Beamforming technology was used to acquire time-domain signals at various beam angles. Layering of time-domain signals was conducted depending on the pulse width of the transmitted signal. Equal window-length layering method is used in the proposed method, as illustrated in Figure 7. The red dashed line illustrates equal-depth layering. The window length of the 0° beam is denoted as w_p. As the beam angle's absolute value increases, the window's length w_c increases correspondingly. In order to balance w_p and w_c, it is necessary to consider the overlap rate between layers and the signal utilization rate within the layers. Equal window-length layering method is used in the proposed method, as illustrated by the blue dots in Figure 7. Beam angles were classified according to uniform processing-window lengths, which facilitates the thorough use of signals within the layer.

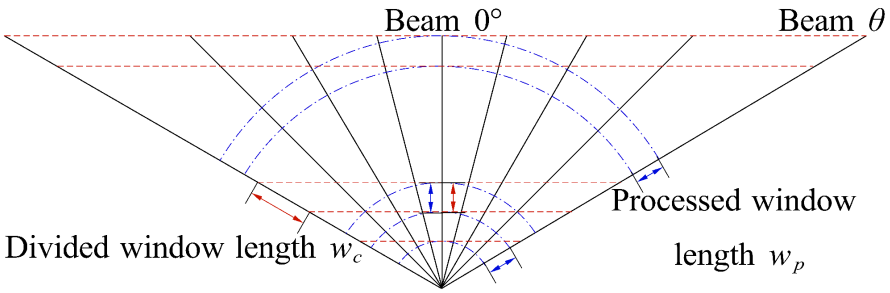


Figure 7. Layered method for selecting different partition window lengths.

The processing-window length depends on the pulse width of the transmitted signal τ, $w_p = \tau$, so the layer thickness over the radial distance is $r_\Delta = c\tau/2$. After beamforming, the layering mode of the time-domain signal is shown in Figure 8. The two coordinate axes were time t and radial distance r, respectively, and the angle between the two coordinate axes was θ, i.e., the time-domain signal waveform with the beam angle of $(90^\circ - \theta)$. T_e represents the interval between two adjacent transmissions, and $t_1 + i\tau$ signifies the time for the MBES to receive the echo signal from the radial distance of r_i. The layer thickness of the unit layer is $r_\Delta = r_i - r_{i-1}(i = 1, 2, \cdots, N)$, where N denotes the number of layers.

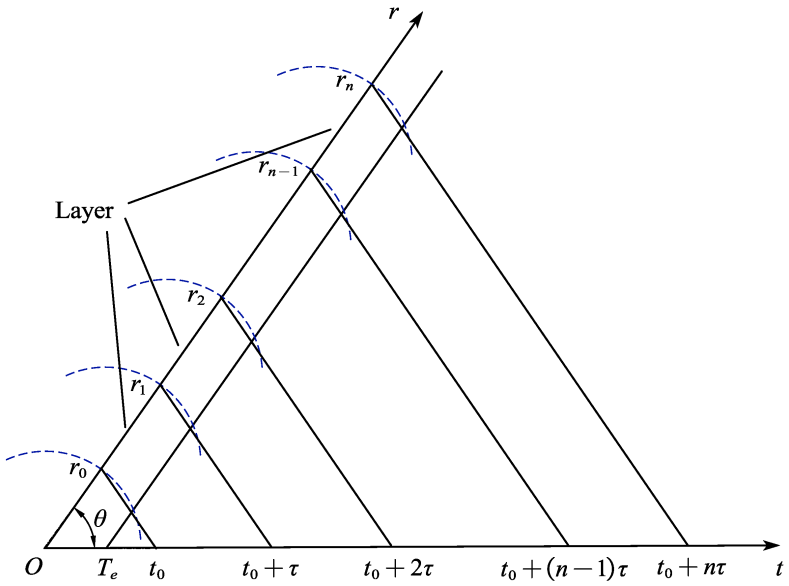


Figure 8. Layered schematic diagram of time-domain signal.

2.5.3. Histogram Equalization

In order to highlight the target area, histogram equalization was used to enhance image contrast [57,58]. Histogram equalization redistributes the pixel values of the original image using a specific transformation. The redistributed pixel values conform to a uniform distribution and enhance the dynamic range of gray values within the original image's pixels.

The probability density function of a digital image with the gray level range $[0, L-1]$ is

$$P(r_k) = \frac{n_k}{N} \quad (0 \le r_k \le L; k = 0, 1, \cdots, L-1) \tag{28}$$

where r_k denotes the k-level gray level, n_k reprensents the number of pixels in the image whose gray level is r_k, and N signifies the total number of pixels in the image. The cumulative distribution function can be calculated as follows:

$$F(r_k) = \sum_{i=0}^{i=k} P(r_i) \quad (k = 0, 1, \cdots L-1, 0 \le F(r_k) \le 1) \tag{29}$$

Each gray level r_k in the original image corresponds to an equalized gray level s_k, which can be written as:

$$s_k = T(r_k) \quad (k = 0, 1, \cdots L-1, 0 \le T(r_k) \le 1) \tag{30}$$

According to probability theory, the cumulative distribution function $F(s_k)$ can be written as

$$F(s_k) = \sum_{j=0}^{j=k} P(s_j) = \sum_{i=0}^{i=k} P(r_i) \quad (k = 0, 1, \cdots L-1, 0 \leq F(s_k) \leq 1) \tag{31}$$

Therefore, the gray level of the equalized image is

$$s_k = T(r_k) = (L-1) \sum_{i=0}^{k} P(r_i) (0 \leq r_i \leq 1; k = 0, 1, \cdots, L-1) \tag{32}$$

2.6. Algorithm Description

In this study, an MBES is fabricated measure water flow fields. The system employs a Mills cross-array configuration, with the transmitter array designed as an arc array and the receiver array as a linear array, as shown in Figure 9. Higher frequency results in faster signal attenuation during propagation, whereas it also provides better system resolution. In designing the MBES, a trade-off between low and high frequencies is made to balance the capability for long-range measurements and high resolution. Consequently, a center frequency of 113.6 kHz is selected.

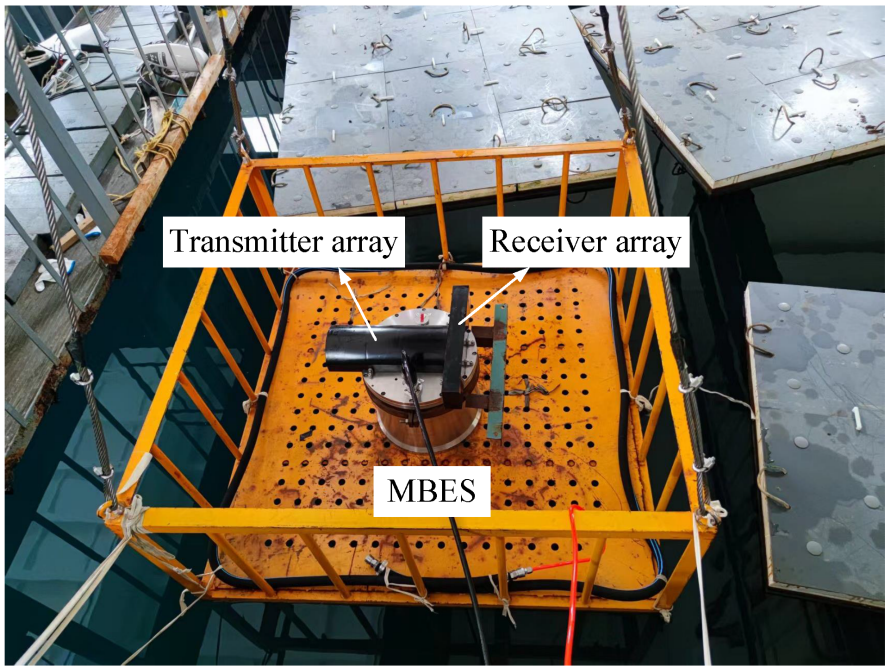


Figure 9. MBES diagram.

The processing workflow of the proposed algorithm is illustrated in Figure 10. Initially, the original data collected by the MBES are processed to obtain the time-domain signals after beamforming. Preprocessing operations are then performed, including layering, backscattering strength computation, and histogram equalization. Finally, three-dimensional data are visualized using the generated opacity transfer function. Section 4 presents a detailed analysis of the simulation performance of the algorithm and the validation results based on experimental data.

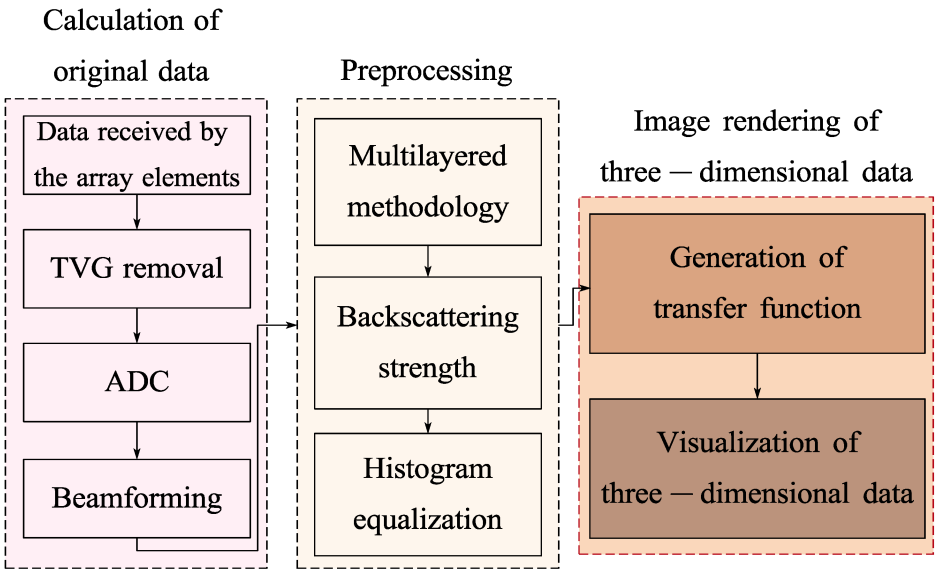


Figure 10. The processing workflow of the proposed algorithm.

3. Numerical Simulation and Test Verification

3.1. Numerical Simulation

3.1.1. Simulation Setup

In three-dimensional modeling, the xyz space is partitioned into multiple measurement blocks, each representing a scatterer utilized to simulate the propagation and reflection properties of sound waves. This paper employs the methodology outlined in [56] to produce echo signals for each array element. The transmission signal consists of four repeated 7-bit Barker codes, comprising ten cycles at a center frequency of 113.6 kHz. According to the Nyquist sampling theorem, the sampling frequency is 769 kHz, and with 48 receiving elements. The parameters for both the simulation and experimental processes are configured identically, with specific details provided in Table 1. The emission angle is $\beta = [-10°, 10°]$, subdivided into 71 segments.

Table 1. Parameter configuration.

Parameter	Center Frequency (kHz)	Sampling Frequency (kHz)	Pulse Width (ms)	Number of Transmit Array Elements	Number of Receive Array Elements
Value	113.6	769	2.5	48	48

As shown in Figure 11, assuming the position of the scatterers is

$$\begin{aligned} &x = x_0, x = x_1,\ x_0 < x_1 \\ &y = y_0, y = y_1,\ y_0 < y_1 \\ &z_0 \le z \le z_1,\ z_0 \ge 0, \end{aligned} \tag{33}$$

Depending on the position of the scatterer, the four critical values of the emission angle can be obtained, which are

$$\beta_{LU} = \arctan(y_0/z_1), \beta_{LD} = \arctan(y_0/z_0) \tag{34}$$

$$\beta_{RU} = \arctan(y_1/z_1), \beta_{RD} = \arctan(y_1/z_0) \tag{35}$$

Here, $\beta_{LU} \leq \beta \leq \beta_{RU}$, and the measurement block for radial distance r contains echoes only at the two measurement blocks corresponding to the beam angles at $\theta_B = \arcsin(x_0/r)$ and $\theta_A = \arcsin(x_1/r)$. When $\beta_{LD} < \beta < \beta_{LU}$, echoes are present in the measurement blocks within the receiving beam angle range of $\theta_B \leq \theta \leq \theta_A$, where $\theta_B = \arctan(x_0/|y_0/\sin\beta|)$ and $\theta_A = \arctan(x_1/|y_0/\sin\beta|)$, with the radial distance of the measurement blocks containing echoes at each angle $r = |y_0/(\sin\beta \cdot \cos\theta)|$. For $\beta_{RD} < \beta < \beta_{RU}$, echoes are present in the measurement blocks within the receiving beam angle range of $\theta_B \leq \theta \leq \theta_A$, specifically at $\theta_B = \arctan(x_0/|y_1/\sin\beta|)$ and $\theta_A = \arctan(x_1/|y_1/\sin\beta|)$, with radial distance $r = |y_1/(\sin\beta \cdot \cos\theta)|$ for the measurement blocks containing echoes at each angle. Throughout the simulation, $x_0 = -4$, $x_1 = 4$, $y_0 = -4$, $y_1 = 4$, $z_0 = 25$, $z_1 = 35$.

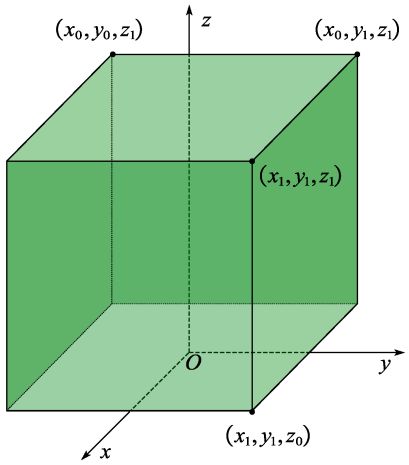


Figure 11. Three-dimensional schematic diagram of the target area during numerical simulation.

3.1.2. Echo Model

The instantaneous backscattering strength of the signals suggested by Medwin and Clay [59] is

$$I(i) = \int_{r_1}^{r_2} \int_{-\pi/2}^{\pi/2} \int_{-\pi/2}^{\pi/2} \frac{I_0 \exp(-2\alpha r(i))}{r^4(i)} \cdot \sigma_{Ibs}(i) \Phi_T^2(\varphi, \theta) \cdot \Phi_R^2(\varphi, \theta) r^2 dr d\varphi d\theta \tag{36}$$

where I_0 represents the emission sound intensity at an equivalent distance of 1 m from the sound source and α denotes the sound absorption coefficient. Throughout the simulation process, we assign $\Phi_T(\varphi, \theta) = 1$ and $\Phi_R(\varphi, \theta) = 1$, and ignore the absorption loss.

The echo level of a microelement can be written as:

$$EL(i) = 10 \log_{10}(I(i)) \tag{37}$$

The expression for the echo signal $r_i(n)$ of the ith element, given the transmitted signal $s(n)$, is

$$r_i(n) = 10^{EL(i)/20} \cdot a(i) s(n - \tau(i)) \exp(j2\pi f_c n) + n_i(n) \tag{38}$$

where $a(i)$ denotes the signal amplitude of the ith element, $\tau(i)$ represents the time delay of the ith element of the receiving transducer, and $n_i(n)$ signifies Gaussian white noise.

The echo signal $y_m(n)$ of the mth element is

$$y_m(n) = \sum_{i=1}^{N} r_i(n - \tau_m(\theta_i)) \tag{39}$$

$$\tau_m(\theta_i) = -(m-1)d\sin\theta_i/c \tag{40}$$

where $\tau_m(\theta_i)$ denotes the time delay difference between the mth element and the reference element for receiving the echo signal from the ith element; d represents the distance between adjacent array elements; and c is the sound velocity in water.

3.1.3. Simulation Results

Under simulation conditions, the beam image for a single transmission angle is shown in Figure 12. Figure 12a and Figure 12b represent two-dimensional images with transmission angles of 0° and −5.7°, respectively. These images are limited to expressing only two-dimensional data information.

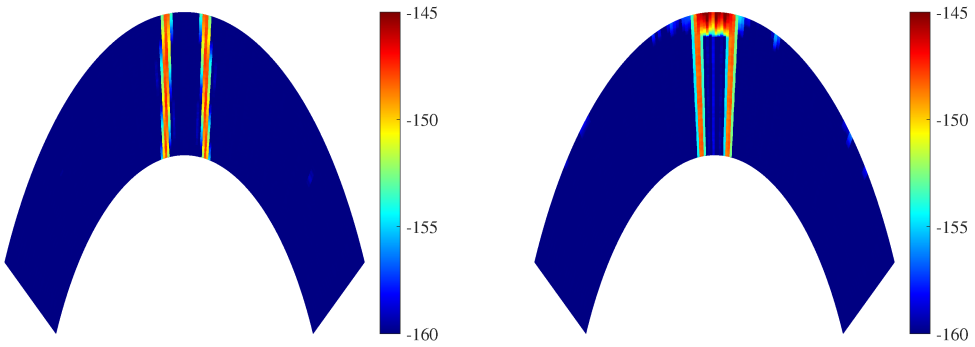


Figure 12. Single-frame beam image. (**a**) Transmission angle of 0°. (**b**) Transmission angle of −5.7°.

A three-dimensional representation of the data was created as a reference image for the algorithm's processing results in the paper—see Figure 13. The color value denotes the normalized gray level of the backscattering strength. The three-dimensional data underwent initial preprocessing, and then critical values in the background and target regions were identified via iterative thresholding. The image was ultimately rendered in a three-dimensional coordinate system. Note that only the delineated target area is shown in the image, while the background remains empty. Figure 13 illustrates that the target area is positioned on all four sides of a rectangular prism with an approximate side length of 8 m, which is consistent with the parameters established in the simulation conditions.

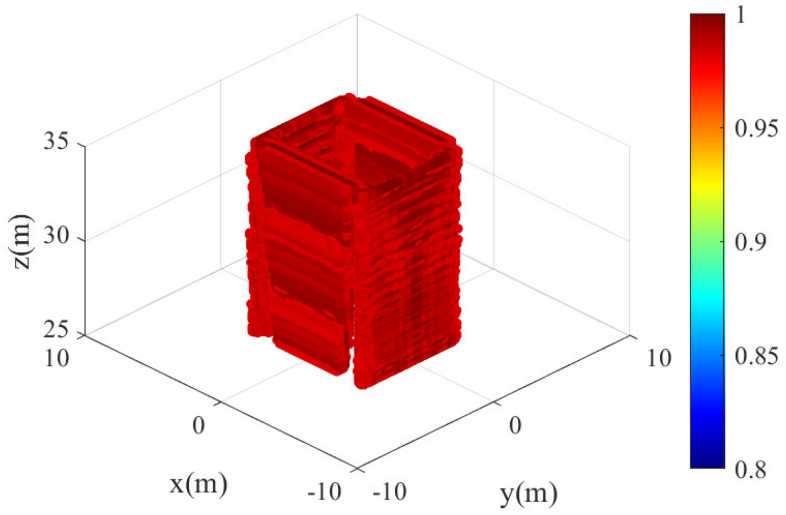


Figure 13. Three-dimensional schematic diagram of the target area during numerical simulation.

The simulation signal was analyzed using the algorithm delineated in the paper, with the results presented in Figures 14 and 15. Figure 14a–c show the joint histogram of data values, gradient amplitude versus data values and the second derivative versus data values, respectively. The background area (A) and the target area (B) are visible in Figure 14a. It is evident from Figure 14b that, despite significant noise in the joint histogram, certain boundaries between the background and the target area (as well as internal boundaries within the target) can be seen. In region A of Figure 14a, the area with lower intensity values reprensents the section not assessed by the MBES during the meshing process. This section is characterized by minimal alterations and low gradient values, which corresponds to A in Figure 14b. The region with elevated intensity values represents the background area assessed by the MBES, while the area exhibiting minimal variations corresponds to B in Figure 14b. At the interface of regions with elevated and diminished intensity values in the background, the intensity values show significant variation and the gradient amplitude is pronounced, which aligns with curve F in Figure 14b. In the target area B of Figure 14a, regions exhibiting minimal variations in intensity values correspond to C, D, and E in Figure 14b. The region with pronounced changes in intensity corresponds to curve G in Figure 14b. The intensity disparity between background area A and target area B in Figure 14a is pronounced, exhibiting a substantial gradient amplitude, which aligns with the contour arc of the entire grayscale region in Figure 14b. Curves H, I, and J in the magnified image of Figure 14b exhibit comparable data values and gradient magnitudes, which can lead to erroneous partitioning during the coloring process. Moreover, the second derivative depicted in Figure 14c can adequately address this category of a boundary assignment problem. Figure 14c shows that curve F aligns with curve F in Figure 14b. At the local maximum gradient amplitude, the second derivative equals zero. The area of boundary attribution ambiguity, depicted in Figure 14b, is clearly shown in Figure 14c.

Figure 15 shows a two-dimensional representation of three-dimensional data from two distinct perspectives, with the orientation angles in Figure 15a being [1, 1, 1] and in Figure 15b being [1, 0, 16]. The results presented in Figure 15 show that the proposed algorithm effectively reconstructs the target area from multiple perspectives while successfully distinguishing the target from the background. A comparative analysis of Figure 15a,b reveals that the target possesses a geometric structure resembling a rectangular prism, with the primary distribution located along its four faces. The intensity profiles in Figure 15a,b further corroborate this observation, showing a peak intensity at the central region that gradually diminishes toward the periphery.

These findings are consistent with the theoretical framework and the three-dimensional reference image provided in Figure 13, thereby validating the algorithm's performance. Additionally, the proposed algorithm offers a significant advantage by eliminating the need for repeated threshold adjustments. The generated two-dimensional image not only isolates the target area with high precision but also preserves the nuanced variations in the target's contours, thereby ensuring the fidelity of the reconstructed imagery.

3.2. Test Validation

Test Scenario

In order to assess the efficacy of the proposed algorithm, the subsequent experiments were devised. The schematic representation of the experimental setup within an anechoic tank is illustrated in Figure 16. The utilized equipment includes an MBES, a nano microporous aeration tube, a polyurethane (PU) tube, a speed regulating valve, an air compressor, and a mounting frame. We positioned the aeration tube and MBES within the mounting frame, as illustrated in Figure 17. The MBES was centrally located within the mounting frame. In contrast, the aeration tube was secured along the bottom edge of the frame.

The size of the mounting frame was 1.2 m × 1.2 m, and it was situated at the base of the pool. The PU tube was linked to a central speed-regulating valve to modulate the size and quantity of bubbles. The air compressor was utilized to produce bubbles. As shown in Figure 16, we positioned the placement frame at the bottom of the pool, activated the air compressor, adjusted the speed control valve to a specified setting, and the aeration tube produced high-density microbubbles.

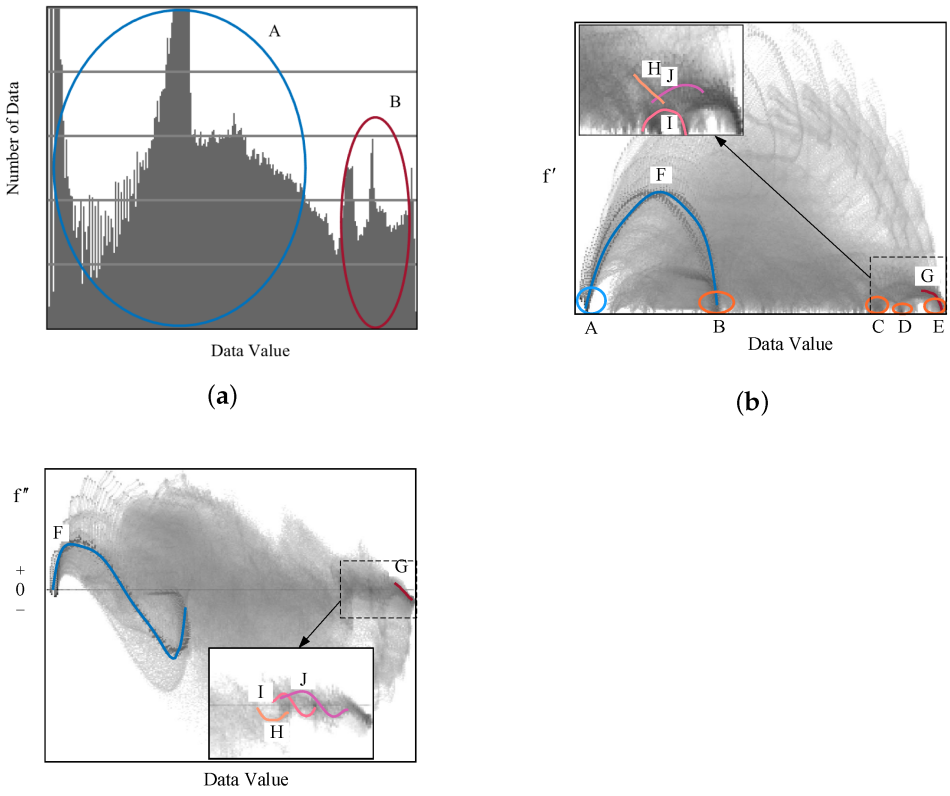


Figure 14. Statistical chart. (**a**) Histogram of data value. (**b**) Gradient magnitude f' versus data value f. (**c**) Second derivative f'' versus data value f.

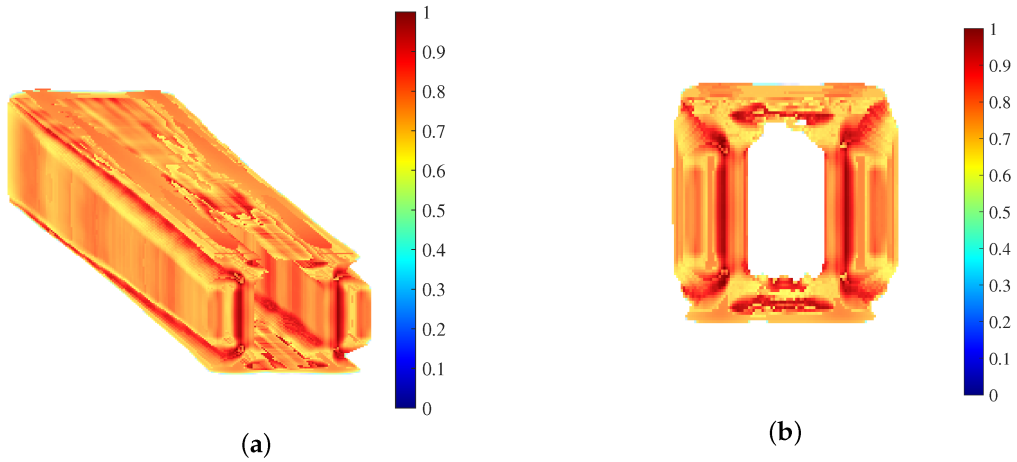


Figure 15. Two-dimensional representation of three-dimensional data. (**a**) Orientation angle of [1, 1, 1]. (**b**) Orientation angle of [1, 0, 16].

The emission angle is $\beta = [-27.5, -21.7, -16.1, -10.6, -5.3, 0, 5.3, 10.6, 16.1, 21.7, 27.5]$. The experimental procedure is as follows: First, start the air compressor and control the generation speed of bubbles by adjusting the speed control valve. After several minutes

of operation, the quantity and motion velocity of the bubbles within the tank gradually stabilize, as shown in Figure 17. In this steady-state condition, the emission angle is altered to acquire three-dimensional echo data.

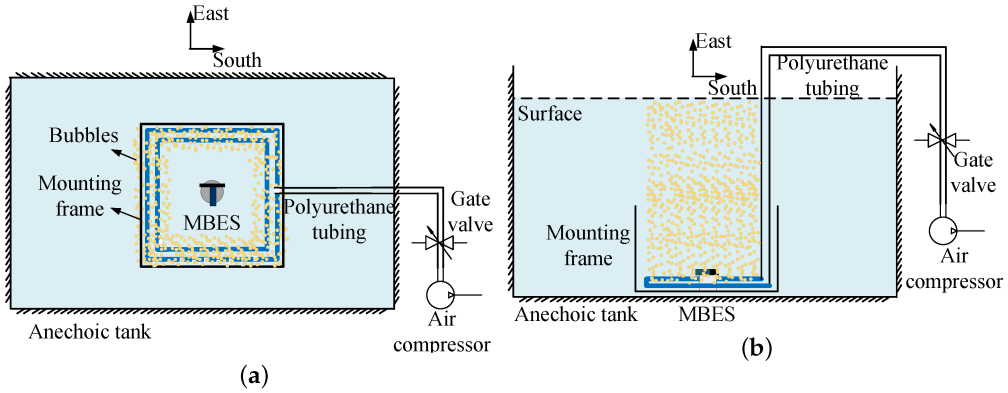


Figure 16. Schematic representation of the experimental setup within the anechoic tank. (**a**) Top view. (**b**) Lateral view.

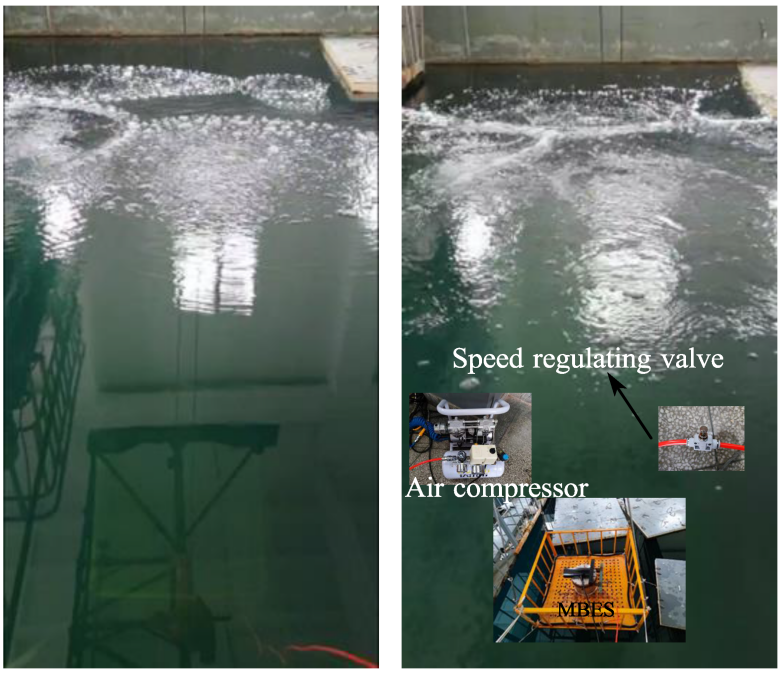
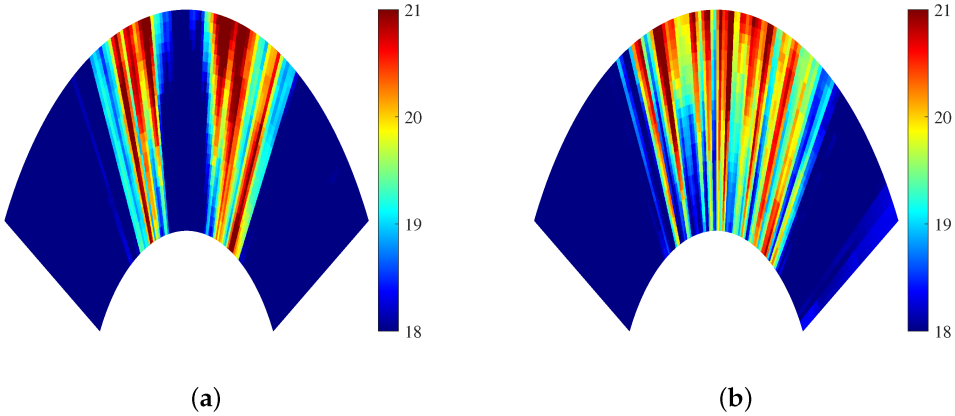


Figure 17. Test equipment and scenes in an anechoic tank.

3.3. Test Results

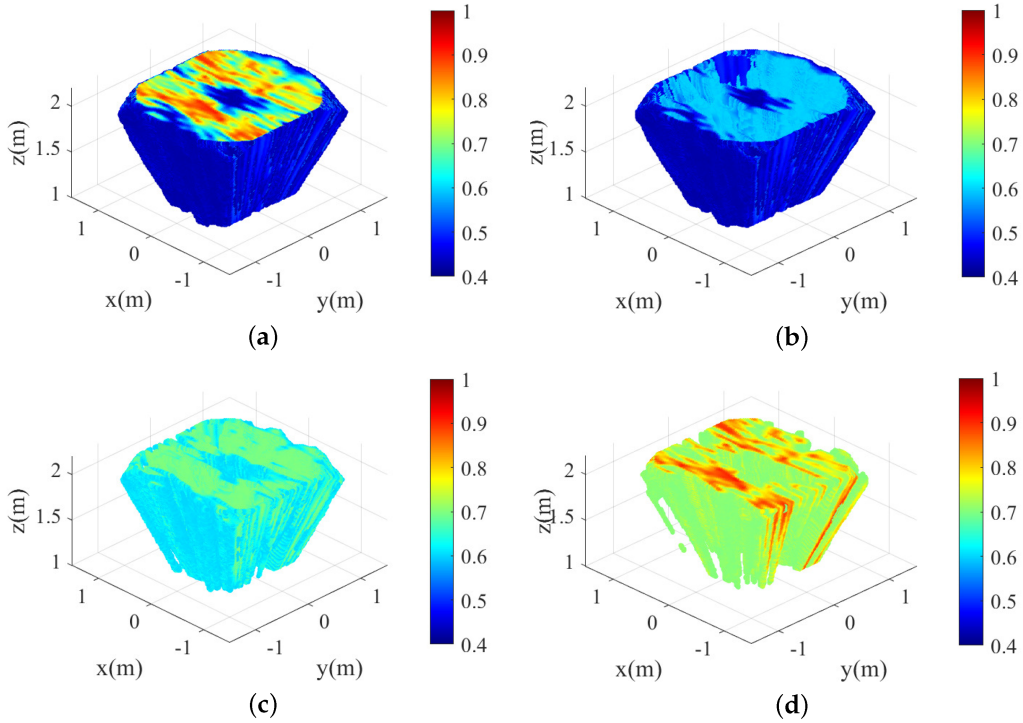
The single-frame beam image obtained from the MBES is shown in Figure 18. Figure 18a,b represent two-dimensional images with transmission angles of 0° and −10.6°, respectively. These images are limited to expressing only two-dimensional data information.

Figure 19 shows the three-dimensional intensity maps obtained by scanning emission angles. Figure 19b–d are intensity maps obtained by splitting the grayscale levels from Figure 19a. The grayscale ranges in Figure 19b–d are from 0.4 to 0.6, from 0.6 to 0.7, and from 0.7 to 1, respectively. Figure 19 shows that, along the z-axis, the range of color blocks increase, which is consistent with the diffusion during the bubble-ascent process in the test scenario. Note that the square formed by $x = \pm 0.5$ m and $y = \pm 0.5$ m shows more color blocks near the edges; the closer it is to the edges of the square, the darker the color. Along the z-axis, the color becomes darker, and the range of color blocks increases. This is consistent with the result of bubbles generated by the aeration tube.



(**a**) (**b**)

Figure 18. Single-frame beam image. (**a**) Transmission angle of 0°. (**b**) Transmission angle of −10.6°.



(**a**) (**b**) (**c**) (**d**)

Figure 19. Three-dimensional intensity maps obtained by scanning the emission angle. (**a**) All data. (**b**) Grayscale range is from 0.4 to 0.6. (**c**) Grayscale range is from 0.6 to 0.7. (**d**) Grayscale range is from 0.7 to 1.

The test data were analyzed using the algorithm delineated in the paper—see Figures 20 and 21. Figure 20a–c show the joint histogram of data values, gradient amplitude versus data values and the second derivative versus data values, respectively. Figure 20a shows the background area (A) and the target area (B). The regions of background area A in Figure 20a show minimal variations in intensity values correspond to A, B, and C in Figure 20b. The boundary between regions with higher and lower intensity values corresponds to curves D and E in Figure 20b. In target areas B and C in Figure 20a, the parts exhibiting substantial variations in intensity values align with curve F in Figure 20b. Futhermore, in the magnified image of Figure 20b, certain regions exhibit data value ranges and gradient sizes of curves G, H, and I that are relatively similar and can lead to partitioning errors during the coloring process. Figure 20c shows that curves D, E, and F correspond to curves D, E, and F in Figure 20b. At the local maximum gradient amplitude point, the second derivative equals zero, and the uncertain boundary attribution depicted in Figure 20b is clarified in Figure 20c.

Figure 21 shows the two-dimensional visualization of the three-dimensional water body data processed by the proposed method. It showcases two different perspectives of the three-dimensional data, with orientation angles of [1, 1, 1] and [1, 0, 16], respectively. From Figure 21, it is evident that the target region is successfully separated from the background. Combining Figure 21a,b, it can be inferred that the geometric shape of the target area is approximately that of a hollow rectangular prism, with the target distributed across the four adjacent faces. Furthermore, Figure 21a shows that the target region gradually expands, which is consistent with the upward motion of the generated bubbles, in accordance with their expected behavior. From Figure 21a,b, it can be seen that the intensity of the target is maximal at the central positions of the four faces, gradually decreasing towards the sides. This result is consistent with the phenomenon where bubbles, during their upward movement, simultaneously diffuse outward to the sides. During the test, two sides of the aerator tube that generate the bubble curtain produce relatively stronger airflow, resulting in the generation of a greater number of bubbles. Accordingly, in Figure 21a,b, it can be seen that two of the four faces where the target is located exhibit higher intensities. Some of the conclusions can be further validated by referring to Figure 19.

In summary, the two-dimensional images obtained using the method proposed in this paper not only successfully separate the bubble region from the water body, but also, from different viewpoints, reveal both the intensity variations and the contours within the target region.

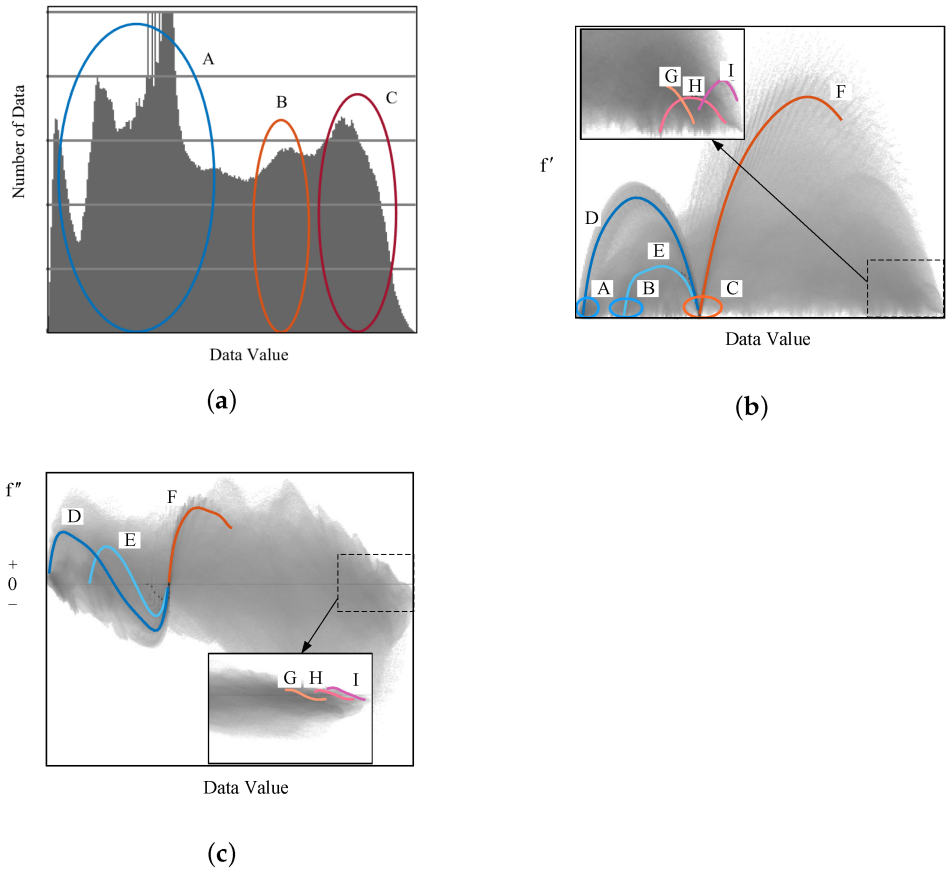
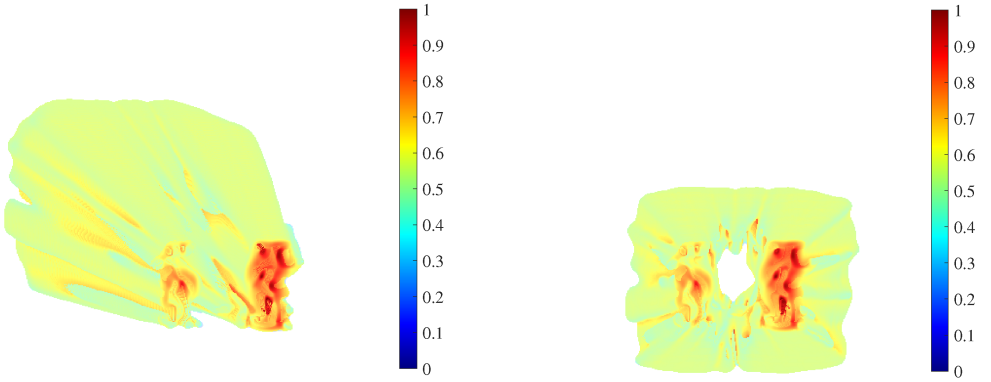


Figure 20. Statistical chart. (**a**) Histogram of data value. (**b**) Gradient magnitude f' versus data value f. (**c**) Second derivative f'' versus data value f.



(**a**) (**b**)

Figure 21. Two-dimensional representation of three-dimensional data. (**a**) Orientation angle of [1, 1, 1]. (**b**) Orientation angle of [1, 0, 16].

4. Discussion and Conclusions

To advance underwater imaging for marine environment monitoring, this paper presents a visualization method for three-dimensional non-uniformly sampled data acquired from MBESs. The proposed method achieves accurate coordinate system transformation while effectively addressing the challenges of beam direction resolution inconsistencies. Additionally, the visualization of three-dimensional data suppresses background and non-interest regions while highlighting boundaries and target areas. Compared to the single-frame beam images obtained from the MBES, the proposed algorithm provides a more intuitive and comprehensive representation of flow field information within three-dimensional data. Furthermore, in contrast to three-dimensional imaging techniques, the proposed method not only captures more detailed and enriched features but also supports multi-perspective imaging capabilities. This study holds significant implications for marine exploration and environmental monitoring while also laying the foundation for further advancement of MBESs. Simulation and experimental results demonstrate that by combining data values with the first- and second-order derivatives in the gradient direction through a joint histogram analysis, the proposed algorithm effectively identifies the critical thresholds between bubble and background regions, as well as detailed features within the target area. The visualized images further enable the inference of the geometric characteristics of bubble regions, which exhibit an approximate hollow rectangular prism shape with bubbles primarily distributed along four adjacent sides. Moreover, the intensity variations reflect the dynamic processes of bubble generation, movement, and diffusion.

However, the algorithm requires traversing all voxels along each ray, increasing computational complexity. By optimizing the algorithm design to improve computational efficiency or reduce complexity, the performance and practicality of the algorithm can be enhanced, thereby better meeting the requirements of practical applications.

Author Contributions: Conceptualization, W.C. and S.F.; methodology, W.C. and S.F.; software, W.C. and S.F.; investigation, W.C., S.F. and C.Z.; writing—original draft preparation, W.C., M. F. and Y.Z.; writing—review and editing, W.C., C.Z., M.F. and Y.Z.; supervision, W.C., M.F., Y.Z. and H.C.; funding acquisition, S.F. and C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant No. 2242024k30004, in part by the National Natural Science Funds of China under Grant No. 12304497.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors thank the anonymous reviewers for their careful reading and valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nylund, A.T.; Arneborg, L.; Tengberg, A.; Mallast, U.; Hassellöv, I.M. In situ observations of turbulent ship wakes and their spatiotemporal extent. *Ocean Sci.* **2021**, *17*, 1285–1302. [CrossRef]
2. Murino, V.; Trucco, A. Three-dimensional image generation and processing in underwater acoustic vision. *Proc. IEEE* **2000**, *88*, 1903–1946. [CrossRef]
3. Tian, Y.; Lan, L.; Guo, H. A review on the wavelet methods for sonar image segmentation. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420936091. [CrossRef]
4. Pailhas, Y.; Petillot, Y.; Capus, C. High-resolution sonars: What resolution do we need for target recognition? *EURASIP J. Adv. Signal Process.* **2010**, *2010*, 205095. [CrossRef]
5. Colbo, K.; Ross, T.; Brown, C.; Weber, T. A review of oceanographic applications of water column data from multibeam echosounders. *Estuar. Coast. Shelf Sci.* **2014**, *145*, 41–56. [CrossRef]
6. Dong, Z.; Liu, Y.; Yang, L.; Feng, Y.; Ding, J.; Jiang, F. Artificial reef detection method for multibeam sonar imagery based on convolutional neural networks. *Remote Sens.* **2022**, *14*, 4610. [CrossRef]
7. Zhang, W.; Zhou, T.; Li, J.; Xu, C. An efficient method for detection and quantitation of underwater gas leakage based on a 300-kHz multibeam sonar. *Remote Sens.* **2022**, *14*, 4301. [CrossRef]
8. Lanzoni, J.C.; Weber, T.C. High-resolution calibration of a multibeam echo sounder. In Proceedings of the OCEANS 2010 MTS/IEEE SEATTLE, Seattle, WA, USA, 20–23 September 2010; pp. 1–7.
9. Li, H.; Li, S.; Chen, B.; Xu, C.; Zhu, J.; Du, W. Research on ship wake acoustic imaging based on multi-beam sonar. In Proceedings of the 2014 Oceans—St. John's, St. John's, NL, Canada, 14–19 September 2014; pp. 1–5.
10. Li, H. Shallow water high resolution multi-beam echo sounder. In Proceedings of the OCEANS 2008-MTS/IEEE Kobe Techno-Ocean, Kobe, Japan, 8–11 April 2008; pp. 1–5.
11. Praet, N.; Collart, T.; Ollevier, A.; Roche, M.; Degrendele, K.; De Rijcke, M.; Urban, P.; Vandorpe, T. The potential of multibeam sonars as 3D turbidity and SPM monitoring tool in the North Sea. *Remote Sens.* **2023**, *15*, 4918. [CrossRef]
12. Urban, P.; Ko, K. Processing of multibeam water column image data for automated bubble/seep detection and repeated mapping. *Limnol-Ocean.-Meth* **2017**, *15*, 1–21. [CrossRef]
13. Xu, C.; Wu, M.; Zhou, T.; Li, J.; Du, W.; Zhang, W.; White, P.R. Optical flow-based detection of gas leaks from pipelines using multibeam water column images. *Remote Sens.* **2020**, *12*, 119. [CrossRef]
14. Zhao, J.; Meng, J.; Zhang, H.; Wang, S. Comprehensive detection of gas plumes from multibeam water column images with minimisation of noise interferences. *Sensors* **2017**, *17*, 2755. [CrossRef]
15. Weber, T.C.; Lyons, A.P.; Bradley, D.L. An estimate of the gas transfer rate from oceanic bubbles derived from multibeam sonar observations of a ship wake. *J. Geophys. Res. Oceans* **2005**, *110*, C04005. [CrossRef]
16. Wilson, D.S.; Leifer, I.; Maillard, E. Megaplume bubble process visualization by 3D multibeam sonar mapping. *Mar. Pet. Geol.* **2015**, *68*, 753–765. [CrossRef]
17. Francisco, F.; Carpman, N.; Dolguntseva, I.; Sundberg, J. Use of multibeam and dual-beam sonar systems to observe cavitating flow produced by ferryboats: In a marine renewable energy perspective. *J. Mar. Sci. Eng.* **2017**, *5*, 30. [CrossRef]
18. Tang, Q.; Li, J.; Ding, D.; Ji, X.; Li, N.; Yang, L.; Sun, W. Deep-sea seabed sediment classification using finely processed multibeam backscatter intensity data in the southwest Indian Ridge. *Remote Sens.* **2022**, *14*, 2675. [CrossRef]
19. Eleftherakis, D.; Berger, L.; Le Bouffant, N.; Pacault, A.; Augustin, J.M.; Lurton, X. Backscatter calibration of high-frequency multibeam echosounder using a reference single-beam system, on natural seafloor. *Mar. Geophys. Res.* **2018**, *39*, 55–73. [CrossRef]
20. Roszkowiak, L.; Korzynska, A.; Zak, J.; Pijanowska, D.; Swiderska-Chadaj, Z.; Markiewicz, T. Survey: Interpolation methods for whole slide image processing. *J. Microsc.* **2017**, *265*, 148–158. [CrossRef]
21. Nürnberger, G.; Zeilfelder, F. Developments in bivariate spline interpolation. *J. Comput. Appl. Math.* **2000**, *121*, 125–152. [CrossRef]
22. Wang, Y.; Liu, X.; Liu, R.; Zhang, Z. Research Progress on Spatiotemporal Interpolation Methods for Meteorological Elements. *Water* **2024**, *16*, 818. [CrossRef]
23. Berger, M.; Li, J.; Levine, J.A. A generative model for volume rendering. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 1636–1650. [CrossRef] [PubMed]
24. De Boer, M.; Hesser, J.; Gröpl, A.; Günther, T.; Poliwoda, C.; Reinhart, C.; Männer, R. Evaluation of a real-time direct volume rendering system. *Comput. Graph.* **1997**, *21*, 189–198. [CrossRef]
25. Mady, A.S.; Abou El-Seoud, S. An overview of volume rendering techniques for medical imaging. *Int. J. Onl. Eng.* **2020**, *16*, 95–106. [CrossRef]

26. Semwal, S.K.; Swann, P.G. Applying ray casting to the flow visualization data using linear and B-spline hyperpatch interpolation. *J. Visual. Comput. Animat.* **1995**, *6*, 33–47. [CrossRef]
27. Liu, Y.; Xue, S.; Qu, G.; Lu, X.; Qi, K. Influence of the ray elevation angle on seafloor positioning precision in the context of acoustic ray tracing algorithm. *Appl. Ocean Res.* **2020**, *105*, 102403. [CrossRef]
28. Goel, V.; Mukherjee, A. An optimal parallel algorithm for volume ray casting. *Vis. Comput.* **1996**, *12*, 26–39. [CrossRef]
29. Binyahib, R.; Peterka, T.; Larsen, M.; Ma, K.L.; Childs, H. A scalable hybrid scheme for ray-casting of unstructured volume data. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 2349–2361. [CrossRef] [PubMed]
30. Ke, H.R.; Chang, R.C. Sample buffer: A progressive refinement ray-casting algorithm for volume rendering. *Comput. Graph.* **1993**, *17*, 277–283. [CrossRef]
31. Ledergerber, C.; Guennebaud, G.; Meyer, M.; Bacher, M.; Pfister, H. Volume MLS ray casting. *IEEE Trans. Vis. Comput. Graph.* **2008**, *14*, 1372–1379. [CrossRef] [PubMed]
32. Thomas, A.R.; Kotadiya, N.B.; Wang, B.; Dhakal, T.P. Physical vapor deposition simulator by graphical processor unit ray casting. *J. Vac. Sci. Technol. B* **2023**, *41*, 064201. [CrossRef]
33. Udupa, J.; Odhner, D. Shell rendering. *IEEE Comput. Grap. Appl.* **1993**, *13*, 58–67. [CrossRef]
34. Knoll, A.M.; Wald, I.; Hansen, C.D. Coherent multiresolution isosurface ray tracing. *Vis. Comput.* **2009**, *25*, 209–225. [CrossRef]
35. Levoy, M. Efficient ray tracing of volume data. *ACM Trans. Graph.* **1990**, *9*, 245–261. [CrossRef]
36. Huang, C.; Lan, Y.; Zhang, G.; Xu, G.; Jiang, L.; Zeng, N.; Tan, J.; Ng, E.Y.K.; Cheng, Y.; Han, N.; et al. A new transfer function for volume visualization of aortic stent and its application to virtual endoscopy. *ACM Trans. Multimed. Comput. Commun. Appl.* **2020**, *16*, 1–14. [CrossRef]
37. Ljung, P.; Krüger, J.; Groller, E.; Hadwiger, M.; Hansen, C.D.; Ynnerman, A. State of the art in transfer functions for direct volume rendering. *Comput. Graph. Forum* **2016**, *35*, 669–691. [CrossRef]
38. Mehaboobathunnisa, R.; Thasneem, A.H.; Sathik, M.M. Fuzzy mutual information-based intraslice grouped ray casting. *J. Intell. Syst.* **2019**, *28*, 77–86. [CrossRef]
39. Lundstrom, C.; Ljung, P.; Ynnerman, A. Local histograms for design of transfer functions in direct volume rendering. *IEEE Trans. Vis. Comput. Graph.* **2006**, *12*, 1570–1579. [CrossRef]
40. Sereda, P.; Bartroli, A.; Serlie, I.; Gerritsen, F. Visualization of boundaries in volumetric data sets using LH histograms. *IEEE Trans. Vis. Comput. Graph.* **2006**, *12*, 208–218. [CrossRef] [PubMed]
41. Kniss, J.; Kindlmann, G.; Hansen, C. Interactive volume rendering using multi-dimensional transfer functions and direct manipulation widgets. In Proceedings of the Proceedings Visualization, 2001, VIS '01, San Diego, CA, USA, 21–26 October 2001; pp. 255–562.
42. Kniss, J.; Kindlmann, G.; Hansen, C. Multidimensional transfer functions for interactive volume rendering. *IEEE Trans. Vis. Comput. Graph.* **2002**, *8*, 270–285. [CrossRef]
43. Park, S.; Budge, B.; Linsen, L.; Hamann, B.; Joy, K. Multi-dimensional transfer functions for interactive 3D flow visualization. In Proceedings of the 12th Pacific Conference on Computer Graphics and Applications, Seoul, Republic of Korea, 6–8 October 2004; pp. 177–185.
44. Kindlmann, G.; Durkin, J. Semi-automatic generation of transfer functions for direct volume rendering. In Proceedings of the IEEE Symposium on Volume Visualization (Cat. No.989EX300), Research Triangle Park, NC, USA, 19–20 October 1998; pp. 79–86.
45. Bozorgi, M.; Lindseth, F. GPU-based multi-volume ray casting within VTK for medical applications. *Int. J. Comput. Assist. Radiol. Surg.* **2015**, *10*, 293–300. [CrossRef]
46. Ping, L.; Tiechang, M.; Xiangzhao, X.; Tianbao, M. A GPU parallel staircase finite difference mesh generation algorithm based on the ray casting method. *Explos. Shock Waves* **2020**, *40*, 024201-1. [CrossRef]
47. Roessler, F.; Botchen, R.P.; Ertl, T. Dynamic shader generation for GPU-based multi-volume ray casting. *IEEE Comput. Graph. Appl.* **2008**, *28*, 66–77. [CrossRef] [PubMed]
48. Sweezy, J.E. A Monte Carlo volumetric-ray-casting estimator for global fluence tallies on GPUs. *J. Comput. Phys.* **2018**, *372*, 426–445. [CrossRef]
49. Levoy, M. Display of surfaces from volume data. *IEEE Comput. Grap. Appl.* **1988**, *8*, 29–37. [CrossRef]
50. Levoy, M. A hybrid ray tracer for rendering polygon and volume data. *IEEE Comput. Grap. Appl.* **1990**, *10*, 33–40. [CrossRef]
51. Max, N. Optical models for direct volume rendering. *IEEE Trans. Vis. Comput. Graph.* **1995**, *1*, 99–108. [CrossRef]
52. Marsden, J.E.; Tromba, A.J. *Vector Calculus*; Macmillan: New York, NY, USA, 2003.
53. Ali, I.; Amat, S.; Trillo, J.C. Point values hermite multiresolution for non-smooth noisy signals. *Computing* **2006**, *77*, 223–236. [CrossRef]
54. Plata, S.A. A review on the piecewise polynomial harmonic interpolation. *Appl. Numer. Math.* **2008**, *58*, 1168–1185. [CrossRef]
55. Schafer, R.; Rabiner, L. A digital signal processing approach to interpolation. *Proc. IEEE* **1973**, *61*, 692–702. [CrossRef]
56. Cao, W.; Fang, S.; Gu, Z.; Feng, M.; Cao, H. Measurement of scattering properties of water body by using a multibeam echosounder system. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 6501112. [CrossRef]

57. Abdullah-Al-Wadud, M.; Kabir, M.H.; Dewan, M.A.A.; Chae, O. A dynamic histogram equalization for image contrast enhancement. *IEEE Trans. Consum. Electron.* **2007**, *53*, 593–600. [CrossRef]
58. Stark, J.A. Adaptive image contrast enhancement using generalizations of histogram equalization. *IEEE Trans. Image Process.* **2000**, *9*, 889–896. [CrossRef]
59. Medwin, H.; Clay, C.S. *Fundamentals of Acoustical Oceanography*; Academic Press: San Diego, CA, USA, 1997.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

A Robust Direction-of-Arrival (DOA) Estimator for Weak Targets Based on a Dimension-Reduced Matrix Filter with Deep Nulling and Multiple-Measurement-Vector Orthogonal Matching Pursuit

Shoudong Wang, Haozhong Wang *, Zhaoxiang Bian, Susu Chen, Penghua Song, Bolin Su and Wei Gao

School of Marine Technology, Ocean University of China, Qingdao 266100, China; wsd@stu.ouc.edu.cn (S.W.); bzx@stu.ouc.edu.cn (Z.B.); css4747@stu.ouc.edu.cn (S.C.); sph2020@stu.ouc.edu.cn (P.S.); subolin@stu.ouc.edu.cn (B.S.); gaowei@ouc.edu.cn (W.G.)
* Correspondence: coolice@ouc.edu.cn

Abstract: In the field of target localization, improving direction-of-arrival (DOA) estimation methods for weak targets in the context of strong interference remains a significant challenge. This paper presents a robust DOA estimator for localizing weak signals of interest in an environment with strong interfering sources that improve passive sonar DOA estimation. The presented estimator combines a multiple-measurement-vector orthogonal matching pursuit (MOMP) algorithm and a dimension-reduced matrix filter with deep nulling (DR-MFDN). Strong interfering sources are adaptively suppressed by employing the DR-MFDN, and the beam-space passband robustness is improved. In addition, Gaussian pre-whitening is used to prevent noise colorization. Simulations and experimental results demonstrate that the presented estimator outperforms a conventional estimator based on a dimension-reduced matrix filter with nulling (DR-MFN) and the multiple signal classification algorithm in terms of interference suppression and localization accuracy. Moreover, the presented estimator can effectively handle short snapshots, and it exhibits superior resolution by considering the signal sparsity.

Keywords: matrix filter; DOA estimation; weak target; pre-whitening operation; sparse representation

Academic Editors: Xuebo Zhang, Haiyong Xu, Jingyue Lu, Lei Zhang, Marc Pinto and Farman Ali

Received: 3 December 2024
Revised: 24 January 2025
Accepted: 27 January 2025
Published: 30 January 2025

Citation: Wang, S.; Wang, H.; Bian, Z.; Chen, S.; Song, P.; Su, B.; Gao, W. A Robust Direction-of-Arrival (DOA) Estimator for Weak Targets Based on a Dimension-Reduced Matrix Filter with Deep Nulling and Multiple-Measurement-Vector Orthogonal Matching Pursuit. *Remote Sens.* **2025**, *17*, 477. https://doi.org/10.3390/rs17030477

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Matrix filters [1–4] have been widely used in sonar and radar applications because of their exceptional data filtering and signal separation capabilities in the spatial domain. A matrix filter projects the observation space onto a subspace by linear transformation, which passes the signals in the sector of interest (passband) while suppressing the out-of-sector (stopband) interference. Vaccaro and Harrison [1] designed a conventional matrix filter (CMF) in the frequency domain. MacInnes [5] improved passband robustness in the beam space by applying the least mean square criterion and solving it using convex optimization. However, this method fails to control the stopband attenuation (SA). Yan et al. [6,7] proposed a second-order cone programming model for CMFs under the stopband constraint, a passband least-squares criterion, and passband minimax criterion. Han et al. [8] proposed three optimization problems using the Lagrange multiplier method to obtain matrix filters with adjustable SA and passband robustness. Hassanien et al. [9] proposed a dimension-reduced conventional matrix filter by constraining the stopband and minimizing the difference between actual and quiescent response matrix filters. These variants of the CMF approach can strictly control the attenuation of the stopband while

maintaining the passband performance. However, data cannot be effectively filtered when strong interfering sources are present in the environment. To solve the problem of weak target direction-of-arrival (DOA) estimation in an environment with strong interfering sources, adaptive matrix filters that can provide nulls toward the direction of the interfering sources and attenuate them adaptively were proposed [9–11]. The presence of noise and perturbation limits the depth of the nulls in the direction of the interfering sources and degrades the passband's robustness. Considering the effect of noise after matrix filtering, Yan [7] limited the power of the noise to reduce its impact on subsequent processing. Nevertheless, non-orthogonal matrix filters transform white noise into colored noise; therefore, the estimation of weak targets is challenging. Hassanien [9] prevented the transformation of noise into colored noise by using an orthogonal pre-whitening (OP) operation. However, this method alters the beam-space characteristics of the matrix filter, particularly by reducing the SA, leading to insufficient interference suppression. To maintain better spatial filtering performance and orthogonality of the matrix filter, Wang [12,13] achieved better performance by synthesizing a set of orthogonal eigenbeams for spatial filtering.

Moreover, the multiple signal classification (MUSIC) algorithm [14], typically employed for DOA estimation with matrix filters, has exhibited poor performance in scenarios with coherent sources or short snapshots [7–9,15–17]. Recently, sparse reconstruction techniques have demonstrated great advantages in solving signal recovery problems and have been widely used in the field of array signal processing [18–25], for example, the sparse spectrum fitting (SpSF) algorithm [26] and matching pursuit algorithms [18–20]. Yang [10,27] proposed estimators based on the SpSF algorithm and matrix filters. However, with these methods, noise power must be estimated. Moreover, determining the selection of the regularization parameter is challenging in practice.

This paper aims to introduce an improved DOA estimator for weak targets. First, considering that white noise at the input may be transformed into colored noise, this study utilizes a Gaussian matrix for the pre-whitening operation. The objective is to prevent the transformation of white noise into colored noise while preserving the characteristics of the beam space. Second, the DOA estimation method is presented based on matrix filters and the multiple-measurement-vector orthogonal matching pursuit (MOMP) algorithm [28]. The MOMP is an extended form of the orthogonal matching pursuit (OMP) algorithm [29] under the multiple-measurement-vector (MMV) model. Next, based on the design method of a dimension-reduced matrix filter with nulling (DR-MFN) proposed in [9], a dimension-reduced matrix filter with deep nulling (DR-MFDN) is presented for DOA estimation in this study. With a projection matrix that enhances the interference component in the array data, the presented matrix filter can provide deep nulls and a lower passband response error when the array output energy is minimized. Additionally, reducing the beam-space dimension can simplify the problem dimension, reducing the computational complexity. An estimator based on the DR-MFDN and MOMP algorithms with the Gaussian pre-whitening (GP) operations can then be obtained. An estimator based on the DR-MFN and MUSIC algorithms is used to verify the effectiveness of the pre-whitening operation. Finally, the DOA estimation performances of four combination estimators, comprising two DOA estimation algorithms (MUSIC and MOMP) and two matrix filters (DR-MFN and DR-MFDN), are compared using the preferred pre-whitening operation.

In this study, we extended our preliminary work in the conference paper [30]. Although the conference paper proposed the concept of DR-MFDN and demonstrated its preliminary performance, due to scope limitations, the paper failed to provide detailed algorithm descriptions, extensive simulation and experimental verification, and analysis of computation time. Simulations and experimental results demonstrate that the presented adaptive matrix filter exhibits greater robustness in the passband and outperforms interfer-

ence suppression. Moreover, the GP operation is better adapted for processing beam-space snapshots than the OP operation. The estimator based on adaptive matrix filters and the MOMP algorithm with the GP operation can achieve high localization accuracy.

2. Materials and Methods

2.1. Signal Model and Pre-Whitening Operation

In this section, sparse models of element-space and beam-space signals are established, and a novel pre-whitening operation based on a Gaussian matrix is introduced.

2.1.1. Element-Space Signal Model

Consider K far-field signals with frequency f impinging on a uniform linear array with M elements distributed along the x direction with spacing d. The received signal at the mth element of the array can be considered as the sum of the finite plane waves as follows:

$$p_m(f) = \Delta\theta \sum_{l=1}^{L} s(f,\theta_j) e^{i(\frac{2\pi f}{c} x_m \sin\theta_j)} \tag{1}$$

where c is the velocity of sound; $s(f,\theta_j)$ is the plane wave amplitude in θ_j; $\Delta\theta = \frac{\pi}{J}$, with J being the number of angular discretizations, and $\theta_j = (j - \frac{1}{2})\Delta\theta - \frac{\pi}{2} h, j = 1, \cdots, J$.

The wideband signals received by the array were partitioned into N segments, and an L-point FFT was applied to each segment. The array output in the frequency domain can be expressed in the discrete form shown in Equation (1) as follows:

$$\mathbf{p}_l(n) = \mathbf{A}(f_l, \mathbf{\Theta})\mathbf{s}_l(n) + \mathbf{e}_l(n), n = 1, 2, \cdots, N \tag{2}$$

where $\mathbf{A}(f_l, \mathbf{\Theta}) \in \mathbb{C}^{M\times J}$ is the steering matrix, $\mathbf{\Theta} = [\theta_1, \cdots, \theta_J]^T$ is the vector that comprises J candidate directions, and the superscript T is the transpose operator. Moreover, $\mathbf{p}_l(n) \in \mathbb{C}^{M\times 1}$, $\mathbf{s}_l(n) \in \mathbb{C}^{J\times 1}$, and $\mathbf{e}_l(n) \in \mathbb{C}^{M\times 1}$ represent the FFT coefficients of the received data, wideband signals, and additive noise at the nth segment, respectively. The steering matrix is given by

$$\mathbf{A}(f_l, \mathbf{\Theta}) = \left[\mathbf{a}(f_l, \theta_1), \cdots, \mathbf{a}(f_l, \theta_j), \cdots, \mathbf{a}(f_l, \theta_J)\right] \tag{3}$$

where

$$\mathbf{a}(f_l, \theta_j) = [1, e^{i(\frac{2\pi f_l}{c} d \sin\theta_j)}, \cdots, e^{i(\frac{2\pi f_l}{c}(M-1) d \sin\theta_j)}]^T \tag{4}$$

Assuming that $\mathbf{e}_l$ is white noise, the array covariance matrix $\mathbf{R}_l \in {}^{M\times M}$ of the array output in the frequency domain is expressed as follows [14,31]:

$$\mathbf{R}_l = \mathbb{E}[\mathbf{p}_l(n)\mathbf{p}_l^H(n)] = \mathbf{A}(f_l, \mathbf{\Theta})\mathbf{R}_\mathbf{s}^l \mathbf{A}^H(f_l, \mathbf{\Theta}) + \mathbf{R}_\mathbf{e}^l = \mathbf{A}(f_l, \mathbf{\Theta})\mathbf{R}_\mathbf{s}^l \mathbf{A}^H(f_l, \mathbf{\Theta}) + {\sigma_l}^2 \mathbf{I}_M \tag{5}$$

where $\mathbb{E}[\,\cdot\,]$ denotes the expectation operation; superscript H represents the Hermitian transpose; $\mathbf{R}_s^l = \mathbb{E}[\mathbf{s}_l(n)\mathbf{s}_l^H(n)] \in \mathbb{C}^{K\times K}$, $\mathbf{R}_\mathbf{e}^l = \mathbb{E}[\mathbf{e}_l(n)\mathbf{e}_l^H(n)] \in \mathbb{C}^{M\times M}$; ${\sigma_l}^2$ is the variance of the white noise; and $\mathbf{I}_M$ is an identity matrix of dimension M.

2.1.2. Beam-Space Signal Model and Pre-Whitening Operation

Conventional array processing methods cannot effectively estimate the DOAs of weak targets under strong interfering sources. Therefore, transforming the signal into the designed beam-space is necessary. The beam-space snapshot vector $\mathbf{z}_l(n) \in \mathbb{C}^{M\times 1}$ in reduced dimensions ($M' < M$ is the beam-space dimension) is defined as a linear transformation of the frequency-domain snapshot $\mathbf{p}_l(n)$ as follows:

$$\mathbf{z}_l(n) = \mathbf{G}_l^H \mathbf{p}_l(n) = \mathbf{G}_l^H \mathbf{A}(f_l, \mathbf{\Theta})\mathbf{s}_l(n) + \mathbf{G}_l^H \mathbf{e}_l(n) \tag{6}$$

where $\mathbf{G}_l^H$ is the matrix filter corresponding to the frequency f_l.

The beam-space array covariance matrix $\mathbf{R_z}$ is given by

$$\mathbf{R_z} = \mathrm{E}[\mathbf{z}_l(n)\mathbf{z}_l^H(n)] = \mathbf{G}_l^H\mathbf{R}_l\mathbf{G}_l = \mathbf{G}_l^H\mathbf{A}(f_l,\boldsymbol{\Theta})\mathbf{R}_\mathbf{s}^l\mathbf{A}^H(f_l,\boldsymbol{\Theta})\mathbf{G}_l + {\sigma_l}^2\mathbf{G}_l^H\mathbf{G}_l \tag{7}$$

If $\mathbf{G}_l$ is not orthogonal, it transforms white noise into colored noise, which is not convenient for subsequent processing. Thus, an additional noise pre-whitening operation is required. Under a conventional OP operation, Equation (6) can be rewritten as [16,17,32]

$$\hat{\mathbf{z}}_l(n) = (\mathbf{G}_l^H\mathbf{G}_l)^{-1/2}\mathbf{G}_l^H\mathbf{p}_l(n) \tag{8}$$

The application of an OP operation alters the characteristics of the beam space. Considering that the SVD decomposition of the matrix filter is performed with $\mathbf{G}_l = \mathbf{u}\boldsymbol{\varepsilon}\mathbf{v}^H$ and $\mathbf{G}_l\mathbf{G}_l^H = \mathbf{u}\boldsymbol{\varepsilon}^2\mathbf{u}^H$, the beam-space response $h(f_l,\theta)$ by the matrix filter $\mathbf{G}_l$ is

$$h(f_l,\theta) = \frac{\left\|\mathbf{G}_l^H\mathbf{a}(f_l,\theta)\right\|_2}{\left\|\mathbf{a}(f_l,\theta)\right\|_2} = [\frac{1}{M}\mathbf{a}^H(f_l,\theta)\mathbf{u}\boldsymbol{\varepsilon}^2\mathbf{u}^H\mathbf{a}(f_l,\theta)]^{1/2} \tag{9}$$

With an OP operation, Equation (9) can be rewritten as

$$\widehat{h}(f_l,\theta) = \frac{\left\|(\mathbf{G}_l^H\mathbf{G}_l)^{-1/2}\mathbf{G}_l^H\mathbf{a}(f_l,\theta)\right\|_2}{\left\|\mathbf{a}(f_l,\theta)\right\|_2} = [\frac{1}{M}\mathbf{a}^H(f_l,\theta)\mathbf{u}\mathbf{u}^H\mathbf{a}(f_l,\theta)]^{1/2} \tag{10}$$

Considering Equations (9) and (10), the OP operation normalizes the eigenvalues of the matrix $\mathbf{G}_l^H$, modifying the beam space. To ensure that the beam-space transformation preserves the original characteristics and prevents white noise from being transformed into colored noise, a Gaussian matrix was employed for the pre-whitening operation. The frequency-domain snapshot is transformed into

$$\tilde{\mathbf{z}}_l(n) = \mathbf{G}_a\mathbf{G}_l^H\mathbf{p}_l(n) \tag{11}$$

where $\mathbf{G}_a = [\mathbf{g}_{a1},\mathbf{g}_{a2},\cdots,\mathbf{g}_{aM'}]^T \in \mathbb{C}^{M'\times M'}$ is a Gaussian matrix, with each vector obeying $\mathbf{g}_{ai}(m) \sim \mathrm{N}(0,\sigma_l^2), i \mid m = 1,2,\cdots,M'$ and any two arbitrary vectors obeying $\mathrm{cov}(\mathbf{g}_{ai},\mathbf{g}_{aj})_{(i\neq j,1\leq i,j\leq M')} = 0$.

With the GP operation, the beam-space response can be reformulated as follows:

$$\hat{h}(f_l,\theta) = \frac{\left\|\mathbf{G}_a\mathbf{G}_l^H\mathbf{a}(f_l,\theta)\right\|_2}{\left\|\mathbf{a}(f_l,\theta)\right\|_2} = \left[\frac{1}{M}\mathbf{a}^H(f_l,\theta)\mathbf{G}_l\mathbf{G}_a^H\mathbf{G}_a\mathbf{G}_l^H\mathbf{a}(f_l,\theta)\right]^{1/2} = h(f_l,\theta) \tag{12}$$

According to Appendix A, noise covariance matrix $\mathbf{R}_\mathbf{e}^l$ is transformed as follows [33]:

$$\mathbf{G}_a\mathbf{G}_l^H\mathbf{R}_\mathbf{e}^l\mathbf{G}_l\mathbf{G}_a^H = \sigma_l^2\mathbf{I}_{M'} \tag{13}$$

where $\mathbf{I}_{M'}$ is an identity matrix of dimension M'.

It is important to note that perfectly uncorrelated Gaussian matrices do not exist in practice. However, Gaussian matrices statistically approximate this property. Therefore, extensive experimentation is needed to approach the ideal performance of a Gaussian matrix, with the average results obtained reflecting the desired outcome.

2.2. Robust DOA Estimator Based on a Matrix Filter and MOMP Algorithm

In this section, we present a robust DOA estimator based on a matrix filter and the MOMP algorithm. Referring to Equation (11), the beam-space snapshot vector with the GP operation can be further expressed as follows:

$$\tilde{\mathbf{z}}_l(n) = \mathbf{B}_l^H \mathbf{p}_l(n) = \mathbf{\Psi}_l \widetilde{\mathbf{s}}_l(n) + \mathbf{B}_l^H \mathbf{e}(n) \tag{14}$$

where $\mathbf{B}_l = (\mathbf{G}_a \mathbf{G}_l^H)^H$; $\mathbf{\Psi}_l = \mathbf{B}_l^H \mathbf{A}(f_l, \mathbf{\Theta}) \in \mathbb{C}^{M' \times J}$ is the dictionary; $\widetilde{\mathbf{s}}_l(n)$ is the amplitude of the plane wave in the candidate direction after filtering, and its sparsity $K'(K' \leq K)$ is the number of residual signals after filtering.

In this study, the MOMP algorithm is an extended form of the OMP algorithm under the MMV model to solve for the column-uniform sparse matrix $\widetilde{\mathbf{s}}_l$. The algorithm is as follows:

Algorithm 1 Multiple-Measurement-Vector Orthogonal Matching Pursuit (MOMP)

ine Step 1	Input: $\tilde{\mathbf{z}}_l$, $\mathbf{\Psi}_l$, K'.
ine Step 2	Initialization: residual $\mathbf{R}_0 = \tilde{\mathbf{z}}_l$, subset $\Lambda_0 = \emptyset$, and $i = 1$.
ine Step 3	At the i-th iteration step: (a) Choose atom $'_i$ satisfying $'_i = \arg\max_{'_i} \left\| \mathbf{\Psi}_l^T \mathbf{R}_{i-1} \right\|_p$, and $p \geq 1$; (b) Update $\Lambda_i = [\Lambda_{i-1}, '_i]$, and $\widehat{\mathbf{s}}_l^i = (\Lambda_i^H \Lambda_i)^{-1} \Lambda_i^H \tilde{\mathbf{z}}_l$; (c) Update $\mathbf{R}_i = \tilde{\mathbf{z}}_l - \mathbf{\Psi}_l \widehat{\mathbf{s}}_l^i$; (d) If $i < \mathrm{K}'$, return to step 3; if $i \geq \mathrm{K}'$, terminate the iteration and switch to step 4.
ine Step 4	Output: The solution $\widetilde{\mathbf{s}}_l = \widehat{\mathbf{s}}_l^i$ and residual $\mathbf{R}_{\tilde{\mathbf{z}}} = \mathbf{R}_{\tilde{\mathbf{z}}}^i$.
ine	

After solving for the matrix $\widetilde{\mathbf{s}}_l$ at every frequency, the final estimation of the spatial spectrum in the beam space can be calculated as follows:

$$\mathbf{s} = \frac{1}{N} \sum_{l=1}^{L} \sum_{n=1}^{N} \widetilde{\mathbf{s}}_l(n) \tag{15}$$

2.3. Design of Dimension-Reduced Matrix Filter with Deep Nulling

This section describes the approach to achieving high localization accuracy of the presented estimator in an environment with strong interfering sources. In particular, a dimension-reduced matrix filter with deep nulling is presented to improve interference suppression and robustness compared to a dimension-reduced matrix filter with nulling.

2.3.1. Design of Dimension-Reduced Matrix Filter with Nulling

The design of DR-MFNs based on quiescent matrix filters (QMFs) was proposed in [9]. This method uses a matrix filter, designed by the discrete prolate spheroidal sequence-based approach [34,35] as a quiescent matrix filter, to design the DR-MFN by constraining the difference between the actual response and QMF and the response of the stopband while minimizing the output power of the filtered array. The design of the MFN can be formulated as follows:

$$\begin{aligned} &\min_{\mathbf{G}_l} tr\left\{ \mathbf{G}_l^H \mathbf{R}_l \mathbf{G}_l \right\} \\ &s.t. \begin{cases} \left\| \mathbf{G}_l^H - \mathbf{G}_{lq}^H \right\|_F \leq \varepsilon \\ \left\| \mathbf{G}_l^H \mathbf{a}(f_l, \theta_s) \right\|_2 \leq \delta,\ \theta_s \in \Theta_S, s = 1, \cdots, S \end{cases} \end{aligned} \tag{16}$$

where, Θ_S combines a continuum of all out-of-sector directions; $\theta_s \in \Theta_S (s = 1, \cdots, S)$ is the angular grid chosen (uniform or nonuniform) that properly approximates the stopband region by a finite number S of directions; $\|\bullet\|_F$ is the matrix Frobenius norm; $\|\bullet\|_2$ is

the vector 2-norm; $\varepsilon > 0$ is the parameter that bounds the passband distortion of the designed matrix filter $\mathbf{G}_l$ with respect to $\mathbf{G}_{lq}$; $\delta > 0$ is a parameter of the user's choice that characterizes the worst acceptable SA; and $\mathbf{G}_{lq}$ is the QMF. The beam space of QMFs varies across different beam-space dimensions, as shown in Figure 1.

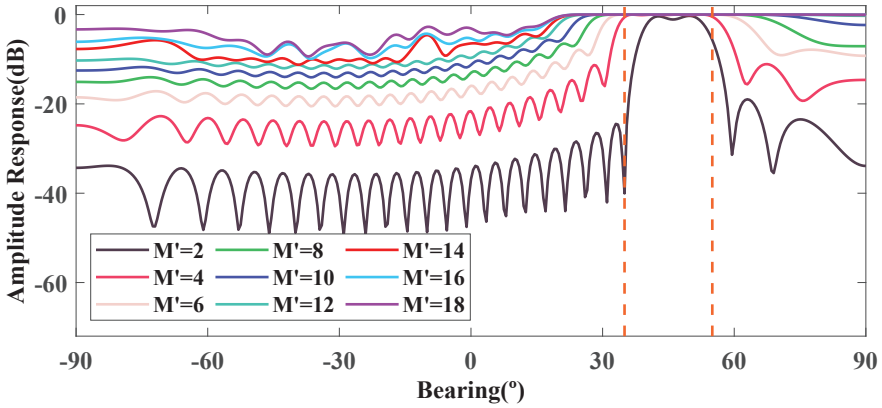


Figure 1. Beam-space amplitude response of the QMFs across various dimensions. The two orange dashed lines represent the left and right limitations of the sector-of-interest area.

Figure 1 shows the beam-space amplitude response of the QMFs across various dimensions for a 32-element uniform line array with a spacing of 1 m when the passband is set to $[33°, 55°]$ with a frequency of 600 Hz. Note that the QMFs have a high passband response error within the preset passband for $M' = 2$. This is because of the low dimensionality, which may lead to an insufficient number of eigenvalues to adequately represent the passband space. Therefore, selecting such a low number of dimensions is not advisable when designing the DR-MFN. Moreover, the dimensions should satisfy

$$M_p \leq M' \leq M \tag{17}$$

where M_p is the minimum number of eigenvalues required to represent passband space.

Equation (16) indicates that DR-MFN is derived by imposing additional constraints based on QMF. Therefore, the beam-space characteristics of the DR-MFN are related to QMF. Applying energy constraints in noninterfering directions may lead to undesired distortions within the passband and insufficient suppression of strongly interfering sources. In particular, obtaining the desired matrix filter when QMF does not satisfy the desired SA is challenging because the stopband and energy constraints are complicated, and the parameters are difficult to establish. Therefore, the robustness of this technique for out-of-sector sources may be insufficient, making it impractical.

2.3.2. Design of Dimension-Reduced Matrix Filter with Deep Nulling

To enhance the robustness of the adaptive matrix filter with nulling, the DR-MFDN was proposed and used for the DOA estimation of the signals of interest. Assuming that all out-of-sector sources satisfy INR ≥ 15 dB, the design procedure is as follows:

Step 1 Input: f_l, Θ_P, Θ_S, SA, $\mathbf{R}_l$, ϵ.
Step 2 Initialization: $\delta = 10^{(\mathrm{SA}/20)}$.
Step 3 Preprocess with MVDR algorithm [36]:
(a) Search spectral peaks in stopband and record their normalized amplitudes $[I_1, \cdots, I_{K''}]$;
(b) Record the directions $[\theta_1, \cdots, \theta_{K''}]$ corresponding to the peaks.
Step 4 Generate the projection matrix:

(a) The sector of the main lobe where each peak is located is identified; the k''th sector of the main lobe is denoted as $\Theta_{k''}$,

$$\Theta_{k''} = [\theta_{k''-l_{k''}}, \cdots, \theta_{k''}, \cdots, \theta_{k''+r_{k''}}], k'' = 1, \cdots, K''. \tag{18}$$

where $\theta_{k''-l_{k''}}$ and $\theta_{k''+r_{k''}}$ are the left and right boundary angulars of the main lobe corresponding to the k''th peak, respectively.
(b) The projection matrix is constructed as follows: $\mathbf{\Psi}_l = [\mathbf{A}(f_l, \mathbf{\Theta}_1), \cdots, \mathbf{A}(f_l, \mathbf{\Theta}_{K''})]$ at frequency f_l.

Step 5 Formulate the optimization problem:
(a) Conventional form:

$$\begin{aligned} &\min_{\mathbf{G}_l} tr\left\{ {\mathbf{G}_l}^H \mathbf{\Psi}_l^H \mathbf{R}_l \mathbf{\Psi}_l \mathbf{G}_l \right\} \\ &\text{s.t.} \begin{cases} \left\| {\mathbf{G}_l}^H - \mathbf{G}_{lc}^H \right\|_F \le \varepsilon \\ \left\| {\mathbf{G}_l}^H \mathbf{a}(f_l, \theta) \right\|_2 \le \delta \end{cases} \end{aligned} \tag{19}$$

Step 6 Transform Equation (19) into SOCP form:

$$\begin{aligned} \mathrm{tr}\left\{ \mathbf{G}_l^H \mathbf{\Psi}_l^H \mathbf{R}_l \mathbf{\Psi}_l \mathbf{G}_l \right\} &= \mathrm{tr}\left\{ \mathbf{G}_l^H \mathbf{\Psi}_l^H \mathbf{R}_l^{1/2} \mathbf{R}_l^{1/2} \mathbf{\Psi}_l \mathbf{G}_l \right\} \\ &= \left\| \mathbf{G}_l^H \left(\mathbf{\Psi}_l^H \mathbf{R}_l^{1/2} \right) \right\|_{\mathbf{F}}^2 \\ &= \left\| \mathrm{vec}\left\{ \mathbf{I}_{M'} \mathbf{G}_l^H \left(\mathbf{\Psi}_l^H \mathbf{R}_l^{1/2} \right) \right\} \right\|_2^2 \\ &= \left\| \left[\left(\mathbf{\Psi}_l^H \mathbf{R}_l^{1/2} \right)^T \otimes \mathbf{I}_{M'} \right] \mathbf{g}_l \right\|_2^2 \end{aligned} \tag{20}$$

$$\left\| \mathbf{G}_l^H - \mathbf{G}_{lc}^H \right\|_{\mathbf{F}} = \| \mathbf{g}_l - \mathbf{g}_{lc} \|_2 \tag{21}$$

$$\left\| \mathbf{G}_l^H \mathbf{a}(f_l, \theta_s) \right\|_2 = \left\| \left[\mathbf{a}^T(f_l, \theta_s) \otimes \mathbf{I}_{M'} \right] \mathbf{g}_l \right\|_2 \tag{22}$$

where $\mathrm{vec}\{\cdot\}$ denotes the vectorization operator stacking the columns of a matrix on top of each other, $\mathbf{g}_l = \mathrm{vec}(\mathbf{G}_l^H)$, $\mathbf{g}_{lc} = \mathrm{vec}(\mathbf{G}_{lc}^H)$, $\otimes$ denotes the Kronecker matrix product and $\mathbf{I}_{M'}$ is an identity matrix of dimension M'. Substituting Equations (20)–(22) into Equation (19), Equation (19) can be rewritten as

$$\begin{aligned} &\min_{\mathbf{g}_l} \left\| \left[\left(\mathbf{\Psi}_l^H \mathbf{U}_l \right)^T \otimes \mathbf{I}_{M'} \right] \mathbf{g}_l \right\|_2 \\ &\text{s.t.} \begin{cases} \| \mathbf{g}_l - \mathbf{g}_{lc} \|_2 \le \varepsilon \\ \left\| \left[\mathbf{a}^T(f_l, \theta_s) \otimes \mathbf{I}_{M'} \right] \mathbf{g}_l \right\|_2 \le \delta, \theta_s \in \Theta_S, s = 1, \cdots, S. \end{cases} \end{aligned} \tag{23}$$

Step 7 Solve the SOCP Problem: Use an SOCP solver (e.g., mosek) to solve for $\mathbf{g}_l$.
Step 8 Output: Reshape $\mathbf{g}_l$ back to matrix form $\mathbf{G}_l = [\mathrm{mat}(\mathbf{g}_l)]^H$.

Through the above steps, a DR-MFDN $\mathbf{G}_l$ at frequency f_l can be obtained. The enhancement principle of the projection matrix is to project all array data onto the subspace corresponding to the main lobe.

An unknown source with direction θ_u is projected as

$$\begin{aligned}\mathbf{x} &= I_u\mathbf{A}(f_l,\Theta_{k''})\mathbf{A}^H(f_l,\Theta_{k''})\mathbf{a}(f_l,\theta_u)\\ &= I_u\mathbf{A}(f_l,\Theta_{k''})[\mathbf{a}^H(f_l,\theta_{k''-l_{k''}});\cdots;\mathbf{a}^H(f_l,\theta_{k''+r_{k''}})]\mathbf{a}(f_l,\theta_u)\\ &= \tfrac{I_u}{l_{k''}+r_{k''}+1}[\mathbf{a}(f_l,\theta_{k''-l_{k''}}),\cdots,\mathbf{a}(f_l,\theta_{k''+r_{k''}})][\Delta_{l_{k''}},\cdots,1,\cdots,\Delta_{r_{k''}}]^T\\ &= I_u\Delta_u\mathbf{a}(\theta_u)\end{aligned} \tag{24}$$

where $\Delta_{l_{k''}} = \frac{\sin[\frac{\pi f_l Md}{c}(\sin\theta_u-\sin\theta_{k''-l_{k''}})]}{M\sin[\frac{\pi f_l d}{c}(\sin\theta_u-\sin\theta_{k''-l_{k''}})]}$; $\Delta_{r_{k''}} = \frac{\sin[\frac{\pi f_l Md}{c}(\sin\theta_u-\sin\theta_{k''+r_{k''}})]}{M\sin[\frac{\pi f_l d}{c}(\sin\theta_u-\sin\theta_{k''+r_{k''}})]}$; and $\Delta_u < 1$ is the amplitude of the transformation of the unknown source into an interfering source after projection.

If $\theta_u = \theta_{k''}$, the spectrum estimation obtained by the MVDR algorithm provides a narrower main lobe; that is, $\Delta_{l_{k''}} \approx \Delta_{r_{k''}} \approx 1$, and Equation (24) can be rewritten as follows:

$$\mathbf{x} \approx \frac{I_{k''}}{l_{k''}+r_{k''}+1}[f_l,\mathbf{a}(\theta_{k''-l_{k''}}),\cdots,\mathbf{a}(f_l,\theta_{k''+r_{k''}})][1,\cdots,1]^T \approx I_{k''}\mathbf{a}(f_l,\theta_{k''}) \tag{25}$$

From Equations (24) and (25), it can be seen that the projection matrix retains the out-of-sector source components while concentrating the signal-of-interest components from other directions onto these out-of-sector sources, thereby directing energy toward the strong out-of-sector sources. Additionally, the interfering sources from different directions reinforce each other.

By enhancing the out-of-sector sources through the projection matrix, the DR-MFDN algorithm, which is designed based on the minimum energy criterion, effectively reduces the impact of perturbations and noise within the passband, achieving deeper nulls in the direction of strong interfering sources. As a result, the robustness of the adaptive matrix filter with nulling is significantly improved.

3. Discussion

In the simulation, a uniform linear array of 32 isotropic array elements with a spacing of 1 m was considered to analyze the effectiveness of the GP operation and DOA estimation performance of the robust estimator based on the MOMP algorithm and DR-MFDN. Three distinct received-signal scenarios were simulated.

3.1. The Influence of the GP Operation

The first scenario assumed that one weak signal of interest with DOA $\theta_1 = 45^\circ$ was present. Moreover, a strong out-of-sector source (interfering) located at $\theta_2 = 0^\circ$ (90° was defined as the end fire direction within the frequency range of 400 to 700 Hz. The signal-to-noise ratio (SNR) was 10 dB, whereas the interference-to-noise ratio (INR) of the interference was 40 dB. The data were partitioned into 10 segments, and a 512-point FFT was performed for each segment. The sector-of-interest area is $\mathbf{\Theta}_p = [35^\circ, 55^\circ]$, the out-of-sector area is $\mathbf{\Theta}_s = [-90^\circ, 28.5^\circ] \cup [64.5^\circ, 90^\circ]$, the beam-space dimension is $M' = 10$, and $\varepsilon = 0.2$, SA $= -20$ dB. In addition, the amplitude responses and errors in the beam space were calculated as $20\lg\{\|\mathbf{G}_l^H\mathbf{a}_l(f_l,\theta)\|_2/\|\mathbf{a}_l(f_l,\theta)\|_2\}$ and $20\lg\{[\|\mathbf{G}_l^H\mathbf{a}_l(f_l,\theta)\|_2 - \|\mathbf{a}_l(f_l,\theta)\|_2]/\|\mathbf{a}_l(f_l,\theta)\|_2\}$, respectively.

The purpose of this simulation scenario is to verify the necessity of the GP operation and to analyze the extent to which both the GP operation and the DR-MFDN affect the DOA estimation.

In the case of DR-MFN, its covariance matrix with no pre-whitening (NP) operation is shown in Figure 2a. $\mathbf{G}_l^H\mathbf{G}_l$ is not a unit matrix, which may result in the transformation of

white noise into colored noise. In Figures 2b and 3, when the OP operation is considered, $(\mathbf{G}_l^H\mathbf{G}_l)^{-1/2}\mathbf{G}_l^H\mathbf{G}_l(\mathbf{G}_l^H\mathbf{G}_l)^{-1/2}$ is the unit matrix, which prevents transforming the noise into colored noise. Meanwhile, the overall attenuation of the out-of-sector area is reduced. Moreover, changes in the beam space may lead to insufficient noise and interference suppression, which may affect the final processing results. Figures 2c,d and 3 consider the case when the GP operation is performed. Note that $\mathbf{B}_l^H\mathbf{B}_l$ is a unit matrix, obtained using numerous Monte Carlo trials, which prevents transforming the noise into colored noise without significant changes in the beam space. Moreover, a Monte Carlo trial was necessary because of the randomness of the Gaussian matrix and the absence of a Gaussian matrix with uncorrelated columns. A large number of Monte Carlo trials is undoubtedly time-consuming. Therefore, determining an appropriate number of trials is necessary. This point will be discussed in the third scenario.

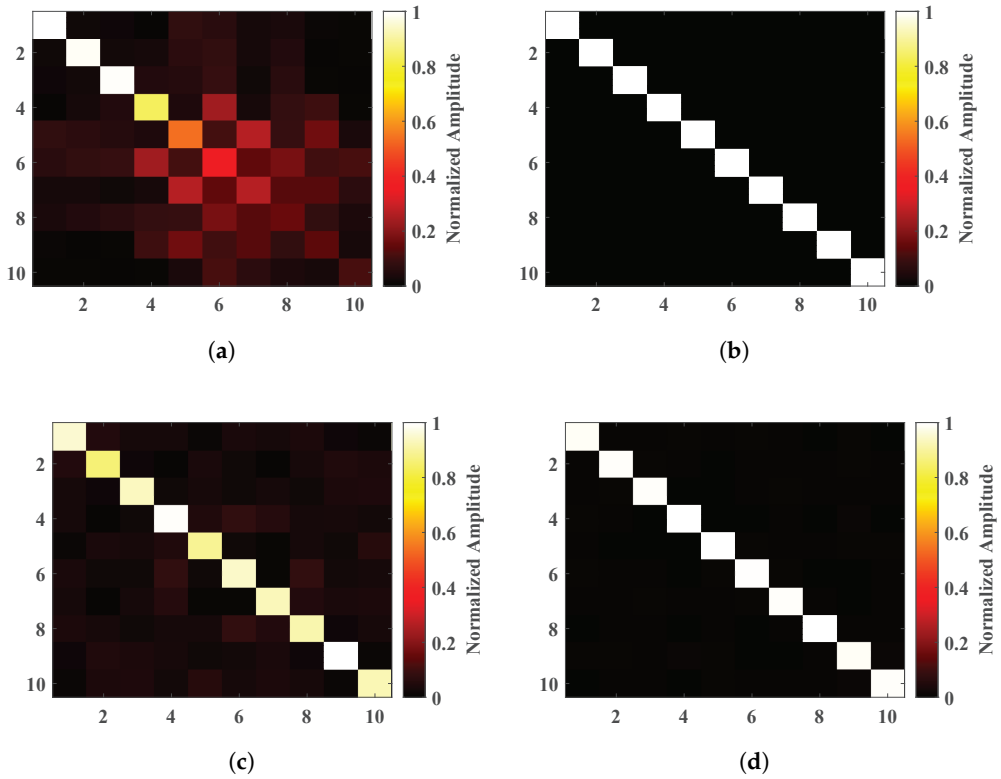


Figure 2. Covariance matrix distributions of array noise. (**a**–**d**) represent the cases of non-pre-whitening (NP) operation, the OP operation, and the GP operation with 100 and 1000 Monte Carlo trials, respectively.

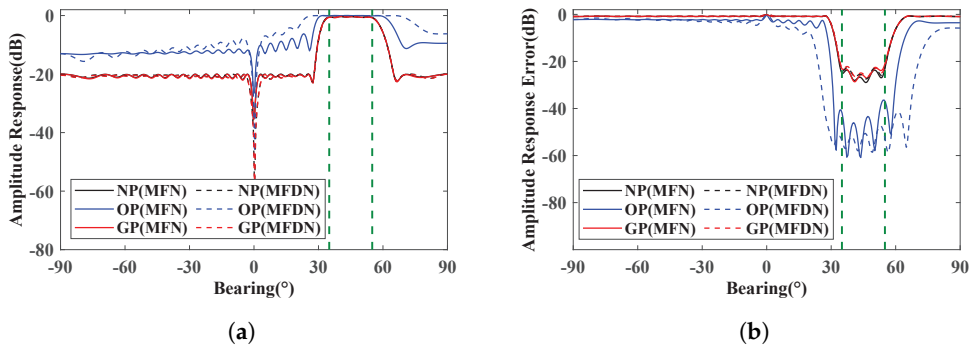


Figure 3. Beam-space characteristics. (**a**,**b**) represent the amplitude response and amplitude response error of DR-MFN with different pre-whitening operations at 600 Hz, respectively. The GP operation conducted 100 Monte Carlo trials. The two orange dashed lines represent the left and right limitations of the sector-of-interest area.

To clarify the difference in DOA estimation performance under different pre-whitening operations and matrix filters, Figure 4 shows the estimation results of the DR-MFN-MUSIC and DR-MFDN-MUSIC algorithms for different pre-whitening operations. The OP operation does not effectively reduce the effect of the strong interfering source, causing the energy of the signal of interest to remain lower than that of the interfering source after filtering. In contrast, the case without the pre-whitening operation shows better performance, indicating that the effect is limited at the current noise levels, even with colored noise. However, note that this trend may change under low SNR conditions. Furthermore, the GP operation significantly improves the performance of DOA estimation with matrix filters.

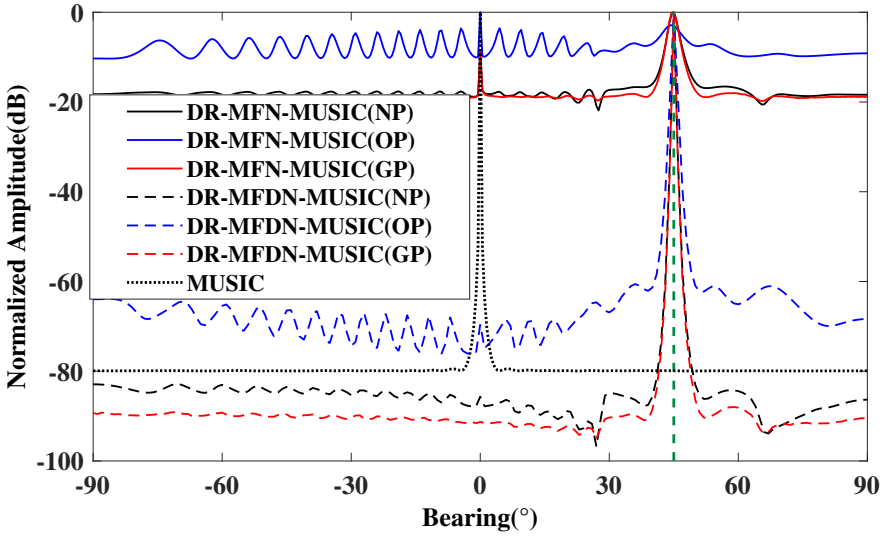
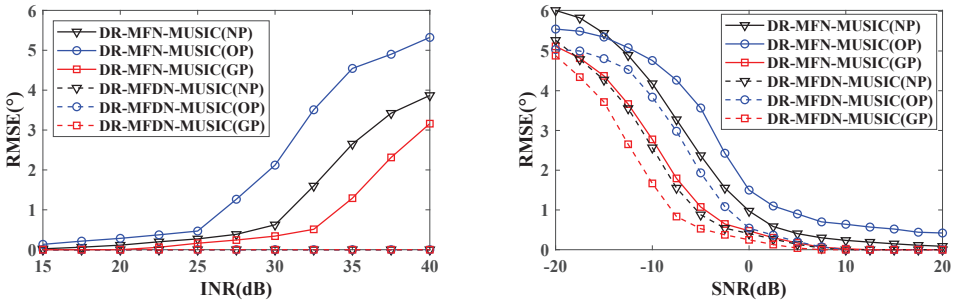


Figure 4. Estimation results of MUSIC without filtering and the DR-MFN-MUSIC and the DR-MFDN-MUSIC algorithms for different pre-whitening operations. The green dashed line represents the DOA of a weak signal of interest.

To analyze the impact of noise on the algorithms, Figure 5 shows the root-mean-square errors (RMSEs) of the DR-MFN-MUSIC and the DR-MFDN-MUSIC algorithms with three pre-whitening operations versus the INR and SNR. Specifically, the INR varies with a fixed SNR of 10 dB, and the SNR varies with a fixed INR of 20 dB. The RMSE curves were averaged over the signal sources and 500 independent simulation runs.



(**a**) (**b**)

Figure 5. DOA estimation of RMSEs of the DR-MFN-MUSIC algorithm versus (**a**) INR and (**b**) SNR with different pre-whitening operations.

In Figure 5a, the following observations are made in the context of using DR-MFN: (1) RMSE increases with INR across all three pre-whitening methods. (2) The GP operation yields the lowest RMSE, while the OP operation yields the highest. This difference is likely due to the OP operation altering the beamspace characteristics (as shown in Figure 3), which weakens DR-MFN's interference suppression. (3) RMSE without pre-whitening is

generally lower than with OP. This may be because, in high-SNR conditions, changes in beamspace characteristics have a greater impact on DOA estimation than colored noise does. In the context of using DR-MFDN with sufficient suppression, noise effects in high-SNR conditions become negligible, leading to no significant performance differences among the three methods.

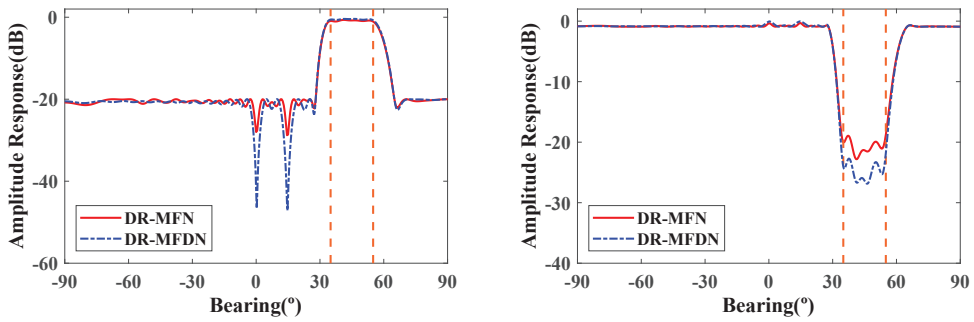
In Figure 5b, further analysis shows the following results. (1) For MFN, the OP operation outperforms the NP operation in the SNR range of $[-20$ dB, -14 dB$]$, which is likely due to the dominant impact of colored noise in low-SNR environments. (2) For DR-MFDN, OP outperforms NP in the range of $[-20$ dB, -18.5 dB$]$, and with identical pre-whitening, DR-MFDN performs better than DR-MFN. This may be attributed to DR-MFDN's lower passband distortion, which makes it less susceptible to noise compared to DR-MFN. (3) Regardless of the filtering method, GP consistently has the smallest impact on DOA estimation, and as SNR increases, the performance differences between the three methods decrease.

In conclusion, based on Figures 4 and 5, DR-MFDN's strong interference suppression and low passband distortion significantly enhance DOA estimation performance for weak signals, while the benefits of the GP operation are relatively limited.

3.2. The Performance of the DR-MFDN-MOMP with the GP Operation

The second scenario assumed that two weak signals of interest with the DOAs $\theta_3 = 44^\circ$ and $\theta_4 = 46^\circ$ were present. Moreover, two strong out-of-sector sources (interfering) were located at $\theta_5 = 0^\circ$ and $\theta_6 = 15^\circ$ within a frequency range of 600 to 700 Hz. The SNR of both signals was 0 dB, whereas the INR of all the interfering sources was 40 dB. All the other parameters of the tested scenario were the same as those of the previous scenario. In addition, a GP operation was applied to this scenario in all the cases for simplicity.

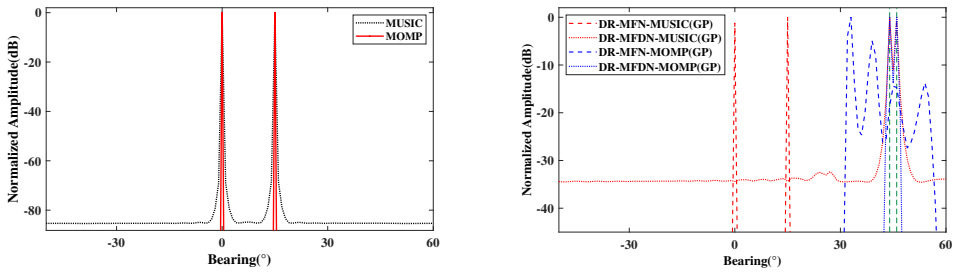
In Figure 6, the DR-MFDN provides deeper nulls in the direction of the interfering sources and a lower passband response error in the sector of interest than the DR-MFN. Figure 7a plots estimation results of the MUSIC and MOMP without filtering, and the directions of the weak signals-of-interest cannot be distinguished. Subsequently, robust estimators based on the MUSIC or MOMP algorithm without filtering and DR-MFN or DR-MFDN, namely, DR-MFN-MUSIC, DR-MFDN-MUSIC, DR-MFN-MOMP, and DR-MFDN-MOMP estimators, with GP operations, were applied for DOA estimation, as shown in Figure 7b. The DR-MFN-MUSIC and DR-MFN-MOMP estimators with the GP operation exhibit poor estimation performance for weak targets, which can be due to the insufficient suppression of interfering sources by the DR-MFN. Comparatively, both algorithms coupled with the DR-MFDN demonstrated effective estimation of weak targets, and the DR-MFDN-MOMP estimator with the GP operation demonstrated a high resolution.



(**a**) (**b**)

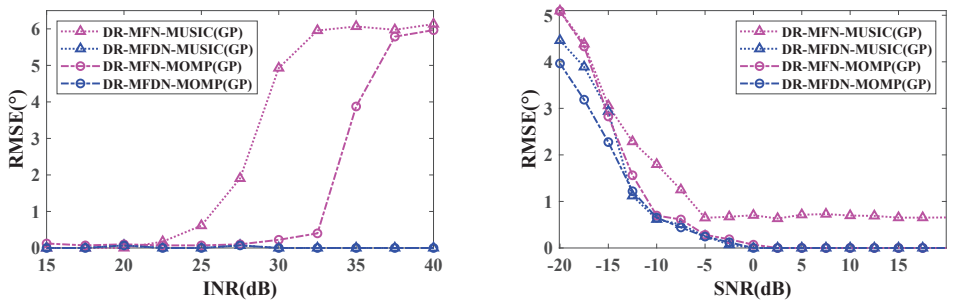
Figure 6. Beam-space characteristics. (**a**,**b**) represent amplitude responses and amplitude response errors of the DR-MFN and DR-MFDN at 600 Hz. The two orange dashed lines represent the left and right limitations of the sector-of-interest area.

To further investigate the performance under varying conditions, the RMSEs of the DR-MFN-MUSIC(GP), DR-MFDN-MUSIC(GP), DR-MFN-MOMP(GP), and DR-MFDN-MOMP(GP) estimator are plotted with respect to INR and SNR in Figure 8. Specifically, the INR was varied for a fixed SNR of 0 dB, and the SNR was varied for a fixed INR of 25 dB. The RMSE curves were averaged over the signal sources and 500 independent simulation runs.



(**a**) (**b**)

Figure 7. Estimation results of (**a**) the MUSIC and MOMP; (**b**) the DR-MFN-MUSIC(GP), DR-MFDN-MUSIC(GP), DR-MFN-MOMP(GP), and DR-MFDN-MOMP(GP) estimators. The two green dashed lines represent the DOAs of two weak signals of interest.



(**a**) (**b**)

Figure 8. DOA estimation RMSEs with respect to (**a**) INR and (**b**) SNR.

Figure 9 shows the probabilities of the source resolution with respect to INR and SNR for the same scenario and methods tested. Similar to Figure 8, all curves were averaged over the signal sources and 500 simulation runs. The signal sources are considered to be resolved in the jth run if [17]

$$\sum_{i=1}^{2} |\widehat{\theta}_i^j - \theta_i| < |\theta_1 - \theta_2| \tag{26}$$

where $\widehat{\theta}_i^j$ is the DOA estimation for the i-th source in the j-th run.

Figures 8a and 9a show that the RMSEs of the DR-MFN-MUSIC(GP) and DR-MFN-MOMP(GP) algorithms increase with increasing INR because of the insufficient suppression of strong out-of-sector sources by the DR-MFN. Conversely, the DR-MFDN provides sufficient suppression, leading to superior DOA estimation capabilities for both the DR-MFN-MUSIC(GP) and DR-MFN-MOMP(GP) estimators. Figures 8b and 9b show that both the DR-MFN and DR-MFDN can sufficiently suppress strong out-of-sector sources for a fixed INR of 25 dB. However, the DR-MFDN improved the performance with the same DOA estimation algorithm, indicating superior signal preservation with a lower passband response error. Figures 8 and 9 show that the MOMP algorithm outperforms the MUSIC algorithm when utilizing the same matrix filter.

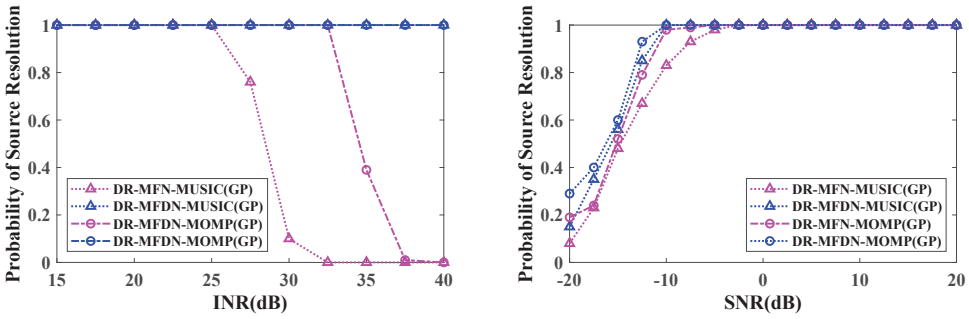


Figure 9. Probabilities of source resolution with respect to (**a**) INR and (**b**) SNR.

3.3. The Efficiency of the Presented Algorithm

In the third simulation scenario, we explore the impact of the number of GP operations on DOA estimation accuracy and algorithm runtime. Given that the GP operation processing is primarily employed to mitigate the nonlinear effects induced by colored noise, we set the SNR to -20 dB, building on the conditions established in Scenario 1. The analysis from the first two scenarios has already demonstrated the necessity of GP processing and the superior performance of the DR-MFDN-MOMP(GP) algorithm for weak-signal DOA estimation. However, the effect of GP on the runtime of different algorithms remains unclear, which is crucial for practical applications. First, we conduct a comparative study of DOA estimation performance for both the DR-MFDN-MOMP(GP) and DR-MFDN-MUSIC(GP) algorithms under varying levels of GP, as shown in Figure 10. Additionally, Figure 10 also illustrates the relationship between the degree of the beam-space response distortion and the number of Monte Carlo trials, where the tolerance is calculated as follows:

$$\text{Tolerance} = \sum_{j=1}^{J} ||\mathbf{G}_l^H \mathbf{a}(f_l, \theta_j) - \mathbf{B}_l^H \mathbf{a}(f_l, \theta_j)||_2 \ (\theta_j \in \Theta) \tag{27}$$

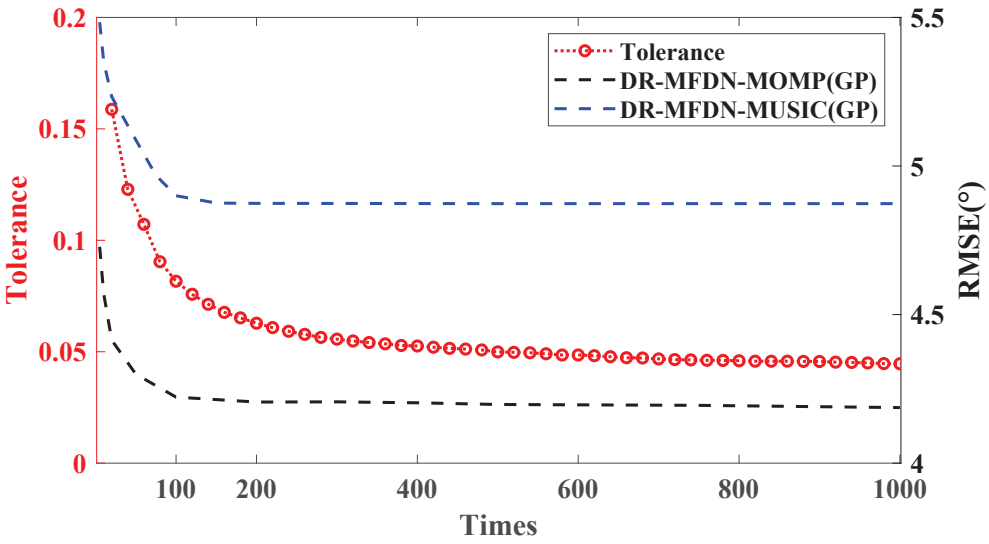


Figure 10. Effect on the beam-space preservation and the DOA estimation RMSEs of the DR-MFDN-MUSIC(GP) and the DR-MFDN-MOMP(GP) algorithms with different numbers of GP operations.

Figure 10 illustrates that as the number of the GP operations increases, the tolerance decreases. The rate of reduction progressively slows down, with convergence not being achieved until after 1000 operations. While an increased number of GP operations may yield a smaller tolerance, it also introduces a substantial computational burden. Upon evaluating the DOA estimation performance of both algorithms across varying GP operation counts, it is evident that the RMSEs for both algorithms reach convergence after approximately 100 GP operations. This phenomenon occurs because, although further increasing the GP operation count reduces the discrepancy between the pre-whitened matrix filter and the original matrix filter, the original matrix filter inherently exhibits a degree of passband distortion. After approximately 100 GP operations, the passband distortion stabilizes and ceases to exhibit significant changes. Consequently, selecting 100 GP operations effectively maximizes the performance gains from the GP processing while minimizing computational overhead.

Figure 11 shows that the RMSE decreases as the number of snapshots increases for both algorithms. This indicates that the accuracy of DOA estimation improves with more data points, leading to a reduction in estimation error. Throughout the range of snapshot numbers tested, the DR-MFDN-MOMP(GP) algorithm consistently exhibits lower RMSE values compared to the DR-MFDN-MUSIC(GP) algorithm. This suggests that DR-MFDN-MOMP(GP) provides more accurate DOA estimations under the same conditions. The rate at which RMSE decreases is relatively steep when the number of snapshots is low (e.g., between 0 and 10). As the number of snapshots continues to increase, the rate of decrease in RMSE begins to slow down, particularly noticeable after surpassing 20 snapshots. This suggests diminishing returns on increasing the number of snapshots beyond a certain point.

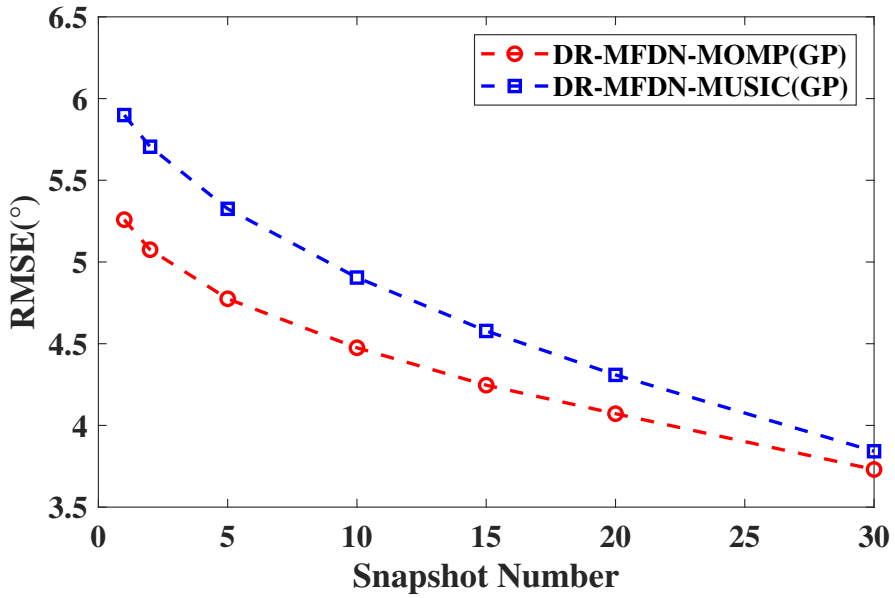


Figure 11. The DOA estimation RMSEs of the DR-MFDN-MUSIC(GP) and the DR-MFDN-MOMP(GP) algorithms with A different number of snapshots.

Based on the above conclusion, we computed the runtime of the two algorithms under different numbers of the GP operations, as shown in Table 1. These simulations were performed on a workstation equipped with an Intel Core i9-13900HX processor (24 cores), 16 GB of DDR5 memory running at 5600 MHz, and an NVIDIA GeForce RTX 4060 Laptop GPU with 8 GB of dedicated memory. The primary storage device was a 1 TB NVMe SSD (WD PC SN810), and the system was running on a Windows operating system. The algorithms were implemented in Matlab R2022b.

Table 1. Runtime (ms) of the DR-MFDN-MUSIC(GP) and DR-MFDN-MOMP(GP) algorithms with different numbers of GP operations.

Method \ Times	1	5	10	20	50	100
MUSIC	55.20	\	\	\	\	\
DR-MFDN-MUSIC(GP)	8.40	39.80	79.40	158.70	392.10	779.5
MOMP	0.54	\	\	\	\	\
DR-MFDN-MOMP(GP)	0.50	2.50	4.60	9	22.8	44.60

From Table 1, we can conclude the following. (1) The runtime of MUSIC without dimensionality reduction is significantly higher than that of DR-MFDN-MUSIC(GP), resulting in greater computational complexity. (2) DR-MFDN-MOMP(GP) with 100 GP operations runs in 44.60 ms, much more efficiently than DR-MFDN-MUSIC(GP) at 779.5 ms. While the runtime of DR-MFDN-MOMP(GP) increases with more GP operations, its computational complexity remains low, and the increase is gradual. (3) The runtime of DR-MFDN-MOMP(GP) after 100 GP operations is lower than the single execution time of MUSIC (55.20 ms), meaning it does not introduce a significant computational burden and is suitable for real-time applications.

Based on Figure 10 and Table 1, the DR-MFDN demonstrates a high cost-effectiveness ratio by effectively reducing computational burden while maintaining accuracy through dimensionality reduction. With 100 GP operations, the maximum computational accuracy benefits are achieved. Additionally, the runtime of the DR-MFDN-MOMP(GP) algorithm is significantly lower than that of DR-MFDN-MUSIC(GP), highlighting its superior performance in high GP operation counts. The lower computational time of DR-MFDN-MOMP(GP) makes it well suited for practical, real-time applications. In summary, the DR-MFDN-MOMP(GP) estimator can accurately and quickly estimate DOA for weak targets in an environment with strong interfering sources.

4. Results

In September 2019, a horizontal array detection experiment was conducted in the sea area of Laoshan Bay, Qingdao. During the experiment, the relative planar position between the horizontal array and the surface target vessel is shown in Figure 12a. The horizontal array was deployed in a region with a water depth of 18 m, with 32 hydrophones uniformly spaced at 1 m. The deployment depth of the horizontal array was 17.8 m, and the relative positions of the array elements are illustrated in Figure 12b. The sound speed remained approximately constant at around 1530 m/s, as indicated by the sound speed profile shown in Figure 12c. The seabed in the experimental area was relatively flat, characterized by fine sand.

The sampling frequency was set to 6 kHz, and the duration of the processed data was 2 min. A short-time Fourier transform (STFT) was applied to the received data of a specific array element, resulting in the time-frequency spectrogram shown in Figure 13a. The STFT used a Hamming window of 4096 samples, with a 12.5% overlap. The time-domain and frequency-domain information of a particular segment are illustrated in Figure 13b and Figure 13c, respectively. Figure 13b shows that the received signal appears disordered. However, multiple line spectra can be observed in Figure 13c. In the 400–700 Hz frequency band, background noise is relatively low compared to the spectral line intensity. To better identify the DOAs of weak targets, this frequency band was selected for the broadband DOA estimation of weak signals.

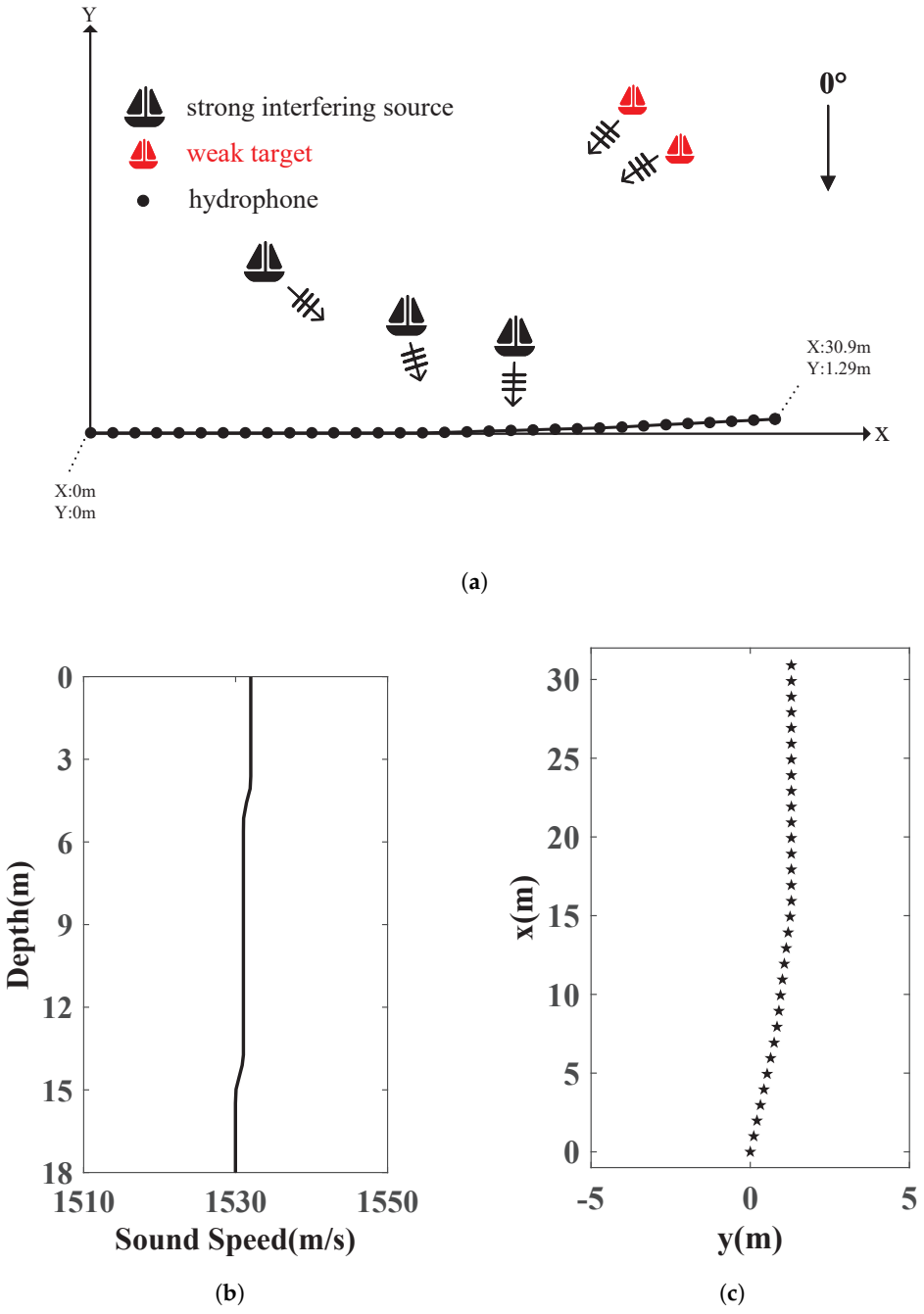


Figure 12. (**a**) Schematic of the relative planar position between the horizontal array and the target vessel located on the water surface. The arrows indicate the arrival directions of the target signals. (**b**) Sound speed profile. (**c**) Relative positions of the elements of the horizontal array.

The observed space, $[-90°, 90°]$, was uniformly discretized into 181 candidate directions. The Bearing-Time Records (BTRs) of the MVDR, MUSIC and MOMP, as shown in Figure 14, reveal information about the interference source and the sector where the weak targets are located. The two interfering sources were moved to the far field of the array. One interference was in the direction of approximately $-17°$, and the other was around $0°$. A few weak targets were located at $-47°$, $42°$, and $54°$, and the first target was clearly tracked. The DR-MFN and DR-MFDN were designed with the passband bearing sector $\boldsymbol{\Theta}_p$ at $[35°, 65°]$ and the stopband bearing sector $\boldsymbol{\Theta}_s$ at $[-90°, 25°] \cup [80°, 90°]$. The SAs were -15 dB. The received data sampled were transformed into the frequency domain by using a 512-point FFT. The sample covariance matrix was obtained by averaging 22 snapshots

with a 50% overlap. The duration of each segment was 1s, and the total analysis time was 2 min. The DR-MFN-MUSIC, DR-MFDN-MUSIC, DR-MFN-MOMP, and DR-MFDN-MOMP estimators with orthogonal or GP operations were applied to localize weak targets. The results are presented in Figures 15 and 16.

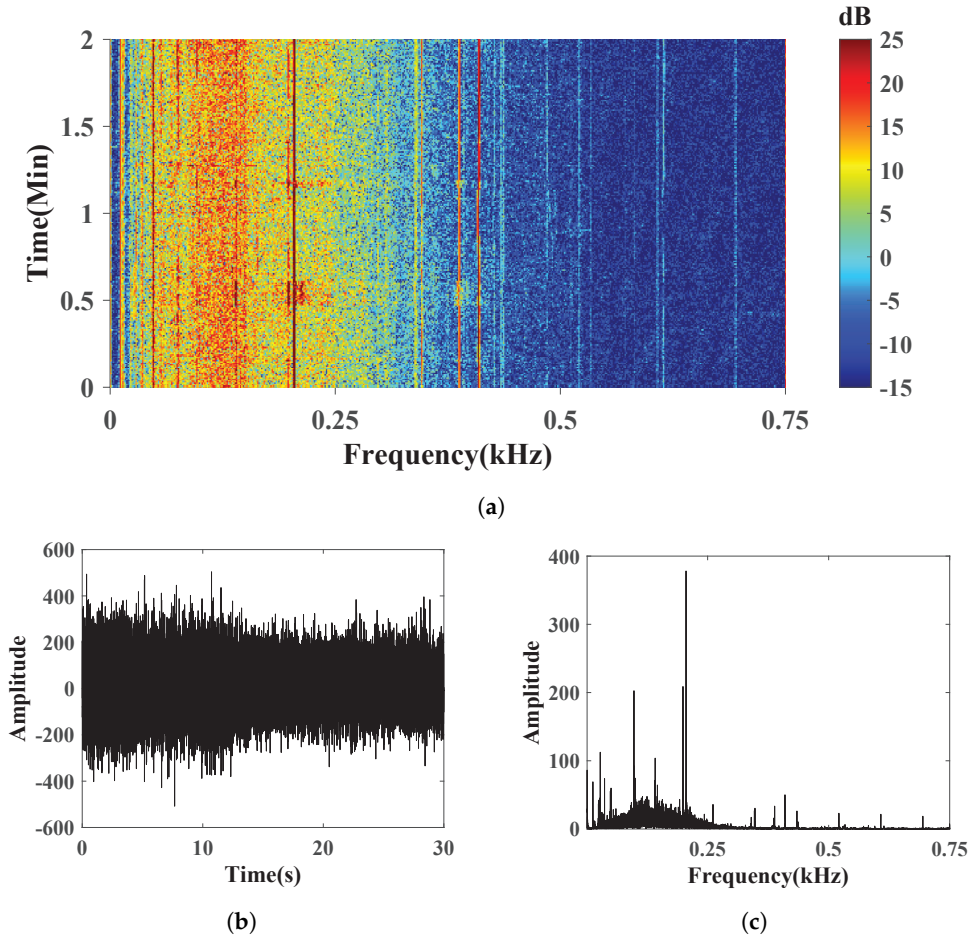


Figure 13. (**a**) Time-frequency spectrogram of the processed data. The information of the data segment in (**b**) time domain and (**c**) frequency domain.

Figure 15 illustrates the DOA estimation results of the four algorithms using the OP operation. Figure 15a,b,e show that the DR-MFN-MUSIC(OP) and DR-MFDN-MUSIC(OP) estimators could not obtain the DOA of a weak target with a broad main lobe. Moreover, many spurious peaks appeared outside these sectors. Figure 15c–e show that the DR-MFN-MOMP(OP) and DR-MFDN-MOMP(OP) estimators could determine the DOAs of the two weak signals. In contrast, occasional spurious peaks were present near the target. Thus, these two algorithms using the OP operation are not robust. Overall, combined with matrix filters, the MOMP algorithm demonstrated superior DOA estimation performance to the MUSIC algorithm, as shown in Figure 15. Figure 16 illustrates the DOA estimation results of the four algorithms with the GP operation. A comparison of Figures 15 and 16 shows that the algorithms with the GP operation outperform those with the OP operation. Figures 15a and 16b,e show that the DR-MFDN-MUSIC(GP) estimator has a smaller response than the DR-MFN-MUSIC(GP) estimator in all non-target sectors and reduces the width of the main lobe. This indicates that the DR-MFDN shows improved interference suppression performance and localization accuracy. In Figure 16c–e, the track obtained using the DR-MFDN-MOMP(GP) algorithm is clearer than that obtained using the DR-MFN-MOMP(GP) estimator. In Figure 16e, the DR-MFDN-MOMP(GP) estimator localizes two weak targets. Owing to the high passband response error of the DR-MFN, the result of the DR-MFN-

MOMP(GP) estimator shows a slight but potentially misleading spurious peak. Therefore, the DR-MFDN-MOMP(GP) estimator achieves superior performance in an environment with strong interfering sources.

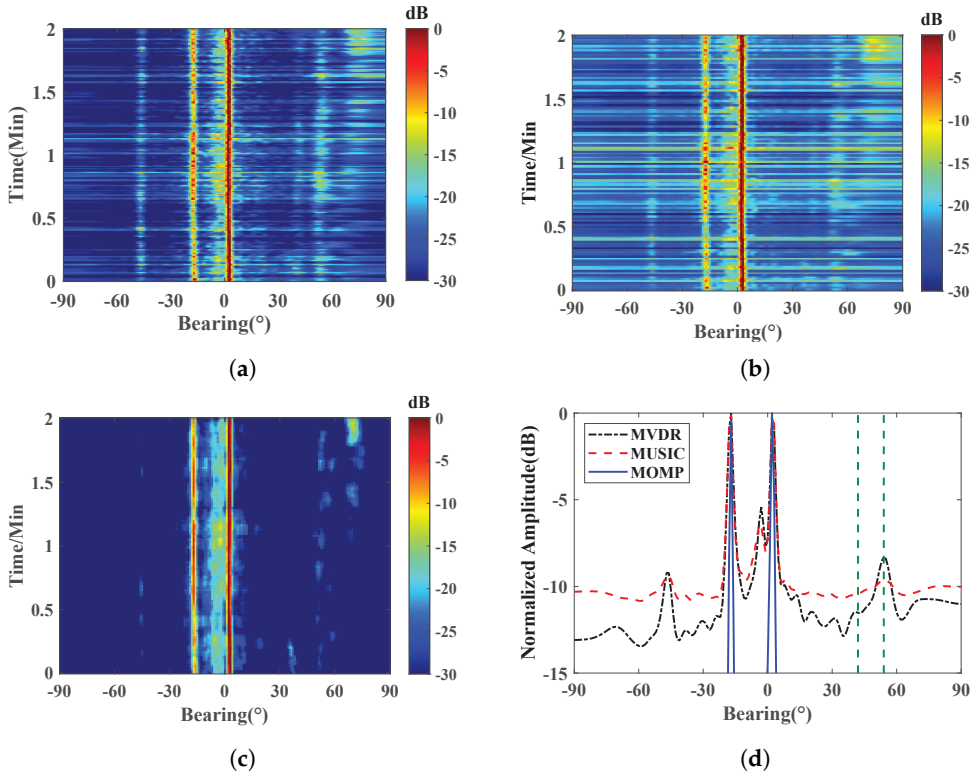


Figure 14. BTRs of the (**a**) MVDR, (**b**) MUSIC, (**c**) MOMP. (**d**) Profile of the results of (**a**–**c**) at t = 57 s. The two green dashed lines in (**d**) represent the directions of weak targets.

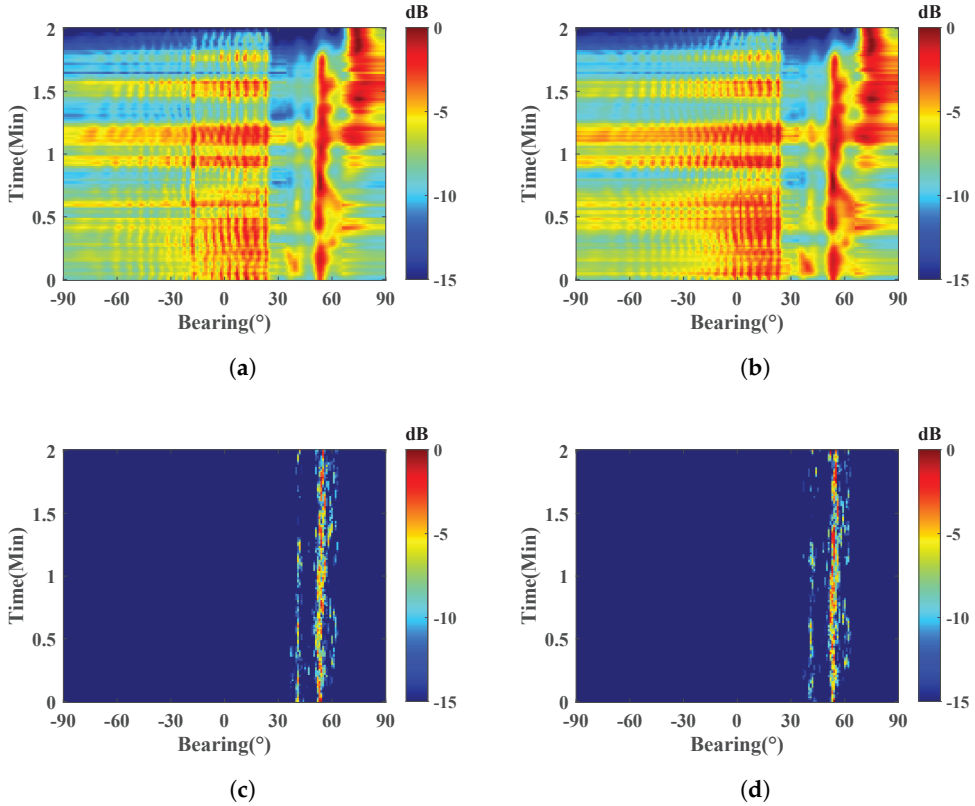
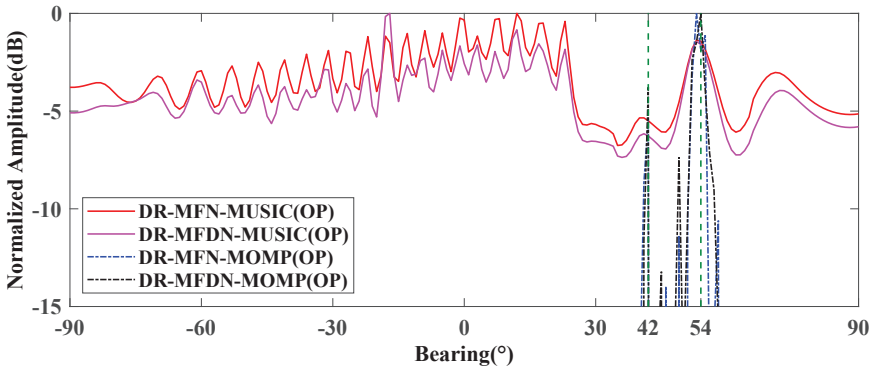


Figure 15. *Cont.*



(**e**)

Figure 15. BTRs of the (**a**) DR-MFN-MUSIC(OP), (**b**) DR-MFDN-MUSIC(OP), (**c**) DR-MFN-MOMP(OP), and (**d**) DR-MFDN-MOMP(OP) estimators. (**e**) Profile of the results of (**a**–**d**) at t = 57 s. The two green dashed lines in (**e**) represent the directions of weak targets.

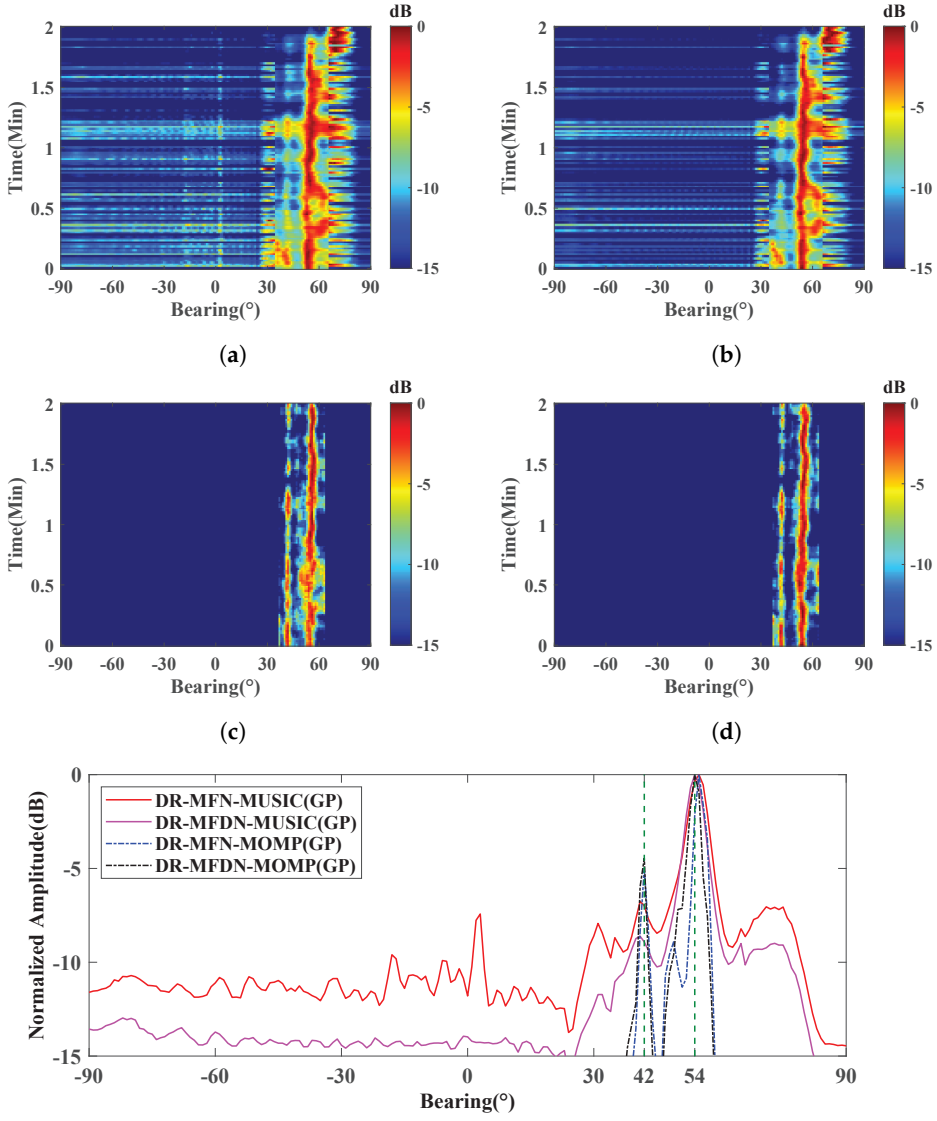


Figure 16. BTRs of the (**a**) DR-MFN-MUSIC(GP), (**b**) DR-MFDN-MUSIC(GP), (**c**) DR-MFN-MOMP(GP), (**d**) DR-MFDN-MOMP(GP) estimators. (**e**) Profile of the results of (**a**–**d**) at t = 57 s. The two green dashed lines in (**e**) represent the directions of weak targets.

5. Conclusions

This study presents a robust DOA estimator for weak targets based on a combination of Gaussian pre-whitening (GP), dimension-reduced matrix filter with deep nulling (DR-MFDN), and multiple-measurement-vector orthogonal matching pursuit (MOMP). The presented method demonstrates superior performance in localizing weak signals in environments with strong interfering sources. The main conclusions can be summarized as follows:

(1) The DR-MFDN effectively suppresses strong interfering sources by forming deep nulls in their directions, thereby significantly improving interference suppression and localization accuracy compared to conventional methods like DR-MFN. This enhancement is crucial for weak target detection.
(2) The Gaussian pre-whitening operation prevents the transformation of white noise into colored noise, preserving the beam-space characteristics and ensuring robustness in DOA estimation. This is particularly important for maintaining accuracy in high-noise environments.
(3) The MOMP algorithm, combined with the DR-MFDN, provides higher resolution and better performance in handling short snapshots compared to traditional algorithms such as MUSIC. This makes the proposed estimator more suitable for real-time applications.
(4) Experimental results from both simulations and sea trials demonstrate that the presented DR-MFDN-MOMP(GP) estimator outperforms existing methods in terms of interference suppression, localization accuracy, and computational efficiency. The method is highly cost-effective and suitable for practical implementation.

Future work may focus on further optimizing computational efficiency and addressing limitations in scenarios with small angular separations between targets and interference sources.

Author Contributions: Conceptualization, S.W. and H.W.; methodology, S.W.; software, H.W.; validation, Z.B., H.W. and S.C.; formal analysis, Z.B.; investigation, H.W.; resources, H.W.; data curation, H.W.; writing—original draft preparation, S.W.; writing—review and editing, W.G. and H.W.; visualization, Z.B.; supervision, S.C.; project administration, S.C., P.S. and B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China under grant nos. 52071309 and 52001296.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Acknowledgments: The authors thank Jie Pang for the insightful discussions and instructive guidance.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BTRs	Bearing-Time Records
CMF	Conventional matrix filter
DOA	Direction-of-arrival
DR-MFDN	Dimension reduced matrix filter with deep nulling
DR-MFN	Dimension-reduced matrix filter with nulling
GP	Gaussian pre-whitening
INR	Interference-to-noise ratio
MMV	Multiple-measurement-vector

MOMP	Multiple-measurement-vector orthogonal matching pursuit
MUSIC	Multiple signal classification
MVDR	Minimum variance distortionless response
NP	None pre-whitening
OMP	Orthogonal matching pursuit
OP	Orthogonal pre-whitening
QMFs	Quiescent matrix filters
RMSE	Root-mean-square errors
SA	Stopband attenuation
SNR	Signal-to-noise ratio
SpSF	Sparse spectrum fitting
STFT	Short-time Fourier transform

Appendix A

The application of the GP operation serves two purposes: preserving the characteristics of beam space and preventing the transformation of white noise into colored noise. The former is evident from Equation (12). Therefore, Appendix A focuses on proving the latter, namely, the covariance matrix of array noise with the GP operation.

The matrix $\mathbf{G}_l^H$ can be rewritten as a combination of column vectors,

$$\mathbf{G}_l^H = [\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_M] \in C^{M' \times M} \tag{A1}$$

where $\mathbf{g}_i \in \mathbb{C}^{M' \times 1}, i = 1, \cdots, M$.

By defining $\mathbf{\Omega} = \sum_{i=1}^{M'} \mathbf{g}_i \mathbf{g}_i^H \in \mathbb{C}^{M' \times M'}$, there is

$$\mathbf{B}_l^H \mathbf{B}_l = \begin{bmatrix} \mathbf{g}_{a1}^H \mathbf{\Omega} \mathbf{g}_{a1} & \mathbf{g}_{a1}^H \mathbf{\Omega} \mathbf{g}_{a2} & \cdots & \mathbf{g}_{a1}^H \mathbf{\Omega} \mathbf{g}_{aM'} \\ \mathbf{g}_{a2}^H \mathbf{\Omega} \mathbf{g}_{a1} & \mathbf{g}_{a2}^H \mathbf{\Omega} \mathbf{g}_{a2} & \cdots & \mathbf{g}_{a2}^H \mathbf{\Omega} \mathbf{g}_{aM'} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_{aM'}^H \mathbf{\Omega} \mathbf{g}_{a1} & \cdots & \cdots & \mathbf{g}_{aM'}^H \mathbf{\Omega} \mathbf{g}_{aM'} \end{bmatrix} \tag{A2}$$

As $\mathbf{\Omega}$ is a symmetric matrix, $\mathbf{\Omega}$ can be rewritten as

$$\mathbf{\Omega} = \mathbf{Q}_\Omega \mathbf{\Lambda}_\Omega \mathbf{Q}_\Omega^H \tag{A3}$$

where $\mathbf{Q}_\Omega \in \mathbb{C}^{M' \times M'}$ is the eigenvector matrix and also an orthogonal matrix; $\mathbf{\Lambda}_\Omega$ is the eigenvalue matrix and is given by

$$\mathbf{\Lambda}_\Omega = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \lambda_{M'} \end{bmatrix} \tag{A4}$$

Define $\mathbf{G}_a' = \mathbf{G}_a \mathbf{Q}_\Omega = [\mathbf{g}_{a1}', \mathbf{g}_{a2}', \cdots, \mathbf{g}_{aM'}']^T \in \mathbb{C}^{M' \times M'}$. Based on the fact that an orthogonal matrix multiplied by a Gaussian matrix does not change its characteristics and therefore $\mathbf{g}_{ai}'(m)$ and $\mathbf{g}_{ai}(m)$ are identically distributed, Equation (A2) can be rewritten as

$$\mathbf{B}_l^H \mathbf{B}_l = \begin{bmatrix} \sum_{m=1}^{M'} \lambda_m \mathbf{g}_{a1}'^2(m) & \sum_{m=1}^{M'} \lambda_m \mathbf{g}_{a1}'(m)\mathbf{g}_{a2}'(m) & \cdots & \sum_{m=1}^{M'} \lambda_m \mathbf{g}_{a1}'(m)\mathbf{g}_{aM'}'(m) \\ \sum_{m=1}^{M'} \lambda_m \mathbf{g}_{a2}'(m)\mathbf{g}_{a1}'(m) & \sum_{m=1}^{M'} \lambda_m \mathbf{g}_{a2}'^2(m) & \cdots & \sum_{m=1}^{M'} \lambda_m \mathbf{g}_{a2}'(m)\mathbf{g}_{aM}'(m) \\ \vdots & \vdots & \ddots & \vdots \\ vspace3pt & & & \\ sum_{m=1}^{M'} \lambda_m \mathbf{g}_{aM'}'(m)\mathbf{g}_{a1}'(m) & \cdots & \cdots & \sum_{m=1}^{M'} \lambda_m \mathbf{g}_{aM'}'^2(m) \end{bmatrix} \tag{A5}$$

The characteristic function of the product sum of two n-dimensional Gaussian vectors with zero means is denoted as [33]

$$\Psi_X(\omega) = \left(\frac{1}{1 - 2j\omega\rho\sigma_1\sigma_2 + \sigma_1^2\sigma_2^2\omega^2(1-\rho^2)} \right)^{n/2} \tag{A6}$$

where X is the inner product of two n-dimensional Gaussian vectors; ω is the conjugate of the Fourier transform of X; ρ is the correlation coefficient of the two Gaussian vectors; a $\sigma_1 dn$ and σ_2 are the variances obeyed by the two Gaussian vectors, respectively. On the basis of the relationship between the eigenfunctions and the origin moments,

$$j^{-k} \frac{d^k \Psi_X(\omega)}{d\omega^k}\bigg|_{\omega=0} = n\mathbb{E}\left[X^k\right] \tag{A7}$$

where k is the order of moment.

After an algebraic simplification, there is

$$\mathbb{E}[X] = \rho\sigma_1\sigma_2 \tag{A8}$$

Based on the relationship between $\mathbf{g}_{ai}'(m)$ and $\mathbf{g}_{ai}(m)$, as well as Equation (A8), it can be deduced that

$$\mathbb{E}[\mathbf{g}_{ai}'^H \mathbf{g}_{ai}'] = \mathbb{E}[\mathbf{g}_{ai}^H \mathbf{g}_{ai}] = \sigma_l^2 \tag{A9}$$

$$\mathbb{E}[\mathbf{g}_{ai}^H \mathbf{\Omega} \mathbf{g}_{ai}] = \frac{1}{M'}\mathbb{E}[\mathbf{g}_{ai}'^H \mathbf{g}_{ai}'] \sum_{m=1}^{M'} \lambda_m = \frac{\sigma_l^2}{M'} \sum_{m=1}^{M'} \lambda_m \tag{A10}$$

When the vectors of the columns of the matrix $\mathbf{G}_a$ satisfy $\|\mathbf{g}_{ai}\|_2 = 1 (i = 1, \cdots, M')$, there is $\sum_{m=1}^{M'} \lambda_m = M'$. Equation (A10) can be further written as

$$\mathbb{E}[\mathbf{g}_{ai}^H \mathbf{\Omega} \mathbf{g}_{ai}] = \sigma_l^2 \tag{A11}$$

By Equations (A5) and (A11),

$$\mathbf{B}_l^H \mathbf{B}_l = \sigma_l^2 \mathbf{I}_{M' \times M'} \tag{A12}$$

The array covariance matrix obtained after linear transformation with matrix $\mathbf{B}_l$ on the array data is given by

$$\begin{aligned} \mathbf{R}_{\tilde{\mathbf{z}}} &= \mathbf{B}_l^H \mathbf{R}_l \mathbf{B}_l \\ &= \mathbf{B}_l^H (\mathbf{A}(f_l, \mathbf{\Theta}) \mathbf{R}_{\mathbf{s}}^l \mathbf{A}^H(f_l, \mathbf{\Theta}) + \mathbf{R}_{\mathbf{e}}^l) \mathbf{B}_l \\ &= \mathbf{B}_l^H \mathbf{A}(f_l, \mathbf{\Theta}) \mathbf{R}_{\mathbf{s}}^l \mathbf{A}^H(f_l, \mathbf{\Theta}) \mathbf{B}_l + \sigma_l^2 \mathbf{I}_{M' \times M'} \end{aligned} \tag{A13}$$

Thus, Equation (13) is proved.

References

1. Vaccaro, R.; Harrison, B. Optimal matrix-filter design. *IEEE Trans. Signal Process.* **1996**, *44*, 705–709. [CrossRef]
2. Vaccaro, R.J.; Harrison, B.F. Matrix filters for short data records. In *Proceedings of the 28th Conference on Information Sciences and Systems*; Princeton University: Princeton, NJ, USA, 1994.
3. Vaccaro, R.; Harrison, B. Matrix filters for passive sonar. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (Cat. No.01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; Volume 5, pp. 2921–2924. [CrossRef]
4. Vaccaro, R.J.; Chhetri, A.; Harrison, B.F. Matrix filter design for passive sonar interference suppression. *J. Acoust. Soc. Am.* **2004**, *115*, 3010–3020. [CrossRef]
5. MacInnes, C. Source localization using subspace estimation and spatial filtering. *IEEE J. Ocean Eng.* **2004**, *29*, 488–497. [CrossRef]
6. Yan, S.; Ma, Y. Matched field noise suppression: A generalized spatial filtering approach. *Chin. Sci. Bull.* **2004**, *49*, 2220–2223. [CrossRef]
7. Yan, S.; Hou, C.; Ma, X. Matrix spatial prefiltering approach for direction-of-arrival estimation. *Chin. J. Acoust.* **2007**, *32*, 151–157.
8. Han, D.; Zhang, X.H. Optimal Matrix Filter Design with Application to Filtering Short Data Records. *IEEE Signal Process. Lett.* **2010**, *17*, 521–524. [CrossRef]
9. Hassanien, A.; Elkader, S.; Gershman, A.; Wong, K. Convex optimization based beam-space preprocessing with improved robustness against out-of-sector sources. *IEEE Trans. Signal Process.* **2006**, *54*, 1587–1595. [CrossRef]
10. Yang, Y.; Zhang, Y.; Yang, L. Wideband sparse spatial spectrum estimation using matrix filter with nulling in a strong interference environment. *J. Acoust. Soc. Am.* **2018**, *143*, 3891–3898. [CrossRef]
11. Feng, J.; Yang, Y.; Sun, C. Adaptive spatial matrix filter design with application to DOA estimation. *J. Syst. Simul.* **2007**, *19*, 4798–4802.
12. Wang, X.; Amin, M.; Wang, X.; Cao, X. Sparse Array Quiescent Beamformer Design Combining Adaptive and Deterministic Constraints. *IEEE Trans. Antennas Propag.* **2017**, *65*, 5808–5818. [CrossRef]
13. Wang, X.; Aboutanios, E. Adaptive Reduced-Dimensional Beamspace Beamformer Design by Analogue Beam Selection. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 4350–4354. [CrossRef]
14. Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280. [CrossRef]
15. Yan, S. *Optimal Array Signal Processing: Modal Array Processing and Direction-of-Arrival Estimation*; Science Press: Beijing, China, 2018; pp. 214–223.
16. Zoltowski, M.; Kautz, G.; Silverstein, S. Beamspace Root-MUSIC. *IEEE Trans. Signal Process.* **1993**, *41*, 344–364. [CrossRef]
17. Gershman, A. Direction finding using beamspace root estimator banks. *IEEE Trans. Signal Process.* **1998**, *46*, 3131–3135. [CrossRef] [PubMed]
18. Karabulut, G.; Kurt, T.; Yongacoglu, A. Angle of arrival detection by matching pursuit algorithm. In Proceedings of the IEEE 60th Vehicular Technology Conference, 2004. VTC2004-Fall. 2004, Los Angeles, CA, USA, 26–29 September 2004; Volume 1, pp. 324–328. [CrossRef]
19. Karabulut, G.Z.; Kurt, T.; Yongaçoglu, A. Estimation of directions of arrival by matching pursuit (EDAMP). *EURASIP J. Wirel. Commun. Netw.* **2005**, *2005*, 197–205. [CrossRef]
20. Cotter, S.F. Multiple snapshot matching pursuit for direction of arrival (DOA) estimation. In Proceedings of the 2007 15th European Signal Processing Conference, Poznan, Poland, 3–7 September 2007; pp. 247–251.
21. Malioutov, D.; Cetin, M.; Willsky, A. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Trans. Signal Process.* **2005**, *53*, 3010–3022. [CrossRef]
22. Fuchs, J.J. On the application of the global matched filter to DOA estimation with uniform circular arrays. *IEEE Trans. Signal Process.* **2001**, *49*, 702–709. [CrossRef]
23. Model, D.; Zibulevsky, M. Signal reconstruction in sensor arrays using sparse representations. *Signal Process.* **2006**, *86*, 624–638. [CrossRef]
24. Gorodnitsky, I.; Rao, B. Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Trans. Signal Process.* **1997**, *45*, 600–616. [CrossRef]
25. Zdunek, R.; Cichocki, A. Improved M-FOCUSS Algorithm with Overlapping Blocks for Locally Smooth Sparse Signals. *IEEE Trans. Signal Process.* **2008**, *56*, 4752–4761. [CrossRef]
26. Zheng, J.; Kaveh, M. Sparse Spatial Spectral Estimation: A Covariance Fitting Algorithm, Performance and Regularization. *IEEE Trans. Signal Process.* **2013**, *61*, 2767–2777. [CrossRef]
27. Yang, L.; Yang, Y.; Zhu, J. Source localization based on sparse spectral fitting and spatial filtering. In Proceedings of the OCEANS 2016 MTS/IEEE Monterey, Monterey, CA, USA, 19–23 September 2016; pp. 1–4. [CrossRef]
28. Chen, J.; Huo, X. Theoretical Results on Sparse Representations of Multiple-Measurement Vectors. *IEEE Trans. Signal Process.* **2006**, *54*, 4634–4643. [CrossRef]

29. Tropp, J.A.; Gilbert, A.C. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *IEEE Trans. Inf. Theory* **2007**, *53*, 4655–4666. [CrossRef]
30. Bian, Z.; Wang, H.; Gao, W.; Song, P. DOA Estimation of Broadband Sources Using Dimension-Reduced Matrix Filter with Deep Nulling in a Strong Interference Environment. *J. Phys. Conf. Ser.* **2024**, *2718*, 012094. [CrossRef]
31. Trees, H.L.V. Adaptive Beamformers. In *Optimum Array Processing*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2002; Chapter 7, pp. 710–916. [CrossRef]
32. Lee, H.; Wengrovitz, M. Resolution threshold of beamspace MUSIC for two closely spaced emitters. *IEEE Trans. Acoust. Speech Signal Process.* **1990**, *38*, 1545–1559. [CrossRef]
33. Simon, M.K. *Probability Distributions Involving Gaussian Random Variables: A Handbook for Engineers and Scientists*; Springer: Boston, MA, USA, 2002; pp. 49–60. [CrossRef]
34. Forster, P.; Vezzosi, G. Application of spheroidal sequences to array processing. In Proceedings of the ICASSP '87 IEEE International Conference on Acoustics, Speech, and Signal Processing, Dallas, TX, USA, 6–9 April 1987; Volume 12, pp. 2268–2271. [CrossRef]
35. Van Veen, B.; Williams, B. Structured covariance matrices and dimensionality reduction in array processing. In Proceedings of the Fourth Annual ASSP Workshop on Spectrum Estimation and Modeling, Minneapolis, MN, USA, 3–5 August 1988; pp. 168–171. [CrossRef]
36. Capon, J. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* **1969**, *57*, 1408–1418. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

Sequential Multimodal Underwater Single-Photon Lidar Adaptive Target Reconstruction Algorithm Based on Spatiotemporal Sequence Fusion

Tian Rong [1], Yuhang Wang [2], Qiguang Zhu [2], Chenxu Wang [1,3,*], Yanchao Zhang [3], Jianfeng Li [3], Zhiquan Zhou [3] and Qinghua Luo [3]

[1] Aerospace College, Harbin Institute of Technology, Harbin 150001, China; 22b921022@stu.hit.edu.cn
[2] College of Information Science and Engineering, Yanshan University, Qinhuangdao 066000, China; q17863136808@163.com (Y.W.); luoj1379@outlook.com (Q.Z.)
[3] College of Information Science and Engineering, Harbin Institute of Technology (Weihai), Weihai 264209, China; zhangyanchao66@sina.com (Y.Z.); lijianfeng@hit.edu.cn (J.L.); zzq@hitwh.edu.cn (Z.Z.); luoqinghua081519@163.com (Q.L.)
* Correspondence: wangchenxu@hit.edu.cn

Abstract: For the demand for long-range and high-resolution target reconstruction of slow-moving small underwater targets, research on single-photon lidar target reconstruction technology is being carried out. This paper reports the sequential multimodal underwater single-photon lidar adaptive target reconstruction algorithm based on spatiotemporal sequence fusion, which has strong information extraction and noise filtering ability and can reconstruct the target depth and reflective intensity information from complex echo photon time counts and spatial pixel relationships. The method consists of three steps: data preprocessing, sequence-optimized extreme value inference filtering, and collaborative variation strategy for image optimization to achieve high-quality target reconstruction in complex underwater environments. Simulation and test results show that the target reconstruction method outperforms the current imaging algorithms, and the built single-photon lidar system achieves underwater lateral and distance resolution of 5 mm and 2.5cm@6AL, respectively. This indicates that the method has a great advantage in sparse photon counting imaging and possesses the capability of underwater target imaging under the background of strong light noise. It also provides a good solution for underwater target imaging of small slow-moving targets with long-distance and high-resolution.

Keywords: ocean; optical detection; slow-moving small targets; single-photon lidar; target reconstruction

Academic Editors: Xuebo Zhang, Haiyong Xu, Jingyue Lu, Lei Zhang, Marc Pinto and Farman Ali

Received: 28 November 2024
Revised: 12 January 2025
Accepted: 13 January 2025
Published: 15 January 2025

Citation: Rong, T.; Wang, Y.; Zhu, Q.; Wang, C.; Zhang, Y.; Li, J.; Zhou, Z.; Luo, Q. Sequential Multimodal Underwater Single-Photon Lidar Adaptive Target Reconstruction Algorithm Based on Spatiotemporal Sequence Fusion. *Remote Sens.* **2025**, *17*, 295. https://doi.org/10.3390/rs17020295

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the deepening of ocean exploration, long-range and high-resolution imaging of small targets has become the basic technology to support many other ocean exploration techniques and has received great attention [1–7]. Traditional optical detection methods suffer from light attenuation and low contrast in the underwater imaging environment, leading to image quality degradation [8–10]. Current mainstream acoustic detection faces limitations of acoustic wave propagation speed and resolution, as well as problems of signal attenuation and distortion caused by the non-uniformity of seawater medium, which reduces the reliability and practicability [11–13]. The emergence of single-photon lidar provides a new way for underwater detection [14–17], which becomes an effective way

to solve the problem of detecting targets in low-light and high-light attenuation intensity environments, and has aroused extensive research interests.

As the key influencing factors of single-photon lidar, the iteration of imaging algorithms and the improvement of hardware systems complement each other. At present, in the field of single-photon lidar development, many researchers and scholars have made good achievements in the hardware system [15,18–23]. However, in practical applications, photon-level detection sensitivity brings about an improvement in detection capability, while noise-induced photon events also cause a problem of increased noise points and bring about randomness of detection, coupled with a lack of information caused by severe optical attenuation. This interferes with the selection and reconstruction of the target information and ultimately affects the detection results [24–26]. So, robust imaging algorithms are indispensable for single-photon detection systems. According to the differences in imaging mechanisms, data processing means, and technical implementations the existing single-photon lidar imaging algorithms are mainly divided into two major categories, namely, neural network-based imaging approaches [27,28], and imaging algorithms based on traditional modeling approaches [29,30]. The neural network-based approach is able to learn nonlinear mapping relationships from complex data by taking advantage of data driven through deep learning techniques, largely detaching itself from a priori experience and showing excellent imaging results. However, the problems of data dependence, high demand for computational resources, limited generalization ability, and insufficient retention of edge information have seriously developed the single-photon imaging based on neural network imaging approach [31–38]. Modeling-based imaging algorithms have high computational efficiency, flexibility in parameter tuning, strong generalization ability, robustness, and cost-effectiveness and are particularly suitable for applications with limited resources or requiring fast response due to their low computational requirements and easy integration [39–41]. Among the current single-photon imaging methods based on modeling approaches, non-fixed pixel scan-time reconstruction methods, represented by first-photon methods [42,43], first-photon group [44,45], and other "firstness" photon information reconstruction targeting methods, are better in terms of resource optimization, but the data processing is more complicated and the system requirements are higher. Also, the current algorithms are not well adapted to complex environments, especially underwater environments, and the characteristics of manual parameterization [42–48] are difficult to adapt to changing environments. Fixed-pixel acquisition time reconstruction methods, although sacrificing some flexibility and resource efficiency, are suitable for standardized imaging due to their consistency and processing simplicity, and methods such as the sequential two-mode fusion imaging algorithm (STMF) also compensate for the echo signal variability [49]. However, due to median filtering [44,45,50,51] the intrinsic spatial and temporal connections of the single-photon data and the distributional characteristics of the signal are ignored. Also, a large amount of fringe information and high-frequency components are lost, and in underwater environments with severe backscattering, non-localized spatial pixel filtering, which is excellent in filtering performance, loses its robustness. Numerous algorithms are applicable to gaze imaging systems [52–55], which further reduce the spatial resolution and flexibility of single-photon lidar systems and increase their information acquisition cost.

An efficient single-photon lidar imaging algorithm is urgently needed for long-range and high-resolution imaging detection of underwater slow-moving small targets. In this study, we report on the sequential multimodal underwater single-photon lidar adaptive target reconstruction algorithm based on spatiotemporal sequence fusion (SUARF), which is based on the pixel-scanning imaging principle. This takes advantage of the fundamental property that the noise photons in the echo signal are differently distributed with the target

information photons in the temporal histogram in decomposing the pulse-accumulated single-photon raw signal, and with the help of an optimization strategy, it settles the target depth and reflectivity information. Compared with algorithms such as the STMF and the first photon group with excellent imaging performance, this method does not require manual parameter adjustment and can efficiently accomplish the target reconstruction task under a variety of challenging imaging conditions, which can provide technical support for underwater target imaging detection.

Our main technical innovations and contributions are as follows:

- We propose the SUARF algorithm consisting of data preprocessing, sequence optimized maximum value inference filtering, and collaborative variation strategy for image optimization to effectively extract target depth and reflection intensity information from complex echo photon time counting. The mask and pixel summation operations are used for data preprocessing to remove the strong backscattering of water near the LiDAR glass window and part of the system noise. Sequence optimization maximum inference filtering combined with matching filter and maximum group filtering can accurately select target depth-reflection information. The collaborative variation strategy integrates the adaptive Wiener–Hopf filter and global morphological constraint into the total variational smoothing constraint to enhance the overall smoothness and effectively deal with abnormal pixel values.
- We built an underwater single-photon LiDAR prototype for testing. The results show that compared with the latest scanning single-photon LiDAR imaging algorithm, the (SSIM, RMSE) of the depth and reflection intensity map was improved by (15%, 57%) and (48%, 71%), respectively. We also proposed an evaluation index R_T to comprehensively evaluate the depth-reflection intensity of the SSIM and RMSE index, which was improved by 24%. All results show that the proposed algorithm has good scene reconstruction performance.

2. SUARF Target Reconstruction Algorithm Model

The SUARF algorithm consists of three steps: data preprocessing, sequence-optimized extreme value inference filtering, and collaborative variation strategy for image optimization to achieve high-quality target reconstruction in complex underwater environments. The schematic diagram of the SUARF algorithm is as in Figure 1.

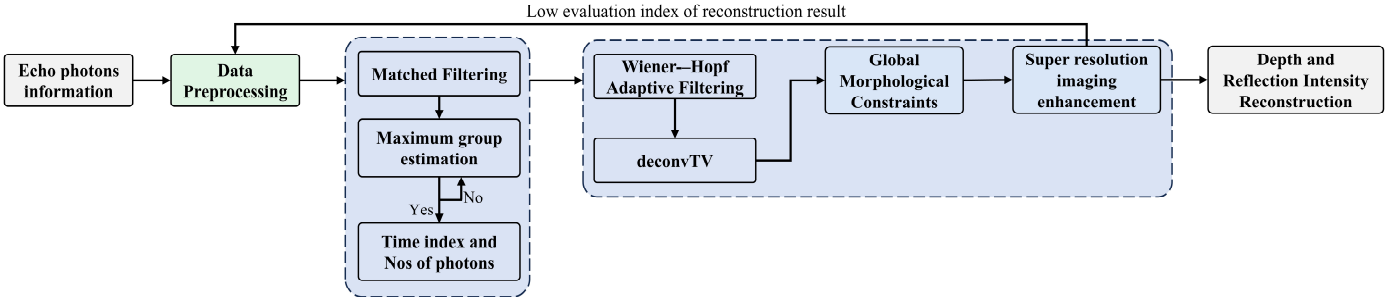


Figure 1. SUARF target reconstruction algorithm flow diagram.

In the original signal of single-photon lidar, the arrival moments of signal photons will be centrally distributed on the time axis. The noise photons do not carry any information about the target, and are independent of each other with respect to the arrival moments of the signal photons, and thus are uniformly distributed in the detection period. According to the above characteristics, the signal photons and noise photons can be initially decomposed using sequence-optimized extreme value inference filtering, and then the outliers can be filtered out using the collaborative variation strategy, while the global morphological

constraints can be used to maintain the structural features. The flowchart of the algorithm is shown in Algorithm 1.

Algorithm 1 Overall framework: SUARF

Input: Photons echo data

Output: Depth and Reflection Intensity Information

/// First Step:
(1)Data coordinate conversion
(2)Data cleaning ← Select the mask

// Second Step:
$y(t) = s^*(-t) \quad r(t)$ // Matched filtering
while $(x, y) \in (X, Y)$ **do**
  if *G is the largest photons group* **then**
    $T_{x,y}^{index} \leftarrow T_{x,y}^{groupindex,max}$ // Time bin information
    $N_{x,y}^{index} \leftarrow \sum N_{x,y}^{groupindex}$ // Reflection intensity information
    $Depth \Leftrightarrow \alpha T_{x,y}^{index}$
    $Reflection\ intensity \Leftrightarrow N_{x,y}^{index}$ // Time bin vs. Depth
  end
end

// Third Step: Wiener-Hopf Adaptive Filtering + deconvTV

$$H_{Depth}(f) = \frac{S_{y_{Depth}x}(f)}{S_{xx_{Depth}}(f) + \sigma^2_{v_{Depth}}} \qquad H_{Refl}(f) = \frac{S_{y_{Refl}x}(f)}{S_{xx_{Refl}}(f) + \sigma^2_{v_{Refl}}}$$

$\rightarrow \mathrm{L} = \mathtt{deconvTV}(P^* \mathcal{F}^{-1}\{H_{\mathrm{p}}(f)\})$

→ Global Morphological Constraints

// Fourth Step: Real-ESRGAN, Super resolution enhancement

First, the raw echo signal is converted into pixel space-arranged data, and the mask length is set according to the system characteristics to filter out the noise generated in the echo signal within the system. Second, matched filtering and extremely large group value estimation are used to obtain the depth and reflectivity information of the acquired target scene. In Algorithm1, $s^*(-t)$ denotes that the impulse response of the matched filter is the time inversion and conjugate of the transmitted impulse signal; $y(t)$ denotes the output of the matched filter; $r(t)$ is the received echo signal; $T_{x,y}^{index}$ and $N_{x,y}^{index}$ denote the time value of the current pixel and the number of the photons in the group; $T_{x,y}^{groupindex,max}$ and

$N_{x,y}^{groupindex}$ denote the time of the bin with the largest number of the photons in the group and the number of the photons in each bin in the group, respectively; α is the time-distance conversion coefficient; and Depth and Reflection intensity are the preliminary reconstructed depth and reflection intensity maps of single-photon lidar, respectively. After that, the Wiener–Hopf filter [55,56] with improved decisionTV [57] smoothing constraints is used to remove anomalies while retaining the edge information. $H_{\mathrm{Depth}}(f)$ and $H_{\mathrm{Refl}}(f)$ denote the frequency response of the Wiener filter for the Depth and Reflection intensity signals, respectively; $S_{y_{\mathrm{Depth}}x}(f)$ and $S_{y_{\mathrm{Refl}}x}(f)$ denote the frequency response of the input Depth and Reflection intensity signals with respect to the desired Depth and Reflection intensity signals, respectively; $\sigma^2_{v_{\mathrm{Depth}}}$, $\sigma^2_{v_{\mathrm{Refl}}}$ denote the noise variance of the Depth and Reflection intensity signals, respectively; P and L refer to the Depth and Reflection intensity signals before and after processing, respectively; and $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. Finally, the clarity and visual quality of the images are enhanced by Enhanced Super-Resolution Generative Adversarial Networks (Real-ESRGAN). The reason for choosing this network is that it currently provides clarity enhancement functions that do not require a large amount of data to train for the time being and can maintain good results.

2.1. Data Preprocessing

The data acquired by the single-photon lidar are stacked in one dimension and need to be converted into a two-dimensional spatial arrangement based on the scanning method to make the data easier to process and understand. In addition, since the presence of large noise within the system interferes with target information extraction, data cleaning is used to remove the accumulation of noisy photons on the time axis due to the system. Second, all photon histograms are assembled to find the peak moments to obtain the approximate distance of the target scene for imaging display. The data preprocessing schematic of the SUARF is shown in Figure 2.

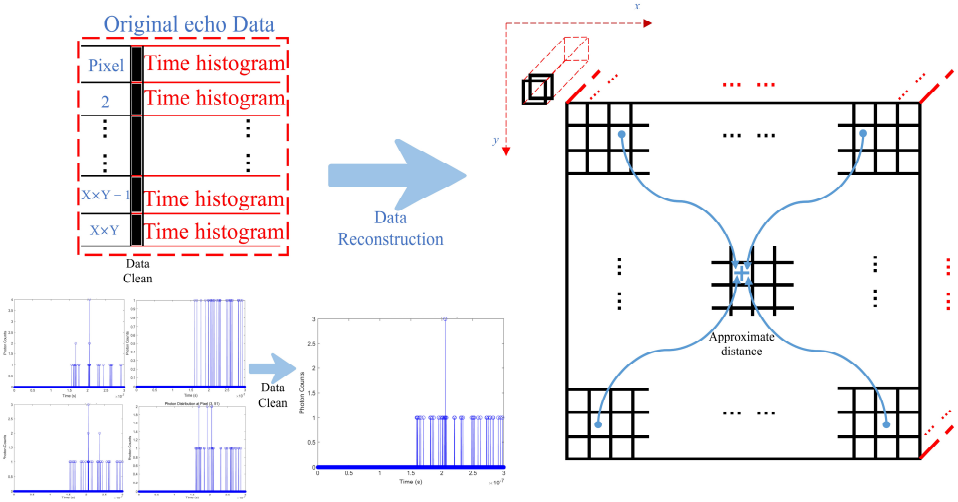


Figure 2. Schematic of data preprocessing of SUARF algorithm.

After setting the mask length BlindBin according to the empirical value of the system, the original data can be covered with this mask to shield specific elements, and the process can be described as Equation (1):

$$B_{\mathrm{masked}}(:,:,k) = \begin{cases} B(:,:,k) & \text{if } k > \mathrm{BlindBin} \\ 0 & \text{if } k \leq \mathrm{BlindBin} \end{cases} \tag{1}$$

B_{masked} is the echo data after data reconstruction and cleaning, and k is the index in the direction of the time axis. In Figure 2, it can be seen that after data preprocessing, the noise in the front part of the time axis has been filtered out. However, it should be noted that the length of BlindBin needs to be set according to the system; if it is arbitrarily set and set too large, it will cause the system to produce a large blind zone, and if it is set too small, the noise caused by the system will affect the determination of the time index and the number of photons.

2.2. Sequence Optimized Maximum Value Inference Filtering

The signal of each pixel after data preprocessing is still time series information, which needs to be calculated by an inference algorithm to find the most compatible time and photon values. In this paper, we optimize previous research work [49] to obtain the information, and the processing schematic is shown in Figure 3.

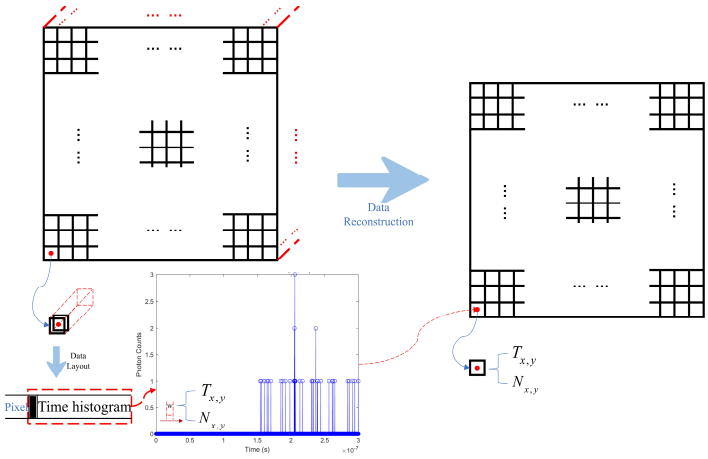


Figure 3. Schematic diagram of the sequence optimized maximum value inference filtering.

Assuming that the input signal is a superposition of the desired signal (target signal) and noise, where s(t) is the desired signal, n(t) is the noise, and the input signal x(t) can be denoted by Equation (2):

$$x(t) = s(t) + n(t) \tag{2}$$

The impulse response of the matched filter is the time reversal and conjugate complex of the desired signal $s(t)$ to make the filter most sensitive in detecting a particular signal. The designed filter and the output signal y(t) obtained after convolution with the input signal can be expressed, respectively, by Equations (3) and (4):

$$h(t) = s^*(-t) \tag{3}$$

$$y(t) = x(t) * h(t) = (s(t) + n(t)) * s^*(-t) \tag{4}$$

The amplitude of the output signal is maximized when the desired signal in the input signal exactly matches the impulse response of the filter, thus enabling the detection of the signal. And the convolution of noise n(t) with $s^*(-t)$ does not produce the same correlation peak as the desired signal $s(t)$, so the effect of noise in the output signal is suppressed, and the target signal in the original signal is more easily detected. Assuming that the size of the data volume after preprocessing and matched filtering is $X \times Y \times Z$, the first two dimensions denote the horizontal and vertical coordinates of the pixel, respectively, and the third dimension denotes the number of bin frames in the echo signal of each pixel, then

the number of bins and intra-bin photons in each pixel of the echo data are respectively expressed by Equations (5) and (6):

$$\left\{T_{x,y}^{1}, T_{x,y}^{2}, T_{x,y}^{3}, \ldots T_{x,y}^{Z}\right\} \tag{5}$$

$$\left\{N_{x,y}^{1}, N_{x,y}^{2}, N_{x,y}^{3}, \ldots N_{x,y}^{Z}\right\} \tag{6}$$

We select the width of the group (i.e., the length of the sliding window w. According to the experimental observation and the characteristics of the signal noise distribution, w = 5 bin can take into account the selection of the signal and avoid the selection of the noise peaks), and slide the group along the time axis with a step of 1 bin. Then we select the group with the highest number of photons in the group as the group where the target signal is located, then the bin and the number of photons in the bin of the group's echo data are numbered, respectively, by Equations (7) and (8); the time value and photon value of the pixel are calculated, respectively, by Equations (9) and (10):

$$\left\{T_{x,y}^{\phi,1}, T_{x,y}^{\phi,2}, T_{x,y}^{\phi,3}, \ldots T_{x,y}^{\phi,\mathrm{w}}\right\} \tag{7}$$

$$\left\{N_{x,y}^{\phi,1}, N_{x,y}^{\phi,2}, N_{x,y}^{\phi,3}, \ldots N_{x,y}^{\phi,\mathrm{w}}\right\} \tag{8}$$

$$T_{x,y} = \max\left\{T_{x,y}^{\phi,1}, T_{x,y}^{\phi,2}, T_{x,y}^{\phi,3}, \ldots T_{x,y}^{\phi,\mathrm{w}}\right\} \tag{9}$$

$$N_{x,y} = \sum\left\{N_{x,y}^{\phi,1}, N_{x,y}^{\phi,2}, N_{x,y}^{\phi,3}, \ldots N_{x,y}^{\phi,\mathrm{w}}\right\} \tag{10}$$

In the Equations (7)–(10), $T_{x,y}^{\phi,\mathrm{w}}$ and $N_{x,y}^{\phi,\mathrm{w}}$ denote the number of bin and bin photons in the ϕth group selected in the (x, y) pixel. Considering that there are very few echo photons in the low illumination environment, the time index of the maximum number of photons and the sum of the number of photons in the group are used as the time value $T_{x,y}$ of the pixel, and the value of the photons $N_{x,y}$, respectively:

$$D_{x,y} = c \times T_{x,y}/2 = c \times \max\left\{T_{x,y}^{\phi,1}, T_{x,y}^{\phi,2}, T_{x,y}^{\phi,3}, \ldots T_{x,y}^{\phi,\mathrm{w}}\right\}/2 \tag{11}$$

Combined with the underwater speed of light c, the depth information for that pixel can be obtained by Equation (11). Repeating the operation (Equations (7)–(11)) for each pixel, the depth and reflectance information of the target scene can be preliminarily reconstructed

2.3. Collaborative Variation Strategy for Image Optimization

Due to the complexity of light attenuation in the underwater environment, there is still some missing information and noise interference in the depth and reflectance map of the initial reconstruction. To eliminate this effect, we improve the idea of total variational smoothing constraints, add adaptive filtering and mixed morphological constraint structures to the deconvTV, and by dynamically adjusting the filtering parameters, we expect to suppress the abnormal pixels while better preserving the image edges and textures and protecting the naturalness and visual details of the image. The output processing flowchart of the collaborative variation strategy for image optimization is shown in Algorithm 2.

Algorithm 2 Co-variation imaging strategy

Function Input Image (*fileP ath*)
Define Blur and Regularization Parameters {Create a blur kernel representing the average blur} {Set the regularization parameter balancing deblurring and Parameters}
$H \leftarrow$ create blur kernel(3)
$\mu \leftarrow 0.015$
End of Define Blur and Regularizationnoise suppression
Filtering and regularization {Apply the Wiener-Hopf filter to the Intput image} {Extract the channel data I_c from the image I} {Simulate the blurred image g by convolving I_c with the blur kernel H}
OutputIm $=$ Wiener $-$ Hopf (InputIm)
for each channel c in the image **do**
$I_c \leftarrow$ extract_channel(I, c)
$g \leftarrow I_c H$
$x^{(k+1)}$ {Refine the image x$^{(k+1)}$ using deconvolution}
apply deconvolution($x^{(k)}$, H, μ)
end for
End of Filtering and regularization
Post-Processing {Combine the deconvolved channels to form the restored image I_{restored}}
$I_{\text{restored}} \leftarrow$ merge_channels($x^{(k+1)}$) {Detect edges E in the restored image using an edge detection method}
$E \leftarrow$ detect_edges(I_ restored)
$E_{\text{closed}} \leftarrow$ refine edges(E) {Refine edge boundaries E_{closed} using morphological closing}
End of Post-Processing
Morphological Reconstruction Compute the distance map D from the closed edges} {Reconstruct the im age $I_{\text{reconstructed}}$ based on the distance map D and edges}
$D \leftarrow$ compute distance map(E$_{\text{closed}}$)
$I_{\text{reconstructed}} \leftarrow$ reconstructed_image(E_{closed},D)
End of Morphological Reconstruction
Region Processing {Label connected regions L in the reconstructed image} {Find pixels in P_k in region k} {Determine the most fre- quent value v_k in region k} {Set all pixels in region k to the most frequent value v_k}
$L,\ N \leftarrow$ label_regions$(I_{\text{reconstructed}})$
for each region k from 1 to N **do**
$P_k \leftarrow$ find_pixels(L, k)
$v_k \leftarrow$ find_mode$(I_{\text{restored}}(P_k))$
$I_{\text{restored}}(P_k) \leftarrow v_k$
end for
End of Region Processing
EndFunction

After obtaining the depth and reflectance 2D matrices of the target scene from the sequence optimized maximum value inference filtering, we use Wiener–Hopf adaptive filtering to adaptively adjust the filtering parameters with the help of the local statistical properties of the image, where the behavior of the filter varies in different regions of the image, thus removing anomalous pixel values while preserving the edges and details and providing clean pixel environments for smoothing constraints The filtered depth

image $D_{\text{Wiener}-\text{Hopf}}$ and the reflectance image $R_{\text{Wiener}-\text{Hopf}}$ can be obtained, respectively, by Equations (12) and (13):

$$D_{\text{Wiener}-\text{Hopf}}(i,j) = \frac{\sum_{k=-\infty}^{+\infty}\sum_{l=-\infty}^{+\infty} H(i-k,j-l)\cdot D_{\text{noisy}}(k,l)}{\sum_{k=-\infty}^{+\infty}\sum_{l=-\infty}^{+\infty} H(i-k,j-l)^2 + \frac{\sigma_{nD}^2}{\sigma_D^2(i,j)}} \tag{12}$$

$$R_{\text{Wiener}-\text{Hopf}}(i,j) = \frac{\sum_{k=-\infty}^{+\infty}\sum_{l=-\infty}^{+\infty} H(i-k,j-l)\cdot R_{\text{noisy}}(k,l)}{\sum_{k=-\infty}^{+\infty}\sum_{l=-\infty}^{+\infty} H(i-k,j-l)^2 + \frac{\sigma_{nR}^2}{\sigma_R^2(i,j)}} \tag{13}$$

In Equations (12) and (13), $D_{noisy}\ (i,j)$ and $R_{noisy}\ (i,j)$ are the depth map and the reflection intensity map pixel value at pixel (*i,j*), respectively. (*i,j*) and (*k,l*) are pixel indices in the image matrix, where (*i,j*) denotes the pixel position currently being processed and (*k,l*) denotes the position in the neighboring area covered by the filter. $H(i-k,j-l)$ is the impulse response of the filter (which can be calculated by Equation (14)), σ_{nD}^2, σ_{nR}^2 are the noise power, and $\sigma_D^2\ (i,j)$, $\sigma_R^2\ (i,j)$ are the local variance of the depth map and reflection intensity map at position (*i,j*), respectively:

$$H(e^{j\Omega}) = \frac{S_{xr}(e^{j\Omega})}{S_{xx}(e^{j\Omega}) + S_{nn}(e^{j\Omega})} \tag{14}$$

$S_{xr}\ (e^{j\Omega})$, $S_{xx}\ (e^{j\Omega})$ and $S_{nn}\ (e^{j\Omega})$ are, respectively, the cross-power spectral density between the estimated signal and the desired response, the power spectral density of the signal, and the power spectral density of the noise from the pixel data.

Wiener–Hopf adaptive filtering can effectively suppress noise, but due to the low signal photon counts (average single-pixel photon counts $\not> 1$ in the reconstructed maps at times), it is inferred that too much variance of the filtered values leads to large uncertainty in the reconstruction of the image, and this uncertainty may lead to significant reconstruction errors. To reduce this uncertainty, the smoothing constraint maintains the edge information by minimizing the sum of the absolute values of the image gradients while suppressing the noise, while further mixed morphological reconstruction is employed to enhance the details by highlighting the edges and textures in the image in order to improve the accuracy of the image reconstruction and to enhance the visual effect. For the acquired depth image $D_{\text{Wiener}-\text{Hopf}}$ and reflectance image $R_{\text{Wiener}-\text{Hopf}}$, the deconvTV algorithm is used to stabilize the image reconstruction results. Considering the underwater application scenario of single-photon lidar, we improve the idea of morphological constraints by extracting the global features of the image and fusing them with the local features, and improving the image quality through the assisting deconvTV algorithm by erosion, expansion, and open/close operations, etc. This module (TV-Morphological Restoration Algorithm, TVMR) can be described by the following equation:

$$\text{FinalImg} = \text{MorphologicalReconstruct}(\text{Edges}(\text{deconvTV}(\text{Img}, \text{H}, \text{Opts}))) \tag{15}$$

Img is the pending $D_{\text{Wiener}-\text{Hopf}}$ and $R_{\text{Wiener}-\text{Hopf}}$, deconvTV (Img,*H*,Opts) denotes the processing of the image using full variational deconvolution, where *H* is the point spread function (PSF) and Opts are the algorithmic options (containing regularization parameter μ, initial penalty parameter *rho_r*, maximum number of iterations *max_itr*, regularization parameter *beta* for weight TV paradigm, update constant *gamma*, tolerance *tol*, etc.). Edges denotes Canny edge detection (enhanced edge display) of the image and

Morphological Reconstruct denotes improved morphological reconstruction of the image. Since the system uses Gaussian beams, its PSF can be expressed as (16):

$$\mathrm{PSF}(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{16}$$

PSF(*x*,*y*) denotes the value of the point spread function at position (*x*,*y*) and σ is the standard deviation of the Gaussian beam, then the deconvTV algorithm can be calculated as Equation (17):

$$\begin{cases} D_{\mathrm{deconvTV}} = \mathrm{argmin}_D \left\{ \lambda \left\| D - D_{\mathrm{Wiener-Hopf}} \right\|_2^2 + \tau \|\nabla D\|_1 \right\} \\ R_{\mathrm{deconvTV}} = \mathrm{argmin}_R \left\{ \lambda \left\| R - R_{\mathrm{Wiener-Hopf}} \right\|_2^2 + \tau \|\nabla R\|_1 \right\} \end{cases} \tag{17}$$

In Equation (17), D_{deconvTV} and R_{deconvTV} are the depth image and reflection intensity image after the constraints of the deconvTV algorithm. D and R are the variables in the optimization problem, λ is the regularization parameter of the data fidelity term, τ is the regularization parameter of the total variation term, and ∇ denotes the gradient operator. Further, the mixed morphological reconstruction method with the fusion of global and local features is used, and the processing sequence is (1) edge detection and morphological closure operation, (2) distance transformation and morphological reconstruction, and (3) connected region processing and enhancement. Since the processing steps of depth image and reflection intensity image are the same, D_{deconvTV} and R_{deconvTV} are standardized to be described by I_{restored} instead. The edge detection and mixed morphological closure operations are as Equation (18):

$$E_{\mathrm{closed}} = \mathrm{close}(\mathrm{C}(I_{\mathrm{restored}}), \mathrm{S}(N)) \tag{18}$$

In Equation (18), $\mathrm{C}(I_{\mathrm{restored}})$ denotes the extraction of edges for the image I_{restored}, 'close' denotes the morphological closure operation, and S(N) is used to define the neighborhood of pixels in the image. Then the distance transformation is performed based on the edge map after the closed operation as Equation (19):

$$I_{\mathrm{Reconstructed}} = \mathrm{Reconstruct}(E_{\mathrm{closed}}, \frac{\mathrm{W}(\sim E_{\mathrm{closed}})}{\max(\mathrm{W}(\sim E_{\mathrm{closed}}))}) \tag{19}$$

In Equation (19), W(~E closed) computes the distance transform of the complement of the edge map after the closure operation, max(·) denotes the normalization operation, and 'Reconstruct' denotes the morphological reconstruction based on the distance transform and the edge map. Finally, the connected regions in the reconstructed image are processed to enhance the visual contrast effect. In the process of enhancing the imaging results, we adjust the specific values of the two parameters and observe their imaging results. We found that when regularization parameter τ is set to about 0.015 and morphological constraint size parameter *se* is set to about 10, the image enhancement effect is better, and the details are saved more completely. Of course, this may need to be adjusted slightly in different LiDAR systems.

$$I_{\mathrm{final}} = \widehat{\mathrm{Sharpen}} \left(\bigcup_{k=1}^{\mathrm{num}} (\mathrm{LabelImg} == \mathrm{k} \Rightarrow \mathrm{mode}(\mathrm{I}_{\mathrm{Reconstructed}}(\mathrm{LabeledImg} == \mathrm{k}))) \right) \tag{20}$$

In Equation (20), 'LabelImg==k' denotes the connected region labeled as k, mode(·) denotes all pixel values in the statistical region, $\cup$ denotes all connected regions, and $\widehat{\mathrm{Sharpen}}(\cdot)$ denotes that while the traditional image processing techniques are mainly used,

Real-ESRGAN [58] can be used as an auxiliary tool, and its pre-trained model can be directly used for image clarity, which further improves image resolution and visual effect.

3. Construction of Test Device and Analysis of Land Imaging

We constructed a single-photon lidar system to verify the algorithm imaging effect. Considering the underwater application scenarios of single-photon lidar, the main reason we chose the scanning imaging single-photon lidar system instead of gaze imaging is its high sensitivity and high spatial resolution. The scanning system is able to adapt to different detection needs by adjusting the scanning mode and speed, which is especially important for the changing underwater environment. In addition, single-photon lidar technology, especially single-photon avalanche diode (SPAD)-based systems, offers unique technological advantages in capturing weak signals and high-precision imaging due to its high temporal resolution, high sensitivity, and ease of integration. These advantages make scanning single-photon lidar systems more effective in underwater target detection and imaging.

The main components of the system are a transmitter module, a receiver module, an optical system module, a scanning imaging module, a time-dependent single-photon counter, and a signal control module. The schematic diagram of the system is shown in Figure 4.

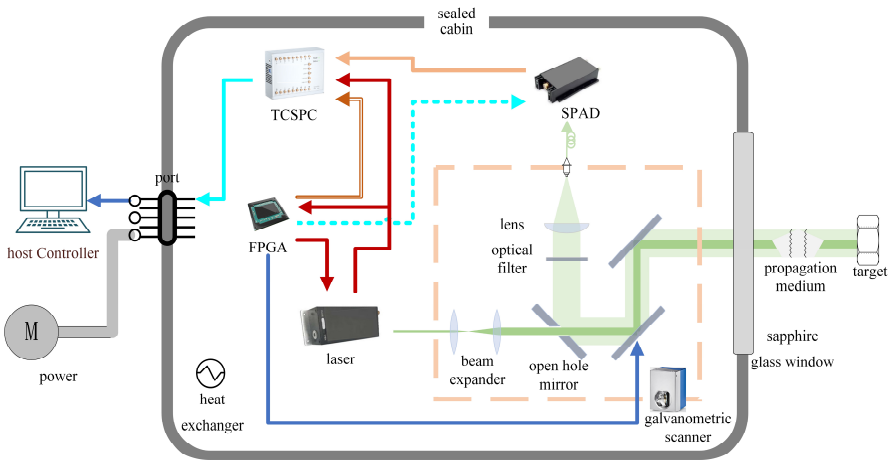
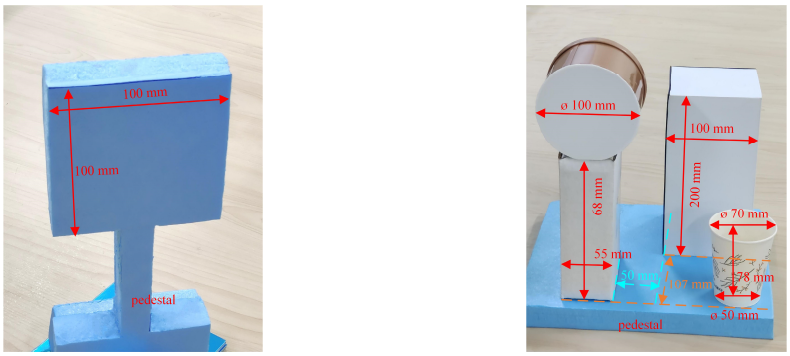


Figure 4. Schematic diagram of the system principle.

The transmitting module uses a 532 nm wavelength solid-state laser as the light source, which is a common green laser suitable for underwater environments, with high stability and reliability. The laser operates in a pulsed mode with a repetition rate of 1–5 kHz, allowing for high-resolution imaging by emitting short bursts of light at regular intervals. The average power of 175 mW and single pulse energy of 35 μJ ensure that the laser can provide sufficient energy for long-distance detection while maintaining the high quality of the beam. The receiving module collects the photons reflected from the target using a flat-convex lens with a focal length of 100 mm and a diameter of 2 inches. This lens focuses the incoming photons onto a multimode fiber with a core diameter of 105 μm and a numerical aperture of 0.22. The fiber then couples the photons to a single-photon detector. A bandpass filter with a central wavelength of 532 nm and a transmittance of 53.58% is used to filter out non-target wavelength light, ensuring that only 532 nm wavelength photons reach the detector. The system adopts a coaxial transceiver combined optical path design, using a 45° open-hole mirror as a transceiver switch. This mirror allows the laser to be transmitted and received in the same optical path, greatly improving the compactness and efficiency of the system.

The scanning imaging module uses a two-axis galvanometer for two-dimensional beam scanning with an optical deflection angle of ±22.5° and an optical resolution of 12 μrad. The galvanometer works by using electromagnetic coils to rotate mirrors, which in turn deflect the laser beam to scan the target area. The time-dependent single-photon counter (TCSPC) module measures the arrival time of each detected photon by recording the time interval between the emission of the laser pulse and the detection of the reflected photon. It converts the electrical signal from the detector into digital time-stamped data, allowing the system to determine the total flight time of the photon and calculate the distance to the target. The signal control module coordinates the gating mode of the single-photon detector, generates the reference time signal required by the TCSPC, and controls the pulse output of the laser. It ensures the synchronization of all components by managing the timing and control signals, thus ensuring the accuracy and reliability of the data.

In the workflow of our system, the 532 nm solid-state laser first emits a pulsed laser. When the photodetector detects the pulsed laser, it outputs an electrical pulse synchronization signal, which is collected by the time-dependent photon counting module and starts timing. The outgoing pulsed laser increases the beam diameter and decreases the divergence angle by shaping the optical path, and changes the direction by the two-axis galvanometer to reach the target point. When the photon returns to the system in the form of diffuse reflection from the target, the propagation direction is changed again through the two-axis galvanometer, and then the incoming echo receiving module is coupled. The SPAD detector generates an electrical pulse signal immediately after receiving the echo photon, and the TCSPC module subtracts the time generated by the electrical signal and the time of the synchronization signal when the pulse laser emits, so as to obtain the total flight time of the outgoing photon; then, we can calculate the numerical distance of the target point. The input and output signals dependent on each module in the system are uniformly managed by the signal control module. The data of the time-dependent photon counting module flows through the host computer and is directly read by the software algorithm for imaging and image enhancement optimization. We first verified the imaging performance of the system in a land trial. The two targets designed in previous work were used to verify the imaging effect, the targets are target A and target B, which are shown in Figure 5.



Target A Target B

Figure 5. Schematic diagram of the target.

In Figure 5, target A is a regular 10 × 10 cm square, and in target B there are circles, rectangles, etc., distributed in different depth areas. It should be noted that, in order to ensure the randomness and fairness of the test, the data of two kinds of targets were randomly selected from the results of multiple data acquisition for imaging analysis. By varying the detection distance, the number of emitted pulses, and the intensity of the

external light intensity, respectively, the imaging results of the SUARF algorithm are as Figures 6–11.

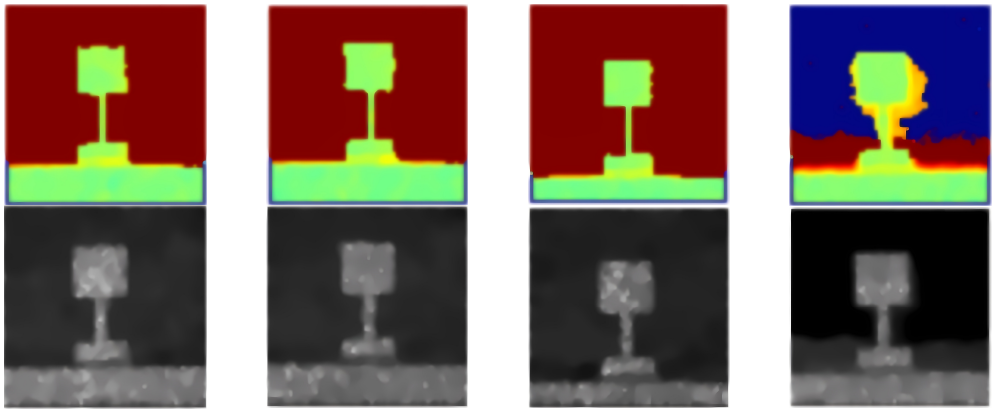


Figure 6. The reconstruction results of the SUARF algorithm on target A when the external light intensity is 0.1/2374/5450/44355 Lux.

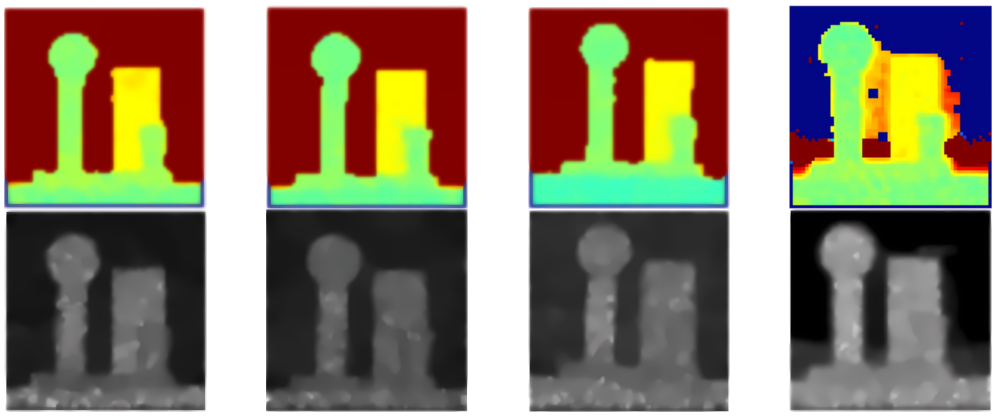


Figure 7. The reconstruction results of the SUARF algorithm on target B when the external light intensity is 0.1/1102/16172/44364 Lux.

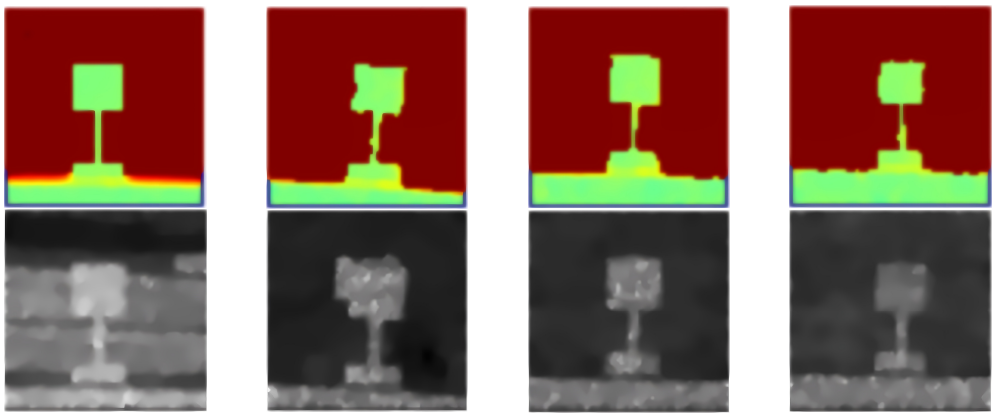


Figure 8. The reconstruction results of the SUARF algorithm for target A when the imaging distance is 10/25/40/55 m.

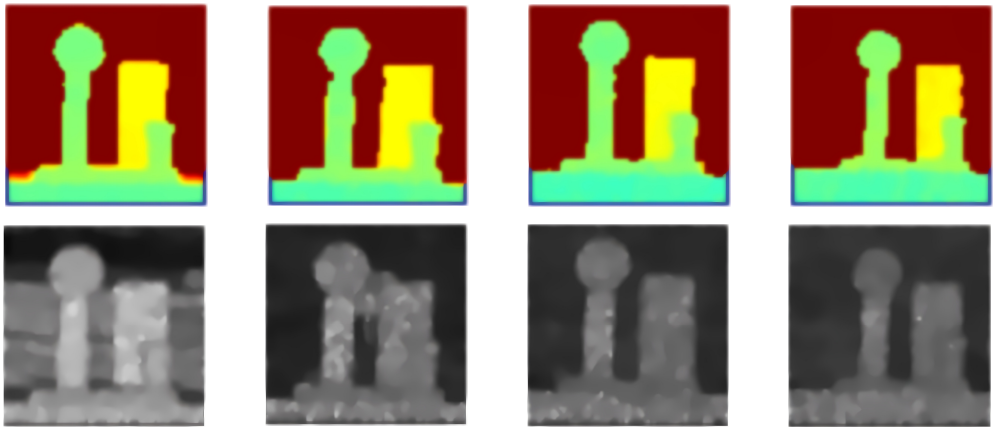


Figure 9. The reconstruction results of the SUARF algorithm for target B when the imaging distance is 10/25/40/55 m.

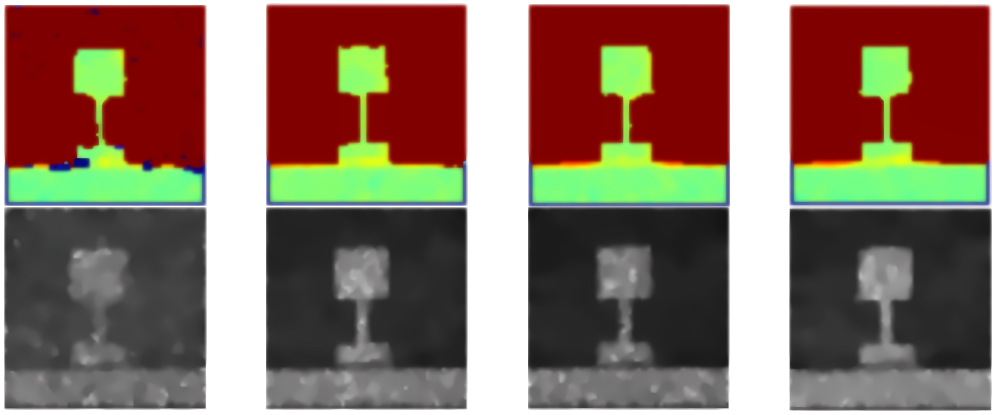


Figure 10. The reconstruction results of the SUARF algorithm for target A when the number of single pixel pulses is 5/50/100/500.

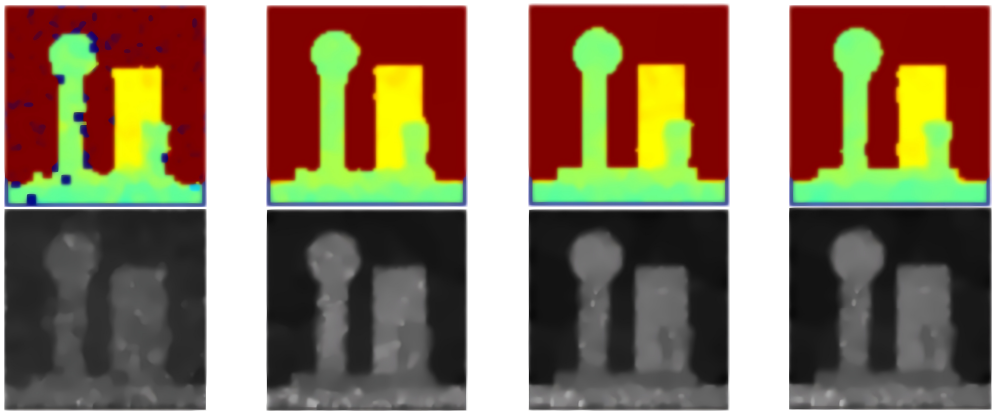


Figure 11. The reconstruction results of the SUARF algorithm for target B when the number of single pixel pulses is 5/50/100/500.

In Figures 6 and 7, the system is tested at a distance of about 40 m, the number of pixels collected is 64 × 64, and the number of single-pixel pulses is 50. From the reconstruction results, it can be seen that the SUARF algorithm maintains a better imaging effect with the enhancement of the external light intensity interference. In Figures 8 and 9, the number of pixels collected by the system is 64 × 64, and the number of single-pixel pulses is 50. From the reconstruction results, it can be seen that the SUARF algorithm maintains a better imaging effect with the increase of the imaging distance (the number of attenuation lengths, AL). In Figures 10 and 11, the number of pixels acquired by the system is 64 × 64, and it can

be seen that the SUARF algorithm still maintains a good imaging effect under the condition of low pulse number. Therefore, the SUARF algorithm is robust and can be tried for underwater imaging. The mainstream imaging algorithms for scanning single-photon lidar systems are first-photon, first-photon group algorithm, peak algorithm, STMF algorithm, etc. The STMF algorithm has strong anti-noise and convenience compared with the first photon, the first photon group, and the peak algorithm, etc., and has obvious advantages in imaging ability under general imaging conditions. Therefore, this paper mainly analyzed the performance of the SUARF algorithm reported here by comparing the STMF algorithm.

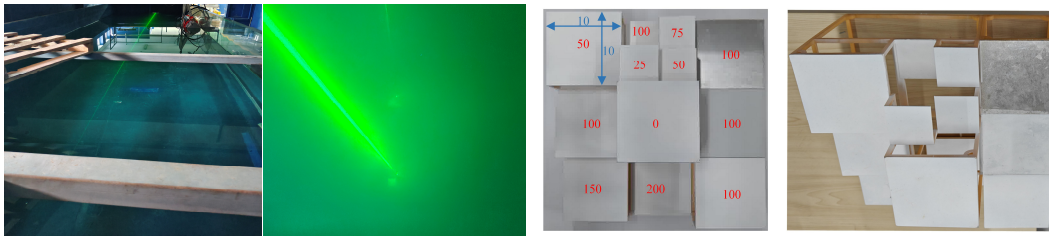
4. Underwater Test and Result Analysis

In underwater environments, conventional imaging techniques face many challenges due to the strong absorption and scattering of light, which make it extremely difficult to acquire high-contrast and high-resolution images. With its extremely high detection sensitivity, single-photon lidar technology is able to capture weak echo signals in this low-light and high-noise environment, thus realizing precise detection and imaging of underwater targets. Therefore, the practical application of single-photon lidar underwater can not only significantly improve the quality and depth resolution of underwater imaging, but also extend its application potential in many fields such as marine exploration, underwater archaeology, environmental monitoring, and underwater navigation, which is of great significance for both scientific research and industrial applications. In order to verify the environmental adaptability of the SUARF algorithm, we built four underwater environments for scene reconstruction experiments, namely, a swimming pool, pipeline (a drain placed horizontally), and the actual waters off the coast of China. According to the performance of the system, the target deployment distance from the lidar is about 5~6 AL.

4.1. Comparative Analysis of the Reconstruction Effect of Each Algorithm

4.1.1. Pool Reconstruction Tests

The underwater experiments were first carried out in a pool in the laboratory. The available dimensions of the pool are about 12 × 13 × 6 m^3, and the lidar and the target were placed at the two ends of the pool, respectively (Figure 12a). Due to the extremely poor visibility, the underwater visible light camera cannot shoot the target at all at the location of the lidar, such as Figure 12b.



(**a**) (**b**) (**c**) (**d**)

Figure 12. Schematic diagram of pool experiment and test target.

The target used for the pool was a multiplanar target (Figure 12c,d), with the highest square plane as the datum plane. The target information of the multi-depth target contains 12 planes in total at depths of 0/25/50/75/100 (three types of reflectivity)/150/200 cm. The pool style is shown in (a) and the water is rather turbid, as shown in (b), and the underwater camera can't capture the target and LIDAR even at very close range. The results of the algorithms are shown in Figures 13 and 14.

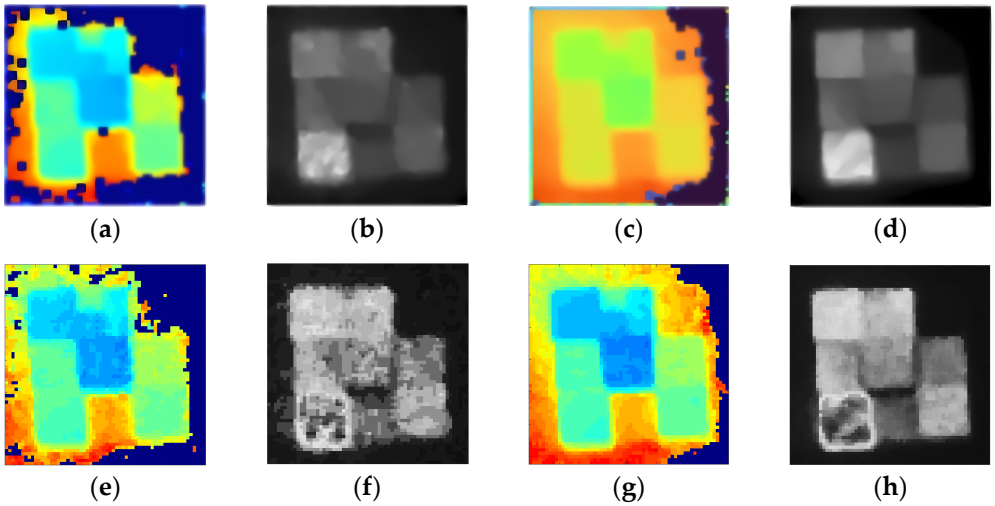


Figure 13. Results of pool target reconstruction for SUARF and STMF algorithms.

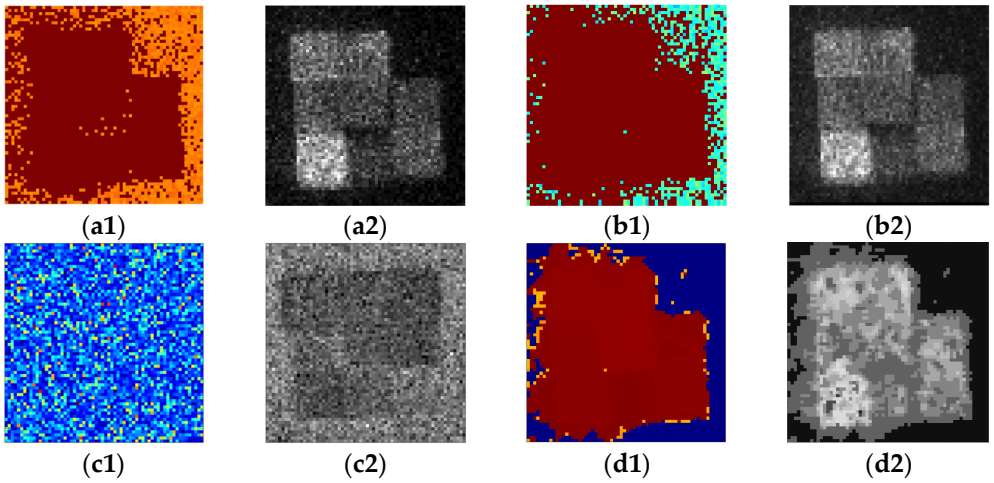


Figure 14. Results of pool target reconstruction for peak (**a1**,**a2**), cross-correlation (**b1**,**b2**), first photon (**c1**,**c2**), and first photon group (**d1**,**d2**) algorithms.

Figure 13a,b shows the results of the SUARF algorithm, and Figure 13e,f the results of the STMF algorithm. Figure 13a,b,e,f shows the imaging results at a pulse number of 50, Figure 13c,d,g,h the imaging results at a pulse number of 500. Figure 14a–d depicts the depth map (1) and reflection intensity map (2) obtained at the pulse number of 50 by using the peak, inter-correlation, first-photon, and first-photon group algorithms, respectively. It can be seen that the SUARF algorithm has the ability to repair anomalies on the basis of strong information extraction capability, and the depth resolution reaches 2.5 cm (the limit of the depth resolution of the device is 3 cm).

4.1.2. Quantitative Evaluation Criteria

Structural similarity index (SSIM) and root mean square error (RMSE) as well as subjective judgment are used to comprehensively evaluate the effectiveness, and the comparative reference is the adopted image at 500 pulses in Figure 15. Visible light struggles to capture clear high-resolution images in the complex experimental water environment, while longer observation times with time-counting single-photon LiDAR generally improve reconstruction quality, as confirmed in the paper's Section 3 air experiment. We chose 500 pulses as the evaluation index, because beyond this point additional pulses in our underwater system did not significantly enhance imaging but increased processing time. In practice, balancing the number of emitted pulses with observation effects is crucial to optimize

imaging performance and efficiency. The reconstruction results for peak, inter-correlation, first-photon, first-photon group, STMF, and SUARF algorithms are shown in Tables 1 and 2.

Table 1. The depth reconstruction results of evaluation index values of each algorithm.

Index	Peak	Cross-Correlation	First Photon	First Photon Group	STMF	SUARF
SSIM	0.373	0.376	0.293	0.512	0.723	0.828
RMSE	134.451	127.358	100.836	111.149	68.431	29.500

Table 2. The reflection intensity reconstruction results of evaluation index values of each algorithm.

Index	Peak	Cross-Correlation	First Photon	First Photon Group	STMF	SUARF
SSIM	0.590	0.630	0.408	0.568	0.563	0.833
RMSE	31.015	28.51	80.833	55.400	58.354	17.125

SSIM is an evaluation index of the similarity of two images, proposed by the Laboratory for Image and Video Engineering at the University of Texas at Austin [59]. It takes into account the brightness, contrast, and structural information of the images. Unlike traditional error metrics (e.g., MSE), SSIM is more in line with the perceptual properties of the human visual system. SSIM has a value between 0 and 1, where 1 means that the two images are exactly the same. The computation of SSIM involves localized windows of an image and assesses the similarity between the images by comparing the statistical properties within these windows. It not only takes into account the pixel-level differences but also the overall structural information of the image and is therefore widely used in image quality evaluation. RMSE [60] is a measure of the difference between predicted and actual values. It is the square root of the mean square error (MSE), which calculates the average of the squares of the differences between the predicted and actual values. It provides a visual measure of the magnitude of the error, with smaller values indicating that the model's predictions are closer to the actual values. RMSE is very sensitive to outliers, and therefore its value increases significantly when there is a large bias in the data, reflecting model variability in these points of variability. These two metrics provide a quantitative assessment of data similarity and prediction accuracy from different perspectives and are very important tools in the field of image processing. SSIM focuses on visual quality and can detect visual distortion and structural changes of images, while RMSE, on the other hand, focuses on the size of the error and can quantify the difference between the predicted value and the actual value. The combination of SSIM and RMSE can comprehensively reflect the visual similarity of images and the size of prediction error, help better understand the overall performance of the imaging effect, and provide more accurate and reliable evaluation results for the optimization and improvement of image processing algorithms. In Tables 1 and 2, the SUARF algorithm reported in this paper has the largest SSIM value and the smallest RMSE value. Compared with the STMF imaging algorithm, with the best performance at present, the imaging results of the SUARF algorithm proposed in this paper increase the SSIM and RMSE of the depth and reflection intensity map by 15%, 57% and 48%, 71%, respectively. It can discriminate 2.5cm@5-6AL in the imaging results, and it can discriminate targets with different reflectance.

Based on the subjective judgment results of 20 randomly selected groups (five people in each group), without observing the target in advance, more than 95% of the personnel think that the reconstruction results of the SUARF algorithm are the most intuitive, and there are very few anomalies. Since both depth and reflection intensity reconstruction are crucial for understanding the target, and RMSE and SSIM are two indicators for comprehensively

evaluating the reconstruction effect, we defined a comprehensive indicator to evaluate the reconstruction results of the algorithm, which is shown in Equation (21):

$$R_{\mathrm{T}} = \omega_1 \times \frac{1}{RMSE_{\mathrm{D}}} + \omega_2 \times \frac{1}{SSIM_{\mathrm{D}}} + \omega_3 \times \frac{1}{RMSE_{\mathrm{R}}} + \omega_4 \times \frac{1}{RMSE_{\mathrm{R}}} \tag{21}$$

In Equation (21), $RMSE_{\mathrm{D}}$, $SSIM_{\mathrm{D}}$, $RMSE_{\mathrm{R}}$, $SSIM_{\mathrm{D}}$ are the evaluation index values of the depth map and reflection intensity map, respectively, and ω_1, ω_2, ω_3, ω_4 are the coefficients of the index values, which can be calculated as Equation (22):

$$\begin{cases} \omega_1 = \frac{1/RMSE_{\mathrm{D}}}{1/RMSE_{\mathrm{D}}+SSIM_{\mathrm{D}}+1/RMSE_{\mathrm{R}}+SSIM_{\mathrm{R}}} \\ \omega_2 = \frac{SSIM_{\mathrm{D}}}{1/RMSE_{\mathrm{D}}+SSIM_{\mathrm{D}}+1/RMSE_{\mathrm{R}}+SSIM_{\mathrm{R}}} \\ \omega_3 = \frac{1/RMSE_{\mathrm{R}}}{1/RMSE_{\mathrm{D}}+SSIM_{\mathrm{D}}+1/RMSE_{\mathrm{R}}+SSIM_{\mathrm{R}}} \\ \omega_4 = \frac{SSIM_{\mathrm{R}}}{1/RMSE_{\mathrm{D}}+SSIM_{\mathrm{D}}+1/RMSE_{\mathrm{R}}+SSIM_{\mathrm{R}}} \end{cases} \tag{22}$$

The value interval of R_{T} is [0, 1], and the closer the value is to 1, the better the comprehensive reconstruction effect of the algorithm. Through the R_{T} evaluation index, we can conduct a comprehensive evaluation of the RMSE and SSIM evaluation values of the depth and reflection intensity of the algorithm. The value of R_{T} for each of the above algorithms is calculated in Table 3.

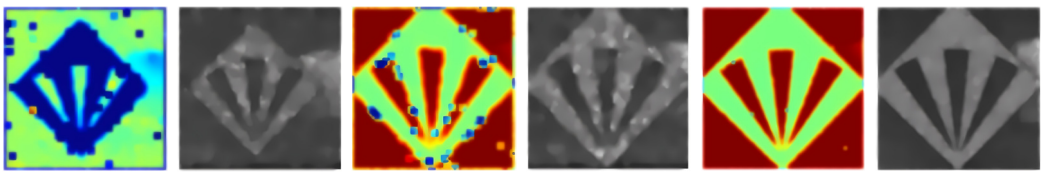
Table 3. Evaluation metrics for each algorithm R_{T} values.

Index	Peak	Cross-Correlation	First Photon	First Photon Group	STMF	SUARF
R_{T}	0.487	0.514	0.349	0.529	0.638	0.789

As can be seen from Table 3, R_{T} is increased by 24%. This index comprehensively reflects that the depth and reflectivity imaging results of the SUARF algorithm have been greatly improved compared with the latest algorithms. Combining the results of SSIM, RMSE, R_{T}, and the subjective judgment of the randomly selected persons, the SUARF algorithm reported in this paper is more effective in reconstructing the target scene.

In addition, in order to further verify the lateral resolution effect of the SUARF algorithm, a hollow target plate (Figure 15) was used, with a lateral resolution of 5 mm at the smallest point. The reconstruction results of the algorithm are shown in Figure 16.

Figure 15. Hollow target board and underwater visible light camera close shot.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 16. Imaging results of the SUARF target reconstruction algorithm for the hollow target plate. (**a**) Depth map, 32 pixels, 10 pulses. (**b**) Reflection intensity map, 32 pixels, 10 pulses. (**c**) Depth map, 64 pixels, 10 pulses. (**d**) Reflection intensity map, 64 pixels, 10 pulses. (**e**) Depth map, 128 pixels, 20 pulses. (**f**) Reflection intensity map, 128 pixels, 20 pulses.

As can be seen from the reconstruction results in Figure 16, the 32-pixels-10-pulses imaging results can already show the overall shape of the target plate, and the zone boundary can be distinguished. The target details for 128-pixels-20-pulses imaging results are fully revealed with a minimum resolution of about 5 mm.

4.2. Imaging Test of SUARF Algorithm in Different Environmental Waters

The water environment is complex and changeable. In order to test whether the SUARF algorithm can be adapted to different waters for imaging, we also conducted tests in the pipeline, swimming pool, and actual water off the coast of China. The target placement distance from the lidar is about 5–6 AL. Affected by the space size of the experimental environment and the target placement sinking and floating, the target used in different environments is different. The test layout is as Figure 17 and the test target is shown in Figures 18–20.

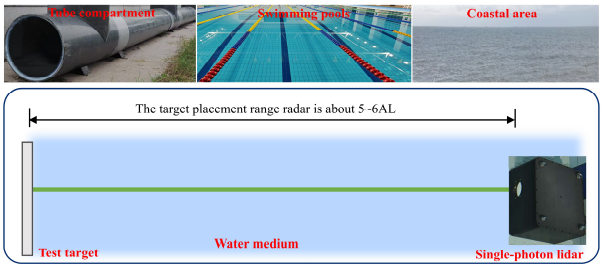
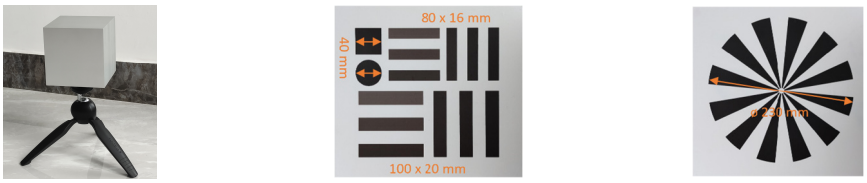
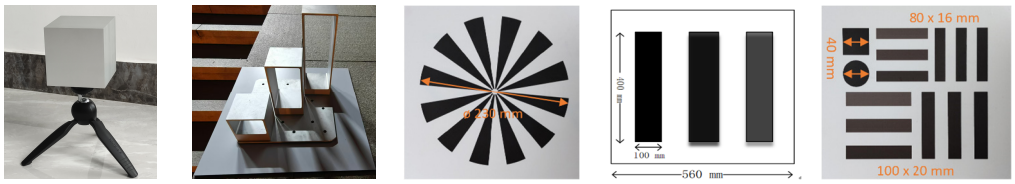


Figure 17. Experimental layout of SUARF target reconstruction algorithm in different waters.



(**a**) (**b**) (**c**)

Figure 18. Targets used in pipeline test. (**a**) Metal cube. (**b**) Band target A. (**c**) Spoke-type target.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 19. Targets used in swimming pool test. (**a**) Metal cube. (**b**) Diagonal ladder distribution of metal blocks. (**c**) Spoke-type target. (**d**) Band target B. (**e**) Band target A.

(**a**) (**b**)

Figure 20. Targets used in the actual water off the coast of China test. (**a**) Metal cube. (**b**) Spoke-type target.

In different water environments, we selected different target combinations to comprehensively evaluate the performance of the SUARF algorithm. In the pipeline, we used a metal cube, a band target A, and a spoke-type target to verify basic imaging capabilities. In

the swimming pool, a diagonal ladder distribution of metal blocks and a band target B were added to test the imaging effect of the algorithm on complex structures and different material targets. In the actual water area in winter, faced with the problems of water complexity, biodiversity, dynamic changes in the environment, artificial obstacles, and insufficient performance of hardware equipment, we chose a metal cube and a spoke-type target to test the performance; this can reduce the complexity of the test and verify the imaging resolution capability. In Figure 18, the metal cube has a volume of $10 \times 10 \times 10$ cm. Band target A contains 10×2 cm and 8×1.6 cm bands with a 20% increase in color contrast from the top to the bottom, as well as squares and circles with a side length and diameter of 4 cm. The spoke-type target was 2.3 cm in diameter and contained 12 black spokes and 12 white spokes. Band target B and Diagonal ladder distribution of metal blocks were used as substitutes in the swimming pool, which contains three 40×10 cm rectangular bands with a contrast difference of 20–30% from left to right, and the stepped metal block contains three height difference of 10 cm rectangles (the shapes above are all 10×10 cm squares), respectively. The situation that needs further clarification is that, affected by the waves generated by underwater activities, the target will produce rapid jitter, which is not conducive to the deployment and observation of targets used in relatively controlled underwater environments such as laboratory pools and swimming pools in actual waters. Therefore, only the more challenging spokes target was used as a representative to test in the actual water. In addition, our goal was to test the adaptability and robustness of the SUARF in several underwater environments, rather than just pursuing the limits of the indexes. The limits of the indexes are closely related to both the hardware system and the algorithm. In the actual water test, we paid more attention to the performance of the algorithm in different environmental conditions, and how to optimize the algorithm to adapt to various complex situations. Therefore, the main purpose of the underwater single-photon lidar test machine built in this report is to cooperate with the algorithm test and verify the performance and effect of the algorithm in the real underwater environment. Therefore, in this case, the use of low-resolution targets is sufficient to meet the needs of testing the adaptability of the algorithm. Although high-precision targets can provide higher accuracy, such high accuracy is not necessary in real-water testing and will increase the complexity and cost of system testing. The reconstruction results in the pipeline are shown below.

In Figure 21, (a) is the depth map of the metal cube, (b) is the reflectance intensity map of the metal cube, (c) is the depth map of the band target A, (d) is the reflectance intensity map of the band target A, (e) is the depth map of the spoke-type target, and (f) is the reflectance intensity map of the spoke-type target. The first line (a1–f1)shows the imaging results using the first photon group algorithm, the second line (a2–f2) showsthe imaging results using the STMF imaging algorithm, and the third line (a3–f3) showsthe imaging results using the SUARF imaging algorithm proposed in this paper. According to the imaging results, for the simple geometry metal cube (a) and (b), the three algorithms can carry out relatively high-quality imaging. However, when the first photon group algorithm is used for imaging under different target needs, different environments, and different single-photon LiDAR systems, the imaging photon threshold needs to be adjusted from time to time. When the threshold is not chosen properly, imaging is not effective, and the poor imaging results are displayed in Figure 22.

In Figure 22, (a4) is the depth map of the metal cube, (b4) is the reflectance intensity map of the metal cube, (c4) is the depth map of the band target A, (d4) is the reflectance intensity map of the band target A, (e4) is the depth map of the spoke-type target, and (f4) is the reflectance intensity map of the spoke-type target. In addition, the first photon group algorithm lacks more information on the imaging edge of simple geometry; there

are more abnormal pixel values, and the imaging effect is not good for the objects (c4), (d4), (e4), and (f4) with higher horizontal resolution. Compared with the first photon group algorithm, the STMF algorithm has a stronger information ex-traction ability, but the boundary retention ability and pixel outlier removal ability need to be further enhanced. The SUARF algorithm proposed in this paper has stronger detail display and outlier processing capabilities. In order to further confirm the imaging ad-vantages of the SUARF algorithm, we conducted comparative tests on the imaging results of the swimming pool, simultaneously, and the results are as Figures 23–25.

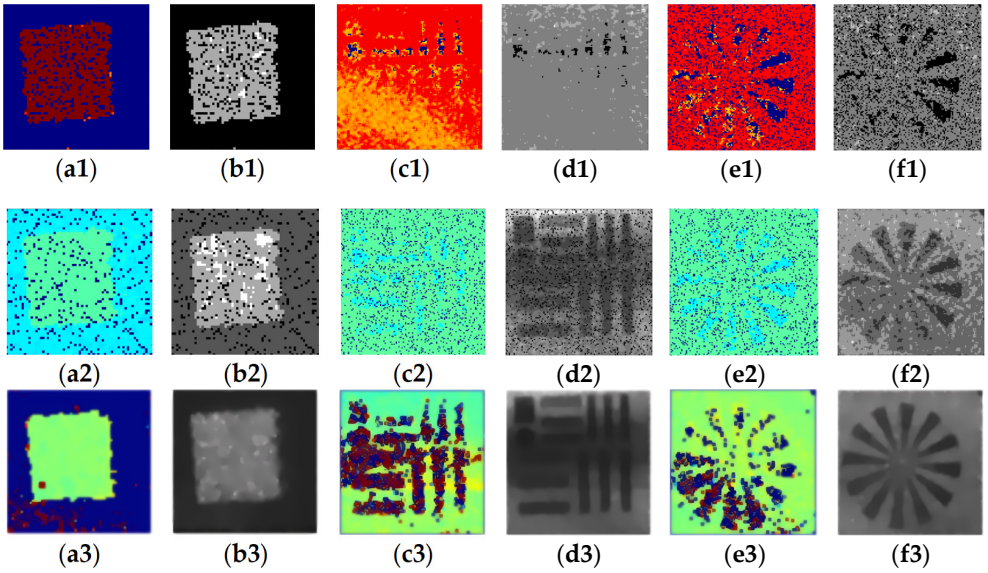


Figure 21. Results of the target reconstruction in the pipeline.

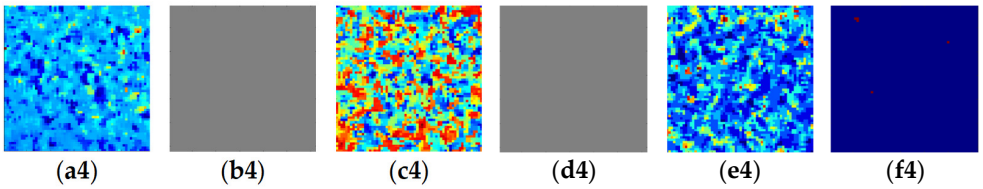


Figure 22. Imaging results under the inappropriate threshold of the first photon group algorithm.

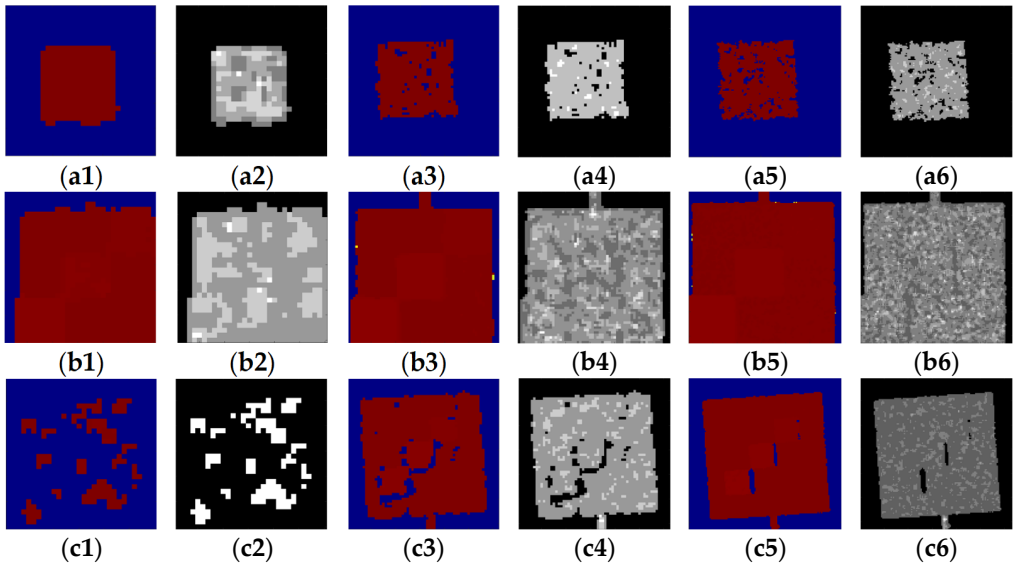


Figure 23. *Cont.*

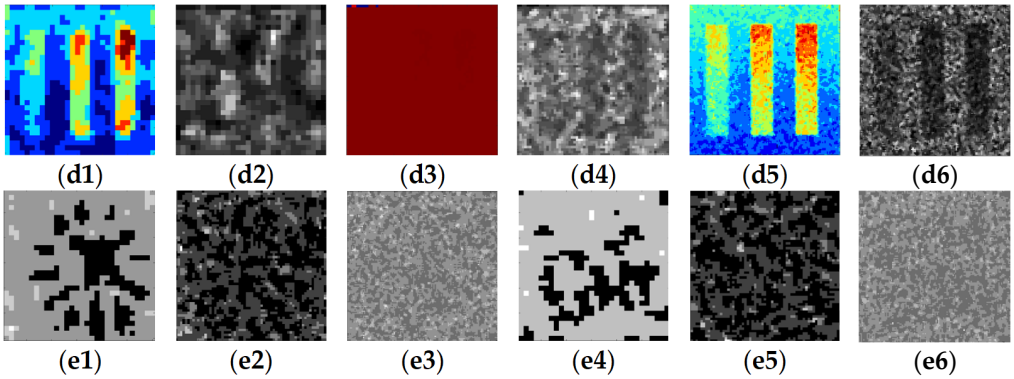


Figure 23. Image result of first photon group algorithm in the swimming pool.

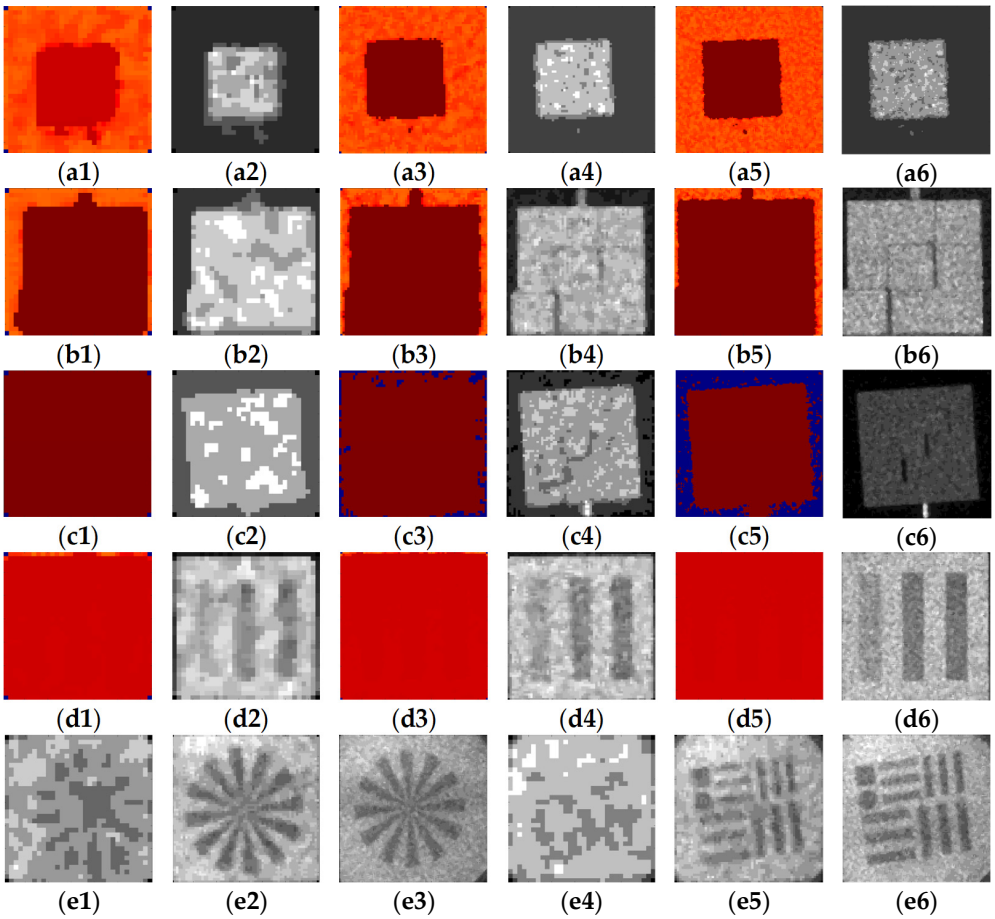


Figure 24. Image result of STMF algorithm in the swimming pool.

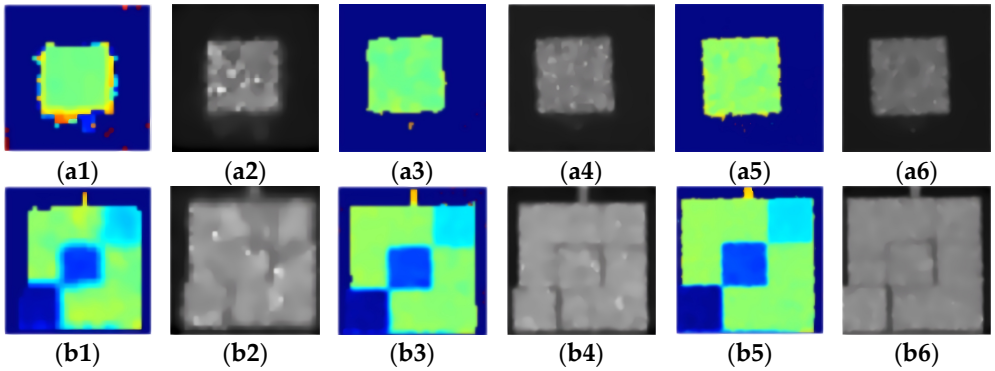


Figure 25. *Cont.*

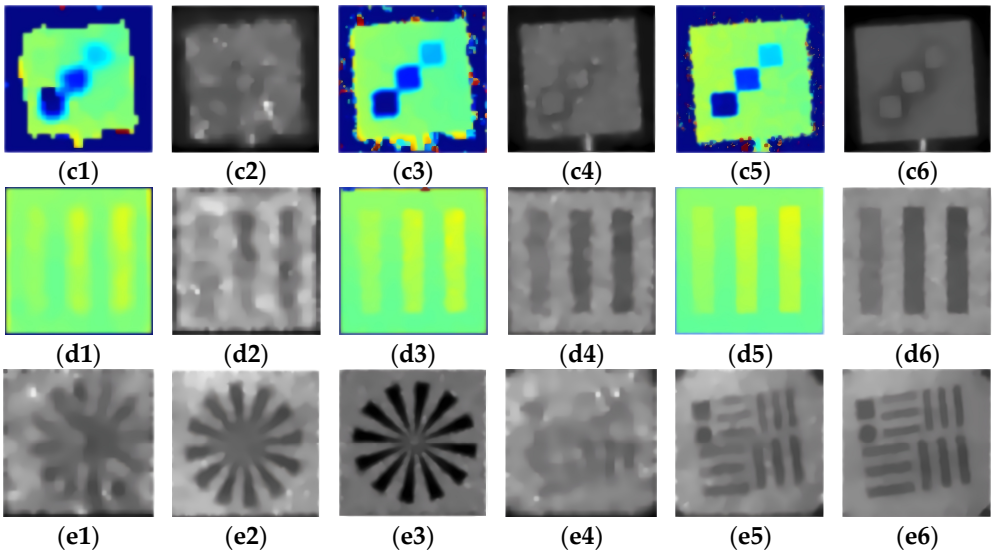


Figure 25. Image result of SUARF algorithm in the swimming pool.

Figures 23–25, (a) depict the depth maps and reflective intensity maps for the metal cube, (b) and (c) the depth maps and reflective intensity maps for the diagonal ladder distribution of metal blocks with different sizes for the base plate, (d) the depth maps and reflective intensity maps for the band target B, and (e) the depth maps and reflective intensity maps for the spoke-type target and band target B. In the above results, 1 and 2 are 32-pixel imaging result maps, 3 and 4 are 64-pixel imaging result maps, and 5 and 6 are 128-pixel imaging result maps. From the results of (a) metal cube, (b) diagonal ladder distribution of metal blocks, and (c) diagonal ladder distribution of metal blocks, the SUARF algorithm is superior to the first photon composition image algorithm and the STMF imaging algorithm in resolving deep targets. As can be seen from the results of (d) band target B and (e) spoke-type target and band target A, the SUARF algorithm still maintains a strong information extraction ability for different targets with the same depth. Moreover, the ability of the SUARF algorithm to smooth pixels is much better than the current first photon group image algorithm and the STMF imaging algorithm. Because the underwater test is difficult, it is not easy to keep the target stationary for a long time and observe the target with a large pulse number to obtain reference results; the above comparison results are easy to obtain intuitively, so we did not use indicators for evaluation, but direct observation. In order to further confirm the imaging advantages of the SUARF algorithm in the actual water environment, we conducted a comparative test on the imaging results of the actual water area, simultaneously, and the results are as follows.

In Figures 26–28, (a) and (b) depict the depth map and reflection intensity map for the metal cube, (c) and (d) the depth map and reflection intensity map for the spoke-type target, and (e) the depth map for the metal cube under a single detection pulse. From the reconstruction results in different water environments of Figures 21–28, it can be seen that in the challenging winter waters off the coast of China, the first-photon group image algorithm loses its stable imaging ability under laboratory conditions, and the STMF imaging algorithm has stronger information extraction ability than the first-photon group imaging algorithm. However, more complex spoke-type targets cannot be effectively imaged, especially in a single pulse, and information cannot be effectively extracted. Therefore, based on the imaging results in the pipeline, swimming pool, and actual waters, the SUARF algorithm proposed in this paper has stronger information extraction and outlier removal capabilities than the latest first-photon group and STMF imaging algorithms. It is able to accomplish the task well and achieve the detection of

small targets in all of the above water environments in which a lower number of detection pulses were used for detection. In the pipeline, because the wall of the tube compartment is very close to the target, it produces a large emission interference, which also simulates the detection of targets in narrow water areas. In the actual waters, after obtaining the single pulse data, the SUARF algorithm can also achieve more accurate detection of simple targets. On the basis of the obtained data, we also carried out a simulation analysis of the environmental influencing factors. We found that with the decrease of the signal-to-noise ratio (SNR), the RMSE value of SUARF imaging results gradually increased and the SSIM value gradually decreased, but it did not affect the observation of the scene contour, which reflects the algorithm's good underwater target reconstruction capability.

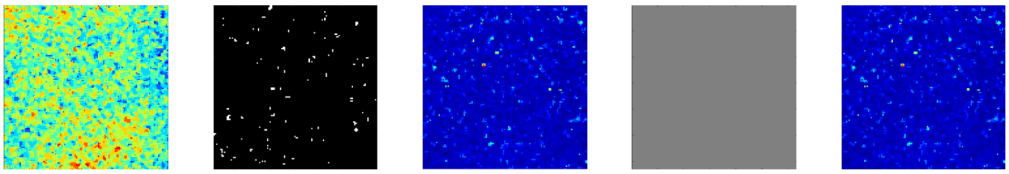


Figure 26. Imaging results of the first photon group algorithm in actual water off the coast of China.

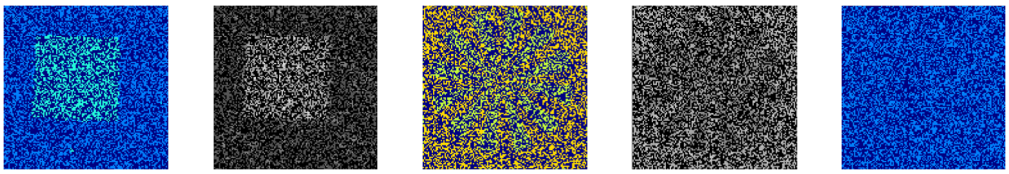


Figure 27. Imaging results of the STMF algorithm in actual water off the coast of China.

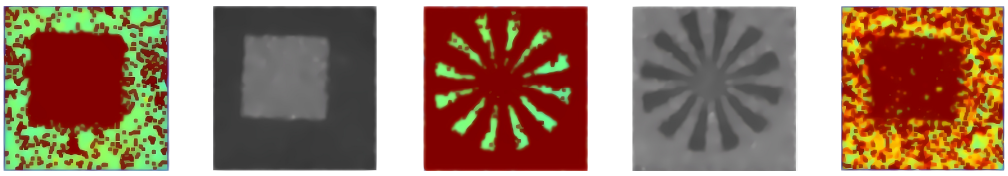


Figure 28. Imaging results of the SUARF algorithm in actual water off the coast of China.

5. Conclusions

This paper reports a sequential multimodal underwater single-photon lidar adaptive target reconstruction algorithm based on spatiotemporal sequence fusion, which consists of three steps: data preprocessing, sequence-optimized extreme value inference filtering, and collaborative variation strategy for image optimization. It has strong information extraction ability and noise filtering ability and can reconstruct target depth and reflection intensity information from complex echo photon time counts and spatial pixel relationships. The test results show that the SUARF algorithm utilizes the built single-photon lidar system to achieve underwater lateral resolution and distance resolution of 5 mm and 2.5cm@6AL, respectively, using a prototype constructed by ourselves, and it has stronger information extraction and compensation of anomalous pixel values compared to the currently available scanning single-photon lidar imaging algorithms. This effective target scene reconstruction algorithm endows underwater single-photon LiDAR with the ability to distinguish signal photons from noise photons in complex underwater environments and then realizes the imaging of small slow-moving underwater targets. It can depict the three-dimensional structure of historical relics such as shipwrecks and ruins in underwater archaeology, helping archaeologists explore the mysteries of ancient civilization. It can also be used in

the field of marine biological monitoring to capture marine organisms and habitats and provide intuitive data for biologists to study marine ecology. It can also detect the integrity and stability of key facilities such as submarine pipelines and bridge piers in underwater infrastructure testing to ensure the safe operation of marine engineering.

In future research, the system hardware facilities will be further analyzed to improve the hardware performance, the algorithm structure will be further optimized with the data results to provide a more optimized detection means for underwater target detection, and we will also design a complete environmental impact analysis experiment to evaluate the stability and adaptability of the algorithm under various environmental factors.

Author Contributions: Conceptualization, T.R.; methodology, T.R. and C.W.; software, T.R., C.W. and Q.L.; validation, T.R., Y.W. and Q.Z.; resources, C.W. and Z.Z.; data curation, T.R., Y.Z., J.L. and Y.W.; writing—original draft preparation, T.R.; writing—review and editing, T.R. and C.W.; project administration, C.W.; funding acquisition, C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Key R&D Program of Shandong Province, China (Major Scientific and Technological Innovation Project), grant number 2022ZLGX04.

Data Availability Statement: All raw and processed data are available after publication by contacting the author.

Acknowledgments: I would like to thank my parents (Jianfang Liu and Yikun Rong), without their encouragement, it would have been difficult for me to have had firm confidence in my research. I also thank my girlfriend (Xiuyue Dong) for encouraging me in my research process.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, B.; Liu, Z.; Men, S.; Li, Y.; Ding, Z.; He, J.; Zhao, Z. Underwater Hyperspectral Imaging Technology and Its Applications for Detecting and Mapping the Seafloor: A Review. *Sensors* **2020**, *20*, 4962. [CrossRef]
2. Song, Y.; Nakath, D.; She, M.; Köser, K. Optical Imaging and Image Restoration Techniques for Deep Ocean Mapping: A Comprehensive Survey. *PFG–J. Photogramm. Remote Sens. Geoinf. Sci.* **2022**, *90*, 243–267. [CrossRef]
3. Mills, M.S.; Ungermann, M.; Rigot, G.; den Haan, J.; Leon, J.X.; Schils, T. Assessment of the utility of underwater hyperspectral imaging for surveying and monitoring coral reef ecosystems. *Sci. Rep.* **2023**, *13*, 21103. [CrossRef] [PubMed]
4. Dumke, I.; Nornes, S.M.; Purser, A.; Marcon, Y.; Ludvigsen, M.; Ellefmo, S.L.; Johnsen, G.; Søreide, F. First hyperspectral imaging survey of the deep seafloor: High-resolution mapping of manganese nodules. *Remote Sens. Environ.* **2018**, *209*, 19–30. [CrossRef]
5. Gu, Z.; Liu, X.; Hu, Z.; Wang, G.; Zheng, B.; Watson, J.; Zheng, H. Underwater computational imaging: A survey. *Intell. Mar. Technol. Syst.* **2023**, *1*, 2. [CrossRef]
6. Chen, Y.; Guo, S.; He, Y.; Luo, Y.; Chen, W.; Hu, S.; Huang, Y.; Hou, C.; Su, S. Simulation and Design of an Underwater Lidar System Using Non-Coaxial Optics and Multiple Detection Channels. *Remote Sens.* **2023**, *15*, 3618. [CrossRef]
7. Daramola, O.; Nootz, G.; Britton, W. Fusion of AUV-Mounted 360-Degree Underwater LiDAR and Side Scan Sonar Data. In Proceedings of the OCEANS 2023—MTS/IEEE U.S. Gulf Coast, Biloxi, MS, USA, 25–28 September 2023; pp. 1–6. [CrossRef]
8. Shen, Y.; Zhao, C.; Liu, Y.; Wang, S.; Huang, F. Underwater Optical Imaging: Key Technologies and Applications Review. *IEEE Access* **2021**, *9*, 85500–85514. [CrossRef]
9. Churnside, J.H.; Shaw, J.A. Lidar remote sensing of the aquatic environment: Invited. *Appl. Opt.* **2020**, *59*, C92–C99. [CrossRef]
10. Hu, X. Underwater Image Enhancement Method Based on Wavelet Transform and Retinex. In Proceedings of the 2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Yibin, China, 20–22 August 2021; pp. 86–90. [CrossRef]
11. Junejo, N.U.R.; Sattar, M.; Adnan, S.; Sun, H.; Adam, A.B.M.; Hassan, A.; Esmaiel, H. A Survey on Physical Layer Techniques and Challenges in Underwater Communication Systems. *J. Mar. Sci. Eng.* **2023**, *11*, 885. [CrossRef]
12. Jekateryńczuk, G.; Piotrowski, Z. A Survey of Sound Source Localization and Detection Methods and Their Applications. *Sensors* **2024**, *24*, 68. [CrossRef]
13. Campo-Valera, M.; Diego-Tortosa, D.; Rodríguez-Rodríguez, I.; Useche-Ramírez, J.; Asorey-Cacheda, R. Signal Processing to Characterize and Evaluate Nonlinear Acoustic Signals Applied to Underwater Communications. *Electronics* **2024**, *13*, 4192. [CrossRef]

14. Lin, Z.; Shangguan, M.; Cao, F.; Yang, Z.; Qiu, Y.; Weng, Z. Underwater Single-Photon Lidar Equipped with High-Sampling-Rate Multi-Channel Data Acquisition System. *Remote Sens.* **2023**, *15*, 5216. [CrossRef]
15. Shangguan, M.; Yang, Z.; Lin, Z.; Lee, Z.; Xia, H.; Weng, Z. Compact Long-Range Single-Photon Underwater Lidar with High Spatial–Temporal Resolution. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1501905. [CrossRef]
16. Yang, X.; Tong, Z.; Dai, Y.; Chen, X.; Zhang, H.; Zou, H.; Xu, J. 100 m full-duplex underwater wireless optical communication based on blue and green lasers and high sensitivity detectors. *Opt. Commun.* **2021**, *498*, 127261. [CrossRef]
17. Gao, Q.; Wang, C.; Wang, X.; Liu, Z.; Liu, Y.; Wang, Q.; Niu, W. Pointing Error Correction for Vehicle-Mounted Single-Photon Ranging Theodolite Using a Piecewise Linear Regression Model. *Sensors* **2024**, *24*, 3192. [CrossRef] [PubMed]
18. Degnan, J.J. Evolution of Single Photon Lidar: From Satellite Laser Ranging to Airborne Experiments to ICESat-2. *Photonics* **2024**, *11*, 924. [CrossRef]
19. Rapp, J.; Tachella, J.; Altmann, Y.; McLaughlin, S.; Goyal, V.K. Advances in Single-Photon Lidar for Autonomous Vehicles: Working Principles, Challenges, and Recent Advances. *IEEE Signal Process. Mag.* **2020**, *37*, 62–71. [CrossRef]
20. Hong, Y.; Liu, S.; Li, Z.P.; Huang, X.; Jiang, P.; Xu, Y.; Wu, C.; Zhou, H.; Zhang, Y.C.; Ren, H.L.; et al. Airborne single-photon LiDAR towards a small-sized and low-power payload. *Optica* **2024**, *11*, 612–618. [CrossRef]
21. Zhang, X.; Wang, F.; Guo, Y.; Wang, J.; Luo, Y.; Wu, W.; Hou, J.; Jiang, Z.; Peng, Z.; Huang, G.; et al. Research on linear array scanning lidar and photon signal processing technology based on InGaAs single-photon detector. *Infrared Laser Eng.* **2023**, *52*, 20220474. [CrossRef]
22. Shi, Y.; Li, Z.; Feng, B.; Yan, P.; Du, B.; Zhou, H.; Pan, H.; Wu, G. Enhanced solar-blind ultraviolet single-photon detection with a Geiger-mode silicon avalanche photodiode. *Chin. Opt. Lett.* **2016**, *14*, 030401. [CrossRef]
23. Xie, J.; Zhang, Z.; Jia, F.; Li, J.; Huang, M.; Zhao, Y. Improved single-photon active imaging through ambient noise guided missing data filling. *Opt. Commun.* **2022**, *508*, 127747. [CrossRef]
24. Zhang, H.; Zhao, X.; Zhang, Y.; Zhang, L.; Sun, M. Review of Advances in Single-Photon LiDAR. *Chin. J. Lasers* **2022**, *49*, 1910003. [CrossRef]
25. Li, Z.P.; Ye, J.T.; Huang, X.; Jiang, P.Y.; Cao, Y.; Hong, Y.; Yu, C.; Zhang, J.; Zhang, Q.; Peng, C.Z.; et al. Single-photon imaging over 200 km. *Optica* **2021**, *8*, 344–349. [CrossRef]
26. Signorelli, F.; Telesca, F.; Conca, E.; Della Frera, A.; Ruggeri, A.; Giudice, A.; Tosi, A. Low-Noise InGaAs/InP Single-Photon Avalanche Diodes for Fiber-Based and Free-Space Applications. *IEEE J. Sel. Top. Quantum Electron.* **2022**, *28*, 1–10. [CrossRef]
27. Tachella, J.; Altmann, Y.; Mellado, N.; McCarthy, A.; Tobin, R.; Buller, G.S.; Tourneret, J.Y.; McLaughlin, S. Real-time 3D reconstruction from single-photon lidar data using plug-and-play point cloud denoisers. *Nat. Commun.* **2019**, *10*, 4984. [CrossRef]
28. Caramazza, P.; Boccolini, A.; Buschek, D.; Hullin, M.; Higham, C.F.; Henderson, R.; Murray-Smith, R.; Faccio, D. Neural network identification of people hidden from view with a single-pixel, single-photon detector. *Sci. Rep.* **2018**, *8*, 11945. [CrossRef]
29. Ma, L.; Sun, J.; Yang, X.; Lu, J.; Lu, W.; Zhou, X.; Ni, H. Reconstruction method of 128 $\times$ 256 array single photon Lidar based on multi-domain stability feature fusion. *Opt. Laser Technol.* **2025**, *181 Pt C*, 111970. [CrossRef]
30. Halimi, A.; Tobin, R.; McCarthy, A.; McLaughlin, S.; Buller, G.S. Restoration of Multilayered Single-Photon 3D Lidar Images. In Proceedings of the 2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM), Sheffield, UK, 8–11 July 2018; pp. 642–646. [CrossRef]
31. Bian, L.; Song, H.; Peng, L.; Chang, X.; Yang, X.; Horstmeyer, R.; Ye, L.; Zhu, C.; Qin, T.; Zheng, D.; et al. High-resolution single-photon imaging with physics-informed deep learning. *Nat. Commun.* **2023**, *14*, 5902. [CrossRef] [PubMed]
32. Zhao, X.; Wu, M.; Zhang, Y.; Wang, C.; Chen, R.; He, W.; Chen, Q. Robust single-photon 3D imaging based on full-scale feature integration and intensity edge guidance. *Opt. Lasers Eng.* **2024**, *172*, 107850. [CrossRef]
33. Cai, R.; Guo, H.; Li, X.; Che, J.; He, J.; Liu, H.; Yang, L.; Guo, Y.; Pu, M.; Luo, X. Sub-diffraction-limited single-photon 3D imaging based on domain features extraction network at kilometer-scale distance. *Opt. Laser Technol.* **2025**, *181 Pt B*, 111868. [CrossRef]
34. Xia, F.; Kim, K.; Eliezer, Y.; Shaughnessy, L.; Gigan, S.; Cao, H. Deep Learning with Passive Optical Nonlinear Mapping. *arXiv* **2024**, arXiv:2307.08558. [CrossRef]
35. Bosse, S.; Maniry, D.; Müller, K.R.; Wiegand, T.; Samek, W. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. *IEEE Trans. Image Process.* **2018**, *27*, 206–219. [CrossRef] [PubMed]
36. Wang, Z.; Chen, J.; Hoi, S.C. Deep Learning for Image Super-Resolution: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3365–3387. [CrossRef] [PubMed]
37. Xu, Y.; Tong, Z.; Jiang, P.; Xu, L.; Wu, L.; Hu, J.; Yang, C.; Zhang, W.; Zhang, Y.; Zhang, J.; et al. Deep-learning based photon-efficient 3D and reflectivity imaging with a 64 $\times$ 64 single-photon avalanche detector array. *Opt. Express* **2022**, *30*, 32948–32964. [CrossRef]
38. Shao, W.; Rowe, S.P.; Du, Y. Artificial intelligence in single photon emission computed tomography (SPECT) imaging: A narrative review. *Ann. Transl. Med.* **2021**, *9*, 820. [CrossRef]
39. Yao, D.; Yang, C.; Li, H. PCA-based real-time single-photon 3D imaging method. *Opt. Commun.* **2022**, *508*, 127777. [CrossRef]

40. Zhu, G.; Nan, Z.; Zhang, X.; Chu, K.; Zhan, S.; Liu, X.; Lin, X. High anti-interference 3D imaging LIDAR system based on digital chaotic pulse position modulation. *Opt. Laser Technol.* **2023**, *163*, 109405. [CrossRef]
41. Xu, R.; Ye, H.; Lv, H.; Sun, J.; Shi, Y.; Liu, W.; Zhao, T.; Song, Y. Single-photon lidar depth imaging based on sparsity adaptive matching pursuit method. *Opt. Lasers Eng.* **2024**, *180*, 108314. [CrossRef]
42. Kirmani, A.; Venkatraman, D.; Shin, D.; Colaço, A.; Wong, F.N.C.; Shapiro, J.H.; Goyal, V.K. First-Photon Imaging. *Science* **2013**, *343*, 58–61. [CrossRef]
43. Peng, X.; Zhao, X.Y.; Li, L.J.; Sun, M.J. First-photon imaging via a hybrid penalty. *Photonics Res.* **2020**, *8*, 325–330. [CrossRef]
44. Hua, K.; Liu, B.; Wang, H.; Fang, L.; Chen, Z.; Luo, J. Efficient and noise robust photon-counting imaging method. In Proceedings of the Seventh Symposium on Novel Photoelectronic Detection Technology and Applications, Kunming, China, 5–7 November 2020; 117632H. SPIE: Bellingham, WC, USA, 2021; Volume 11763. [CrossRef]
45. Hua, K.; Liu, B.; Chen, Z.; Fang, L.; Wang, H. Efficient and Noise Robust Photon-Counting Imaging with First Signal Photon Unit Method. *Photonics* **2021**, *8*, 229. [CrossRef]
46. Goudreault, F.; Scheuble, D.; Bijelic, M.; Robidoux, N.; Heide, F. LiDAR-in-the-Loop Hyperparameter Optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 13404–13414. [CrossRef]
47. Liu, X.; Shi, J.; Chen, H.; Zeng, G. First-Photon Ghost Imaging at Low Light Level. In Proceedings of the Conference on Lasers and Electro-Optics, OSA Technical Digest (online) (Optica Publishing Group 2017), San Jose, CA, USA, 4–19 May 2017; pp. 1–2. Available online: https://ieeexplore.ieee.org/document/8082923 (accessed on 26 November 2024).
48. Xu, Y.; Jiang, X.; Jiang, P.; Xu, L.; Wu, L.; Hu, J.; Zhang, Y.; Zhang, J.; Zou, B. S2O-FSPI: Fourier single pixel imaging via sampling strategy optimization. *Opt. Laser Technol.* **2023**, *166*, 109651. [CrossRef]
49. Rong, T.; Wang, Y.; Zhu, Q.; Wang, C.; Zhang, Y.; Li, J.; Zhou, Z.; Luo, Q. Sequential Two-Mode Fusion Underwater Single-Photon Lidar Imaging Algorithm. *J. Mar. Sci. Eng.* **2024**, *12*, 1595. [CrossRef]
50. Draz, H.H.; Elashker, N.E.; Mahmoud, M.M.A. Optimized Algorithms and Hardware Implementation of Median Filter for Image Processing. *Circuits Syst. Signal Process.* **2023**, *42*, 5545–5558. [CrossRef]
51. Lone, M.R.; Sandhu, A.K. Enhancing image quality: A nearest neighbor median filter approach for impulse noise reduction. *Multimed. Tools Appl.* **2024**, *83*, 56865–56881. [CrossRef]
52. Shin, D.; Kirmani, A.; Goyal, V.K.; Shapiro, J.H. Photon-Efficient Computational 3-D and Reflectivity Imaging with Single-Photon Detectors. *IEEE Trans. Comput. Imaging* **2015**, *1*, 112–125. [CrossRef]
53. Rapp, J.; Goyal, V. A Few Photons Among Many: Unmixing Signal and Noise for Photon-Efficient Active Imaging. *IEEE Trans. Comput. Imaging* **2017**, *3*, 445–459. [CrossRef]
54. Maccarone, A.; Drummond, K.; McCarthy, A.; Steinlehner, U.K.; Tachella, J.; Garcia, D.A.; Pawlikowska, A.; Lamb, R.A.; Henderson, R.K.; McLaughlin, S.; et al. Submerged single-photon LiDAR imaging sensor used for real-time 3D scene reconstruction in scattering underwater environments. *Opt. Express* **2023**, *31*, 16690–16708. [CrossRef] [PubMed]
55. Noether, F. Über eine Klasse singulärer Integralgleichungen. *Math. Ann.* **1920**, *82*, 42–63. [CrossRef]
56. Döttling, M.; Mohr, W.; Osseiran, A. WINNER II Channel Models. In *Radio Technologies and Concepts for IMT-Advanced*; Wiley: Hoboken, NJ, USA, 2010; pp. 39–92. [CrossRef]
57. Chan, S.H.; Khoshabeh, R.; Gibson, K.B.; Gill, P.E.; Nguyen, T.Q. An augmented Lagrangian method for total variation video restoration. *IEEE Trans. Image Process.* **2011**, *20*, 3097–3111. [CrossRef] [PubMed]
58. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 1905–1914. [CrossRef]
59. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
60. Estévez, J.; Bellido-Jiménez, J.A.; Liu, X.; García-Marín, A.P. Monthly Precipitation Forecasts Using Wavelet Neural Networks Models in a Semiarid Environment. *Water* **2020**, *12*, 1909. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

Modeling and Data Analysis of Bistatic Bottom Reverberation from a Towed Horizontal Array

Zhaohua Su [1,2], Jie Zhuo [1,2,*] and Chao Sun [1,2]

[1] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China; suzh@mail.nwpu.edu.cn (Z.S.); csun@nwpu.edu.cn (C.S.)
[2] Shaanxi Key Laboratory of Underwater Information Technology, Xi'an 710072, China
* Correspondence: jzhuo@nwpu.edu.cn

Abstract: The spatial-temporal structures of bottom reverberation are associated with seafloor features. In a bistatic bottom reverberation experiment involving a vertical transmitting array and a towed horizontal receiving array, stable stripe structures were observed within the beam-time domain. In this study, a bistatic reverberation model based on ray theory is presented to interpret the experimental phenomena. The conventional empirical scattering function is primarily applicable to small grazing angles. Moreover, the regional segmentation method simulates reverberations across various receiving beams, ignoring scatterers in other areas. To address these issues, we substitute the empirical scattering function with a small-slope approximation (SSA) that is appropriate for full grazing angles. Furthermore, we utilize the beam pattern of arrays to incorporate the effects of each scatterer, and derive the expression for bottom reverberation intensity in both the array and beam domains. The established model demonstrates its applicability in simulating and interpreting the stripe structures of bottom reverberation, and the comparison shows that the model outputs are in agreement with the experimental results. The analysis indicates that the vertical stripes within the structures originate from eigenrays in the mirror reflection direction. Furthermore, the convex stripes are predominantly affected by the direct ray and the surface reflection ray among the scattered eigenrays, whereas the concave and elliptical stripes are primarily affected by the bottom-surface reflection ray and the surface-bottom-surface reflection ray within the scattered eigenrays.

Keywords: bottom reverberation; spatial-temporal structures; bistatic sonar; model establishment

Academic Editors: Xuebo Zhang, Haiyong Xu, Jingyue Lu, Lei Zhang, Marc Pinto and Farman Ali

Received: 20 November 2024
Revised: 6 January 2025
Accepted: 6 January 2025
Published: 8 January 2025

Citation: Su, Z.; Zhuo, J.; Sun, C. Modeling and Data Analysis of Bistatic Bottom Reverberation from a Towed Horizontal Array. *Remote Sens.* **2025**, *17*, 192. https://doi.org/10.3390/rs17020192

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As a critical boundary for ocean waveguides, the seabed exhibits significant reflection and scattering effects on acoustic waves. Bottom reverberation results from the superposition of scattering echoes generated by numerous seafloor scatterers, encapsulating rich information about the ocean environment [1,2], which is widely utilized in analyzing the seafloor substrate and the inversion of parameters [3,4]. Owing to multipath propagation, each combination of incident and scattered rays contributes to bottom reverberation, resulting in a multi-peak structure [5]. By employing a vertical or horizontal array, the receiving signals can be beamformed to obtain the spatial-temporal structures of a bottom reverberation. The complex spatial-temporal distribution is an important feature that is closely associated with seabed topography [5–7]. Utilizing numerical methods to model bottom reverberation, researchers can analyze the shape characteristics of spatial-temporal struc-

tures. This approach is significant and useful for engineering applications investigating the generation of these structures and their association with seafloor features.

Two approaches are prevalently employed to model reverberation. The first is the full-wave model [8–11], which utilizes the wave equation to depict the propagation and scattering of reverberation, offering a rigorous theoretical expression. The second is a combination of the sound propagation model and the scattering model. In the sound propagation model, bottom reverberation models mainly include normal mode-based models [12–14], ray-based models [15,16], and parabolic equation models [17,18]. This latter approach is more pragmatic and applicable in analyzing the characteristics of measured reverberation.

Previous studies have demonstrated that large-scale topography and coastal slopes give rise to intense bottom reverberation [19–23]. The above approach has been employed to identify unknown features by modeling reverberation and comparing model outputs with data [22,23]. Among these outputs, local peaks in reverberation are associated with bathymetric features, while abnormal peaks align with unknown seafloor features. In the Mid-Atlantic Ridge (MAR) experiment, a range-dependent parabolic equation model was employed to calculate transmission loss (TL). The comparison with the echoes reveals that significant scattered echoes originated from the steep escarpments and cliffs, and are highly correlated with TL [19–21]. Moreover, a bistatic sonar system with a horizontal receiving array and two low-frequency vertical source arrays was employed in the continental shelf area [6]. The obtained beam data were projected onto Cartesian coordinates, revealing that the echoes in the reverberation mainly came from a preset tubular target and an underground channel. Subsequently, the researchers utilized a range-dependent parabolic equation model to rectify the scattering intensity of the inflated target [24,25]. Additionally, Preston and Ellis [22,23] employed ray theory and the Lambert scattering function to model the reverberation received by the towed array. They superimposed a polar map of the reverberation onto a topographic map, identifying unknown features and extracting seabed information through model–data comparisons. Furthermore, an adiabatic normal mode-based model was employed to analyze and handle seafloor scattering and clutter echoes in the Target and Reverberation Experiment 2013 (TREX13) [26,27]. In addition, researchers have employed the scattering model and the statistical model to investigate the causes and statistical characteristics of strong scattering in seafloor reverberation on the Malta Plateau [28–31]. Furthermore, in an experiment conducted on the Ulleung Basin [32,33], a depth-bistatic bottom reverberation model was developed to predict reverberation from an active triple-towed array, and the influence of continental shelf topography on the drop-off regions in the bearing-time records was analyzed.

In the aforementioned studies, the Lambert scattering function was typically employed to compute the scattering intensity of the interface [15,34]. This empirical function mainly applies to small grazing angles, and its coefficients have no quantitative correlation with seabed parameters. Seabed areas are divided by azimuth angles to simulate reverberation at different receiving angles, ignoring scatterers in other regions [35]. Additionally, the scattering region that contributes to reverberation at a specific time is regarded as an elliptic ring in the conventional method. This approach disregards the disparity in the propagation time of eigenrays at each seafloor scatterer. Owing to multipath structures, scatterers that impact reverberation do not lie exactly on the same ellipse.

In a bistatic bottom reverberation experiment conducted in the northern South China Sea in August 2020, with a vertical transmitting array and a towed horizontal array, the stable spatial-temporal structures of bottom reverberation were observed at different receiving positions. These structures are related to the paths of the incident and scattered rays, presenting vertical, curved, and elliptical stripes in the beam-time domain.

This study presents a bistatic bottom reverberation model based on ray theory. These stripes are modeled, and the experimental phenomena are analyzed and interpreted. Instead of the empirical scattering function, the small-slope approximation (SSA) [36,37] is adopted for full grazing angles. The seabed is sectioned into rectangular grids to obtain accurate propagation delays, and the beam patterns of the transmitting array and the receiving array are utilized to account for each grid's influence on reverberation. The established model enables the calculation of reverberation intensity within the element or beam domain, and simulate the spatial-temporal structures of reverberation from the horizontal array. The model outputs reproduce the stripe structures observed in the experiment, interpreting the generation of the structures by specifying the types of incident and scattered rays.

This paper is organized as follows: Section 2 establishes a bistatic bottom reverberation model with a vertical transmitting array and a horizontal receiving array. Section 3 presents an overview of the experiment. Section 4 compares the model results and the experimental results, and explains the experimental phenomena. Finally, Section 5 summarizes the conclusions.

2. Bistatic Bottom Reverberation Model

2.1. Bistatic Bottom Reverberation Scene and the Parameters

The bistatic system consists of a vertical transmitting array and a towed horizontal receiving array; the bottom reverberation scene is depicted in Figure 1. Each transmission is referred to as a single ping. During each transmission, the sound waves emitted by the source are scattered on the seafloor and then propagated to the receiver, resulting in bottom reverberation. Typically, sound propagation models are combined with scattering models to simulate reverberation. In this section, the ray model is employed to describe underwater acoustic propagation, and the small-slope approximation (SSA) model is utilized to address the scattering process of the eigenrays with different grazing angles. The influence of multipath propagation is considered in this approach, with the scattered signals along various paths from scatterers contributing to the reverberation signal.

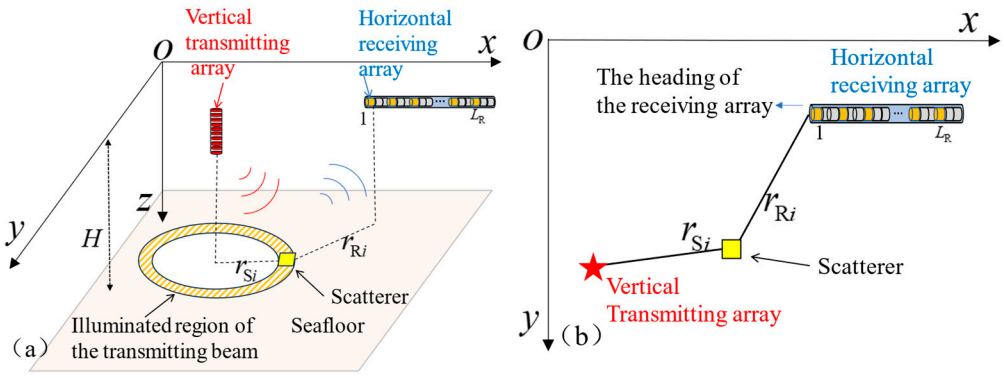


Figure 1. The scene of the bistatic bottom reverberation. (**a**) Bistatic sonar geometry and an element scattering area. (**b**) Top view of the scene.

The water depth in a horizontally layered environment is denoted as H, and the sound speed is c. Assume that the source is a vertical linear array of L_S elements with an element spacing of d_S. The center coordinate of the source is denoted as (x_s, y_s, z_S). A horizontal linear array is the receiving array, consisting of L_R elements with an element spacing of d_R. The first receiving element is the reference element, with the coordinate denoted as (x_R, y_R, z_R). Additionally, the coordinate of the unit scatterer is denoted as (x_i, y_i, z_i), and the horizontal distances to the source and the reference element of the receiving array are r_{Si}

and r_{Ri}, respectively. The heading angle of the receiving array is denoted as γ. In this study, γ is defined as 0° in the positive direction of the x-axis, with the counterclockwise direction corresponding to positive angles. To derive and present the equations in Section 2.2, we introduce the following vectors: $\mathbf{r}_{Si} = (x_i - x_S, y_i - y_S)$, $\mathbf{r}_{Ri} = (x_R - x_i, y_R - y_i)$, and $\mathbf{r}_\gamma = (\cos\gamma, \sin\gamma)$.

Figure 2a,b illustrates the schematic diagrams of the transmitting beam angle and the receiving beam angle, respectively. For the vertical linear array in Figure 2a, the beam pattern is symmetrical around the axis of the array within the xy-plane. In the vertical dimension, the transmitting beam angle is designated as θ_S, and the clockwise direction corresponds to the positive angle of the transmitting beam. Specifically, θ_S is denoted as $-90°$ when the beam angle is toward the sea surface, and 90° when the beam angle is toward the seafloor. For the horizontal linear array in Figure 2b, the beam pattern exhibits symmetry around the axis of the array within the z plane, and the receiving beam angle is denoted as θ_R. θ_R is equal to $-90°$ when the beam is directed toward the head of the receiving array, and 90° when the beam points at the tail of the array.

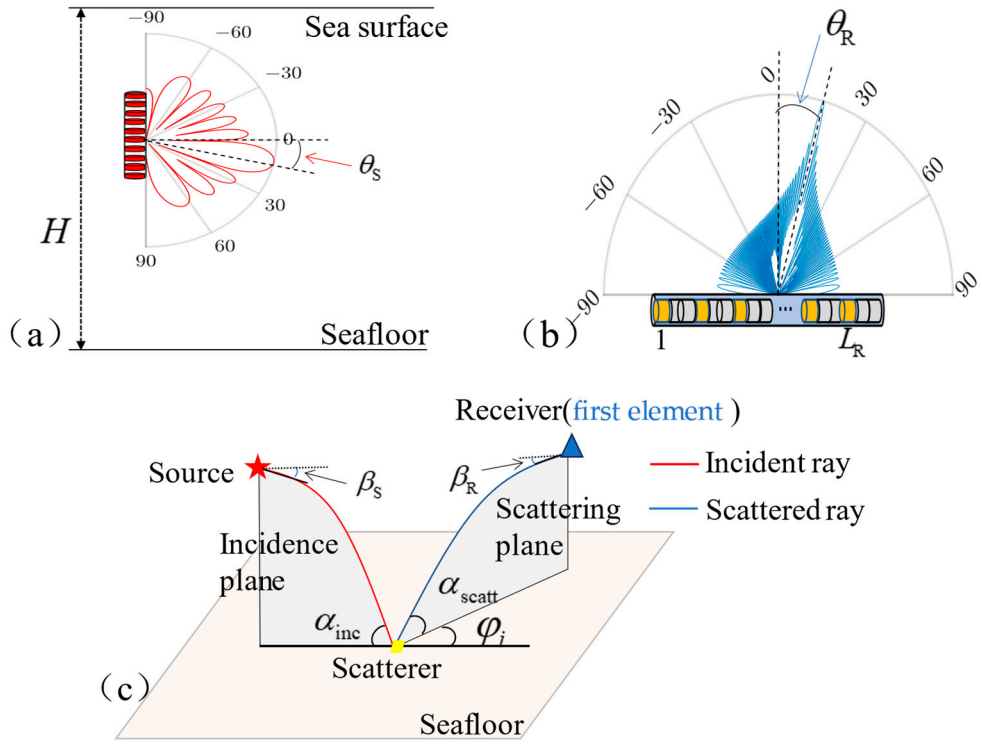


Figure 2. Angle parameters involved in the model. (**a**) The transmitting beam angle θ_S. (**b**) The receiving beam angle θ_R. (**c**) The exit grazing angle β_S of the incident ray, the arrival grazing angle β_R of the scatter, the incident grazing angle α_{inc}, the scattered grazing angle α_{scatt}, and the scattered azimuth angle φ_i at the seafloor scatterer.

The incident plane and the scattering plane are typically non-coplanar, as illustrated in Figure 2c. The variables β_S and β_R represent the exit grazing angle of the incident ray at the source and the arrival grazing angle of the scattered ray at the first receiving element, respectively. In the three-dimensional scattering function $g(\alpha_{inc}, \alpha_{scatt}, \varphi_i)$ of the seafloor interface, the variables α_{inc}, α_{scatt}, and φ_i represent the incident grazing angle, the scattered grazing angle, and the scattered azimuth angle, respectively. In addition, the incidence plane and the scattering plane are coplanar within the vertical plane when $\varphi_i = 0°$, which is referred to as the mirror reflection direction.

2.2. Theoretical Derivation of Expressions for Bistatic Bottom Reverberation Model

According to ray theory, we can suppose that the coordinate of the point source is (x_S, y_S, z_S); thus, the sound pressure at the unit scatterer with the coordinate (x_i, y_i, z_i) can be expressed as

$$\begin{cases} p(x_i, y_i, z_i) = \sum_{m=1}^{M} A_{\mathrm{inc},m} e^{j\omega\tau_{\mathrm{inc},m}} \\ A_{\mathrm{inc},m} = \sqrt{\dfrac{W\cos\beta_{Sm}}{r_{Si}\left|\partial r_{Si}/\partial\beta\right|_{\beta_{Sm}}\sin\alpha_{\mathrm{inc},m}}} \\ \tau_{\mathrm{inc},m} = \left(r_{Si}\cos\beta_{Sm} + \int_{z_S}^{H}\sqrt{\eta^2(z)-\cos^2\beta_{Sm}}dz\right)/c \end{cases} \tag{1}$$

where M is the number of incident eigenrays, and $A_{\mathrm{inc},m}$ and $\tau_{\mathrm{inc},m}$ are the amplitude and propagation delay of the mth incident eigenray, respectively. $\omega = 2\pi f$ is the angular frequency and f is the frequency of the source. W represents the radiated power per unit solid angle, and $r_{Si} = |\mathbf{r}_{Si}|$ represents the horizontal distance between the source and the seafloor scatterer. β_{Sm} and $\alpha_{\mathrm{inc},m}$ are the exit grazing angle of the mth incident eigenray at the source and the incident grazing angle at the seafloor, respectively. In addition, $\eta(z) = c(z)/c(z_S)$ is the refractive rate.

For the vertical transmitting array, the transmitting signal of each element can be weighted through beamforming methods. Consequently, the sound pressure in Equation (1) can be represented as

$$p(x_i, y_i, z_i) = \sum_{m=1}^{M} A_{\mathrm{inc},m} B_S(\beta_{Sm};\theta_S) e^{j\omega\tau_{\mathrm{inc},m}} \tag{2}$$

where $B_S(\beta_{Sm};\theta_S)$ is the beam pattern of the transmitting array, and θ_S represents the polar angle of the transmitting beam.

According to the reciprocity theorem, the sound pressure received by the reference element of the receiving array can be expressed as follows (the coordinates (x_i, y_i, z_i) are omitted)

$$p_1 = \sum_{m=1}^{M}\sum_{n=1}^{N} A_{\mathrm{inc},m} A_{\mathrm{scatt},n} B_S(\beta_{Sm};\theta_S) g(\alpha_{\mathrm{inc},m}, \alpha_{\mathrm{scatt},n}, \varphi_i) e^{j\omega(\tau_{\mathrm{inc},m}+\tau_{\mathrm{scatt},n})} \tag{3}$$

where N is the number of the scattered eigenrays, and $A_{\mathrm{scatt},n}$ and $\tau_{\mathrm{scatt},n}$ are the amplitude and propagation delay of the nth scattered eigenray, respectively. The amplitudes and delays of both the incident and scattered eigenrays are computed utilizing the Bellhop model [38].

The three-dimensional scattering model in the literature [37] is expressed as

$$\begin{aligned} G(\alpha_{\mathrm{inc},m}, \alpha_{\mathrm{scatt},n}, \varphi_i) &= \left|\mathrm{g}(\alpha_{\mathrm{inc},m}, \alpha_{\mathrm{scatt},n}, \varphi_i)\right|^2 \\ &= \frac{k_{ww}^2 |R_{ww}(\alpha_{\mathrm{inc},m}, \alpha_{\mathrm{scatt},n}, \varphi_i)|^2}{2\pi(\Delta Q)^2(\Delta Q_z)^2} \cdot \int_0^{+\infty} J_0(u) u e^{-qu^{2v}} du \end{aligned} \tag{4}$$

where

$$\Delta Q = k_{ww}\sqrt{\cos^2\alpha_{\mathrm{inc},m} + \cos^2\alpha_{\mathrm{scatt},n} - 2\cos\alpha_{\mathrm{inc},m}\cos\alpha_{\mathrm{scatt},n}\cos\varphi_i} \tag{5}$$

$$\Delta Q_z = k_{ww}(\sin\alpha_{\mathrm{inc},m} + \sin\alpha_{\mathrm{scatt},n}) \tag{6}$$

$$q = \frac{\pi w_2 (\Delta Q_z)^2}{(2\cdot\Delta Q)^{2v}} \cdot \frac{\Gamma(2-v)}{v(1-v)\Gamma(1+v)} \tag{7}$$

$$v = \frac{\gamma_2}{2} - 1 \tag{8}$$

where k_{ww} is the wavenumber, ΔQ represents the magnitude of the difference between the horizontal components of the incident wave vector and that of the scattered wave vector, and ΔQ_z corresponds to the difference between the vertical components of these two vectors. In the "Kirchhoff integral" to the right of Equation (4), J_0 is the 0-order Bessel function of the first kind, and Γ is the gamma function. w_2 and γ_2 are the spectral strength and spectral exponent for the rough interface, respectively. The expression of variable $R_{ww}(\alpha_{\mathrm{inc},m}, \alpha_{\mathrm{scatt},n}, \varphi_i)$ varies depending on the type of seafloor utilized, and the specific representation employed in this study is detailed in Appendix A.

By employing the far-field assumption, the horizontal range from the bottom scatterer to the lth element of the receiving array can be approximately expressed as

$$\begin{cases} r_{\mathrm{R}i,l} \approx r_{\mathrm{R}i} + (l-1)d_{\mathrm{R}} \sin\theta_i, & l = 1, \ldots, L_{\mathrm{R}} \\ \theta_i = \langle \mathbf{r}_\gamma, -\mathbf{r}_{\mathrm{R}i} \rangle - 90^\circ \end{cases} \tag{9}$$

where $\langle \cdot \rangle$ denotes the calculation of the angle between two vectors.

Consequently, the propagation delay of the nth scattered eigenray to the lth element is expressed as

$$\tau_{\mathrm{scatt},n}(l) = \tau_{\mathrm{scatt},n} + (l-1)d_{\mathrm{R}} \sin\theta_i \cos\beta_{Rn}/c_0 \tag{10}$$

where c_0 is the sound speed at the receiver. The variable $\theta'_{i,n}$ is defined as $\sin\theta'_{i,n} = \sin\theta_i \cos\beta_{Rn}$, representing the arrival angle of the nth scattered eigenray at the receiving array.

Combined with Equations (3) and (9), the sound pressure of the lth element can be expressed as

$$p_l = \sum_{m=1}^{M}\sum_{n=1}^{N} A_{\mathrm{inc},m} A_{\mathrm{scatt},n} B_{\mathrm{S}}(\beta_{\mathrm{S}m}; \theta_{\mathrm{S}}) g(\alpha_{\mathrm{inc},m}, \alpha_{\mathrm{scatt},n}, \varphi_i) e^{j\omega(\tau_{\mathrm{inc},m}+\tau_{\mathrm{scatt},n})} e^{j\omega(l-1)d_{\mathrm{R}}\sin\theta'_{i,n}/c_0} \tag{11}$$

Accordingly, the incoherent reverberation intensity of the element is expressed as

$$I_{l,i} = \sum_{m=1}^{M}\sum_{n=1}^{N} |A_{\mathrm{inc},m}|^2 |A_{\mathrm{scatt},n}|^2 |B_{\mathrm{S}}(\beta_{\mathrm{S}m}; \theta_{\mathrm{S}})|^2 |G(\alpha_{\mathrm{inc},m}, \alpha_{\mathrm{scatt},n}, \varphi_i)| \tag{12}$$

At a specific time t, the reverberation of the lth element is represented as the superposition of scattered signals from different scatterers. Then, the reverberation intensity is expressed as

$$RL_l(t) = I_0 \cdot \sum_{i=1}^{K} I_{l,i} \Delta S_i \tag{13}$$

where I_0 is the intensity of the source, t is denoted as $t = \tau_{\mathrm{inc},m} + \tau_{\mathrm{scatt},n}$, K is the number of scatterers contributing to the reverberation, and ΔS_i is the area of the ith scatterer.

The sound pressure of each element in Equation (11) is weighted using the beamforming method. When the angle of the receiving beam is denoted as θ_{R}, the beam output is represented as

$$\begin{aligned} p_{\mathrm{beam}}(\theta_{\mathrm{R}}) &= \sum_{l=1}^{L_{\mathrm{R}}} w_l{}^* p_l \\ &= \sum_{m=1}^{M}\sum_{n=1}^{N} A_{\mathrm{inc},m} A_{\mathrm{scatt},n} B_{\mathrm{S}}(\beta_{\mathrm{S}m}; \theta_{\mathrm{S}}) B_{\mathrm{R}}\left(\theta'_{i,n}; \theta_{\mathrm{R}}\right) g(\alpha_{\mathrm{inc},m}, \alpha_{\mathrm{scatt},n}, \varphi_i) e^{j\omega(\tau_{\mathrm{inc},m}+\tau_{\mathrm{scatt},n})} \end{aligned} \tag{14}$$

where the superscript * denotes the conjugation, w_l is the weighted value of the lth element, and $B_{\mathrm{R}}\left(\theta'_{i,n}; \theta_{\mathrm{R}}\right)$ is the beam pattern of the receiving array.

Consequently, the corresponding incoherent reverberation intensity can be expressed as

$$I_{\mathrm{beam},i}(\theta_{\mathrm{R}}) = \sum_{m=1}^{M}\sum_{n=1}^{N} |A_{\mathrm{inc},m}|^2 |A_{\mathrm{scatt},n}|^2 |B_{\mathrm{S}}(\beta_{\mathrm{S}m};\theta_{\mathrm{S}})|^2 \left|B_{\mathrm{R}}\left(\theta'_{i,n};\theta_{\mathrm{R}}\right)\right|^2 |G(\alpha_{\mathrm{inc},m},\alpha_{\mathrm{scatt},n},\varphi_i)| \quad (15)$$

Similar to Equation (13), the reverberation intensity of the beam angle θ_{R} is expressed as

$$RL_{\mathrm{beam}}(\theta_{\mathrm{R}},t) = I_0 \cdot \sum_{i=1}^{K} I_{\mathrm{beam},i}(\theta_{\mathrm{R}})\Delta S_i \quad (16)$$

where the definitions of parameters I_0, t, K, and ΔS_i are consistent with those in Equation (13).

The theoretical derivation indicates that there is a phase difference in the signal waveform of each element, while the incoherent reverberation intensity of each element is consistent in the array domain. In the beam domain, the signal from each scatterer is modulated by the beam pattern of the receiving array. As a result, the incoherent reverberation intensity within the beam varies with the angle. In summary, Figure 3 presents a flow chart illustrating the calculation of bistatic bottom reverberation in both the array and beam domains.

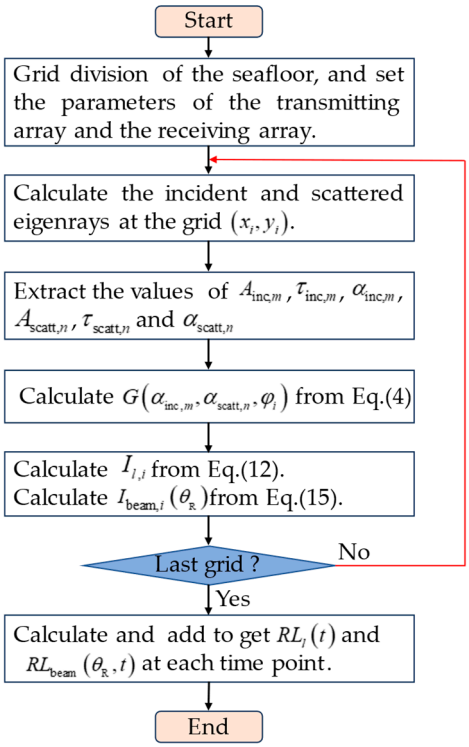


Figure 3. Flow chart for calculating the bistatic bottom reverberation.

3. Overview of Sea Trial Experiment Conducted in the South China Sea

3.1. Experiment Description

In August 2020, an experiment was conducted in the northern South China Sea. The topography of the experimental area, the source site, and the tracks of the receiver are shown in Figure 4a, where "☆" indicates the position of the source and "↑" represents the heading direction of the receiving array. The water depth of the area is approximately 1100 m, with islands and reefs situated to the northwest. In addition, the coordinates of the source are 111.56°E, 16.88°N.

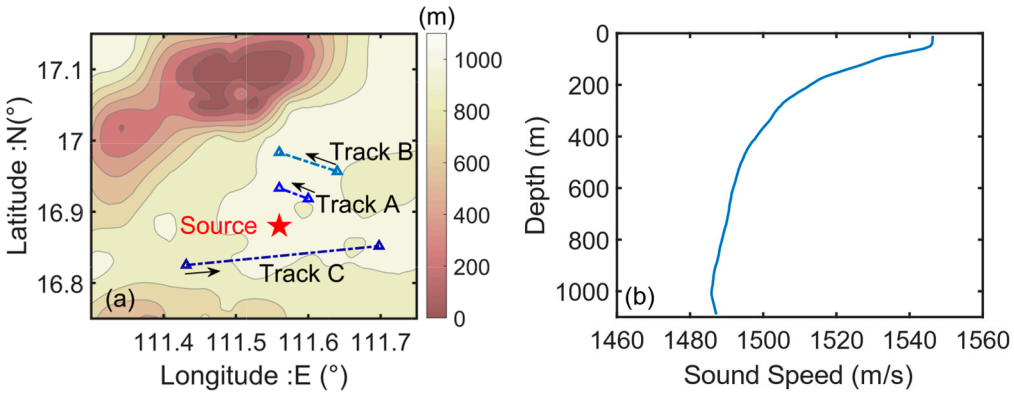


Figure 4. Experimental environment. (**a**) Bathymetry of the area, with depth data obtained from the ETOPO1. (**b**) The sound speed profile measured in the experiment.

The sound speed profile obtained from the experiment is depicted in Figure 4b, exhibiting a typical negative gradient. The sound speed was constant at approximately 1544 m/s within a depth range of 0 to 50 m, and it decreased to about 1484 m/s at the seabed.

The transmitting and receiving arrays utilized in the experiment are shown in Figure 5. The specific parameters are as follows: the source was a 10-element vertical linear array, with an element spacing of 0.42 m, and the receiver was a 96-element towed horizontal linear array, with an element spacing of 0.416 m. Furthermore, the transmitted signal was a linear frequency modulation (LFM) with a frequency ranging from 1700 Hz to 1900 Hz. The duration of each transmission was 1 s, with an interval of 50 s. The source level for each element in the transmitting array was 190 dB, and the sensitivity level of the receiving hydrophone was −142 dB. Additionally, the signal sampling frequency was set 16 kHz.

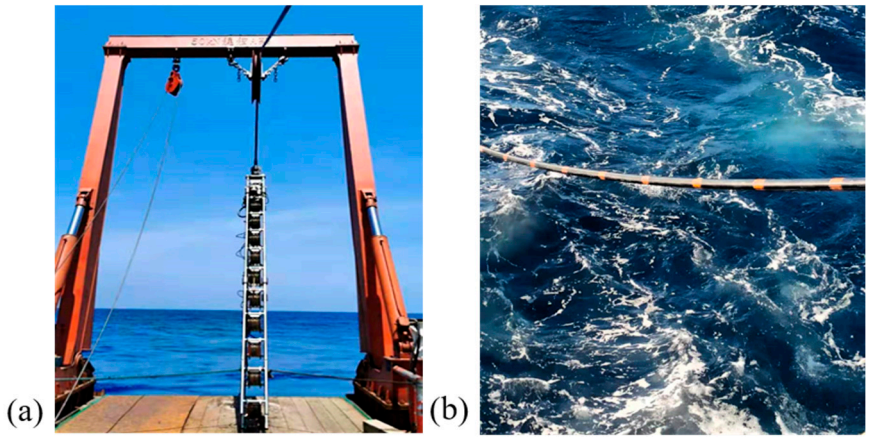


Figure 5. The equipment used in the experiment. (**a**) The transmitting array. (**b**) The receiving array.

During the experiment, the position of the transmitting array was fixed, with a central depth of approximately 52 m. Moreover, the conventional beamforming method was employed to adjust the transmitting beam angle in each transmission. The transmitting beam angle was set to either 0°or 12° in different received tracks. Specifically, the receiving vessel traveled twice along Track A with these transmitting beam angles. Subsequently, the receiving vessel moved along Tracks B and C once each, with transmitting beam angles of 12° for Track B and 0° for Track C.

3.2. Experimental Data Processing

In this section, the data processing is briefly described to obtain the reverberation intensities of the array and beam domains. The received signal of the lth element of the receiving array is denoted as ${p'}_l(t)$. After being processed through matched filtering, the Fourier transform is applied to derive the spectrum ${pp'}_l(f)$. At a specific time t, the signal power of the lth element at the center frequency $f_0 = 1800$ Hz is expressed as

$$ {p'}_l(f_0) = \frac{1}{J' f_s (nf_2 - nf_1 + 1)} \sum_{j'=nf_1}^{nf_2} \left| {pp'}_l(j') \right|^2 \tag{17} $$

where J' represents the number of data sampling points within the period centered on time t, and f_s is the sampling frequency. The variables nf_1 and nf_2 denote the number of frequency points corresponding to the lower and upper bounds of the signal frequency, respectively.

In logarithmic form (unit: dB), the reverberation intensity of the element is as follows

$$ {RL'}_l(t) = 10 \log_{10} \left[{p'}_l(f_0) \right] - SL - DI - M_v - E_c \tag{18} $$

where SL denotes the source level of a single emitter, and DI represents the directivity index of the transmitting array. M_v is the sensitivity level of the receiving hydrophone and E_c is the processing gain of the matched filter.

Assume that the receiving beam angle is denoted as θ_{rec}. The received signal is processed through conventional beamforming and matching filtering to obtain the beam output, denoted as ${p'}_{\text{beam}}(\theta_{\text{rec}}, t)$, and the variable $pp_{\text{beam}}(\theta_{\text{rec}}, f)$ represents the spectrum of ${p'}_{\text{beam}}(\theta_{\text{rec}}, t)$. The average power of the beam output at the center frequency $f_0 = 1800$ Hz is expressed as

$$ {p'}_{\text{beam}}(\theta_{\text{rec}}, f_0) = \frac{1}{J' f_s (nf_2 - nf_1 + 1)} \sum_{j'=nf_1}^{nf_2} \left| pp_{\text{beam}}(\theta_{\text{rec}}, j') \right|^2 \tag{19} $$

where the definitions of the variables J', f_s, nf_1, and nf_2 in Equation (19) are consistent with those in Equation (17).

Similarly, the variable ${p'}_l(f_0)$ in Equation (18) is substituted with the variable ${p'}_{\text{beam}}(\theta_{\text{rec}}, f_0)$. Accordingly, the expression of reverberation intensity at the receiving beam angle θ_{rec} is expressed as

$$ {RL'}_{\text{beam}}(\theta_{\text{rec}}, t) = 10 \log_{10} \left[{p'}_{\text{beam}}(\theta_{\text{rec}}, f_0) \right] - SL - DI - M_v - E_c \tag{20} $$

where the definitions of the variables SL, DI, M_v, and E_c accord with those in Equation (18).

4. Comparison and Interpretation of Simulation and Experimental Results

Following the processing steps in Section 3, the reverberation intensities at a source level of 0 dB in both the array and beam domains are obtained. The results for the experimental data and a comparison between the model and experimental results are presented in this section.

4.1. Parameters of the Simulation Environment

Figure 6 illustrates the simulation environment and the parameters. A single-layer seafloor model was employed, with a seafloor depth of 1100 m. In simulations, we take into account the bathymetry of the experimental area. An $N \times 2D$ ray model is utilized to calculate incident and scattered eigenrays on the seabed grids at different distances,

with the calculations centered on the positions of the source and reference element of the receiving array.

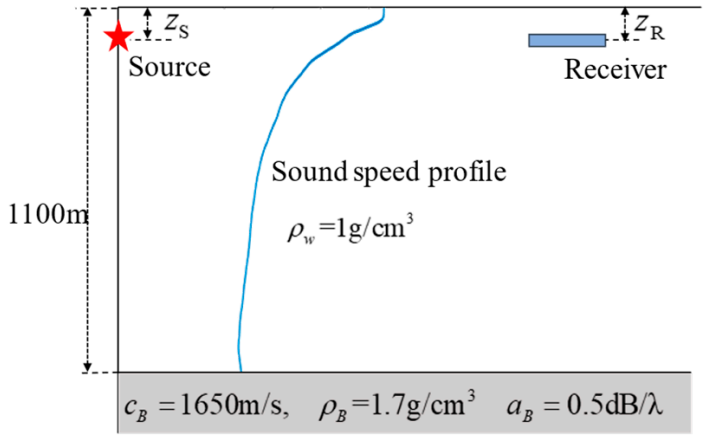


Figure 6. The parameters of the simulation environment.

As the seafloor substrate was not collected during the experiment, a seafloor with sandy substrate was employed and the parameters were appropriately modified. The parameters of the bottom are detailed in Table 1, where a_ρ and a_{wp} represent the ratios of the density and sound speed at the bottom to those in the water, respectively. The variable a_B denotes the attenuation expressed in decibels per wavelength. In addition, w_2 and γ_2 are the spectral strength and spectral exponent for the rough interface, respectively.

Table 1. The parameters of the bottom, where a_ρ, a_{wp}, and γ_2 are dimensionless values.

$a_\rho = \rho_B/\rho_w$	$a_{wp} = c_B/c_w$	a_B (dB/λ)	w_2 (m$^{4-\gamma_2}$)	γ_2
1.7	1.1	0.5	0.006957	3.3

The sound speed profile obtained from measurements is utilized in the simulation. Furthermore, the parameters of the source and the receiving array used in the simulation are consistent with the equipment parameters in the experiment, and the frequency of the sound source is set at 1800 Hz.

4.2. Comparison of Simulation and Experimental Results

This section compares the model output and the experimental results. Regarding the source position as the origin, the latitude and longitude coordinates in Figure 4a were converted into Cartesian coordinates, as depicted in Figure 7. Here, the positive directions of the x-axis and y-axis correspond to the east and north, respectively. We selected experimental data from eight receiving positions, with one signal transmission conducted at each position. The receiving signals at each of these positions were processed according to the procedure described in Section 3.2. The specific parameters are presented in Table 2, including the transmitting beam angle θ_S, the receiving range r_{SR}, depth z_R, the heading angle γ, and the azimuth angle ϕ of the receiver relative to the source in the xoy-plane. Additionally, ϕ is defined as 0° in the positive direction of the x-axis, with the counterclockwise direction corresponding to the positive angles.

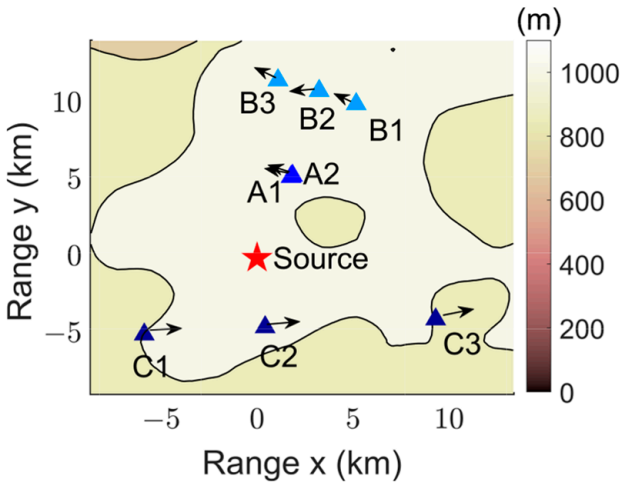


Figure 7. The receiving positions selected for the tracks.

Table 2. Parameter details at different receiving positions.

Position	Transmitting Beam Angle θ_S (°)	Receiving Range r_{SR} (km)	Depth of the Receiver z_R (m)	γ (°)	ϕ (°)
A1	0	5.67	50.39	158.87	70.98
A2	12	5.57	59.30	163.20	71.47
B1	12	11.28	60.50	159.14	62.74
B2	12	11.36	53.93	190.70	73.45
B3	12	11.64	56.86	161.38	84.62
C1	0	7.78	66.64	2.82	219.85
C2	0	4.53	66.15	8.48	274.08
C3	0	10.12	65.10	11.95	338.62

4.2.1. The Results of Track A

Figure 8a,b presents the measured reverberation intensity and model outcomes for the first element of the receiving array at positions A1 and A2, respectively. Both of the figures demonstrate that the model calculations accord with the experimental results. The reverberation signals appear within 4.1 s to 15 s, and the intensity gradually decreases with time. According to the eigenrays and the corresponding arrival times presented in Figures A1 and A2 in Appendix B, the first peak in Figure 8a corresponds to the direct (D) ray and the surface reflection (SR) ray, and the second peak corresponds to the primary bottom reflection rays. In Figure 8b, three peaks correspond to D and SR, primary and secondary bottom reflection rays. Additionally, the primary bottom reflection rays consist of four types, namely, the bottom reflection (BR) ray, the bottom-surface reflection (BSR) ray, the surface-bottom reflection (SBR) ray and the surface-bottom-surface reflection (SBSR) ray.

Table 3 presents the value of the exit grazing angle at the source. At positions A1 and A2, although minor variation exists in the receiving range (r_{SR}) and the receiving depth (z_r), the arrival structures of the received signal are similar in both figures. However, when the transmitting beam angle is adjusted from 0° to 12°, the amplitudes of the multipath components change significantly. In Figure 8a, the exit grazing angles of D and SR are close to 0°, that is, the two rays lie on the main lobe of the transmitted beam, where the amplitudes are the highest. Whereas, the exit grazing angles of the four bottom reflection rays are far away from the main lobe, thus, their amplitudes are not as strong.

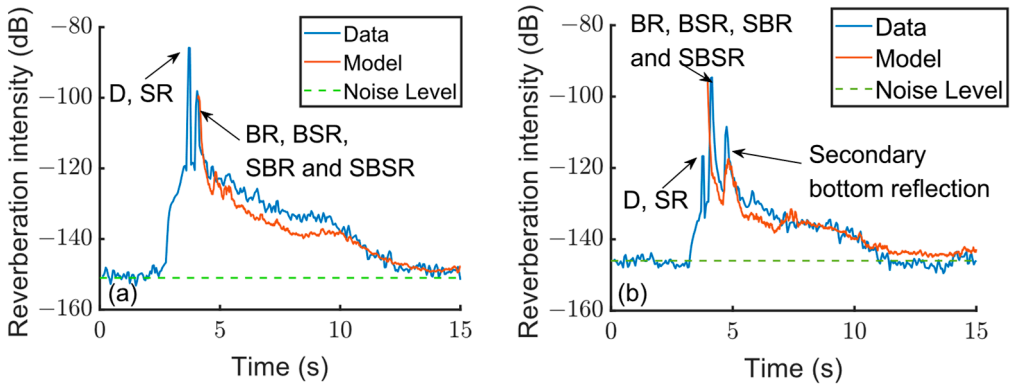


Figure 8. Comparison of the measured reverberation intensity with the model result for the first element of the receiving array at positions A1 and A2. (**a**) Position A1. (**b**) Position A2.

Table 3. The exit grazing angle of the eigenrays at the source.

Position	The Exit Grazing Angle at the Source					
	D	**SR**	**BR**	**BSR**	**SBR**	**SBSR**
A1	−1.85°	−2.02°	14.87°	16.46°	−16.51°	−17.92°
A2	−1.83°	−2.10°	15.12°	16.91°	−16.77°	−18.35°

The transmitting beam angle is 12° in Figure 8b. In this scene, the exit grazing angles of BR and BSR lie within the main lobe, while the exit grazing angles of D and SR are away from the beam angle. Consequently, the amplitudes of BR and BSR are higher than those of D and SR. Additionally, the amplitudes of the secondary bottom reflection increase. Only the signals scattered from the seafloor are taken into account; therefore, the model is incapable of calculating the direct and sea surface reflections between the source and the receiver. Due to the difference between the recorded parameters (such as source depth and receiving range) and the actual values in the experiment, as well as the relief of the seafloor, the intensity curve output by the model presents a certain deviation from the measured data.

Figure 9 shows the variations in reverberation intensity over time in beam domain on the Track A. Here, Figure 9a,c presents the experimental results at positions A1 and A2, and Figure 9b,d shows the corresponding simulations, respectively. The simulations present similar spatial-temporal structures and the intensity distribution of the bottom reverberation to the experimental results. The overall structures exhibit a "| < >" shape in the beam-time domain, corresponding to vertical, convex, and concave stripes. Moreover, the reverberation intensity gradually decreases with time at each beam output.

Figure 9a shows that peaks appear at 3.84 s and 4.1 s at a receiving beam angle of −6.8 degrees. Here, the first peak corresponds to D and SR, and the second peak corresponds to the four paths of the primary bottom reflection. In the mirror reflection direction, the propagation time of the aforementioned eigenrays reaching the receiver is earlier than in other directions. When the receiving beam angle varies, the amplitudes of the eigenrays entering from the side lobes remain high at these specific times. Therefore, the amplitude distribution of these eigenrays across the whole receiving beam angles constitutes the vertical stripes within the spatial-temporal structures of the bottom reverberation.

The convex and concave stripes are formed by the contribution of the scattered eigenrays from the seafloor grids. The convex stripe emerges after the primary bottom reflection, ranging from 4.1 s to 7.5 s, and the extent of the stripe gradually expands over time in the beam domain. Subsequently, the concave stripe emerges within the period from 7.5 s to 10 s.

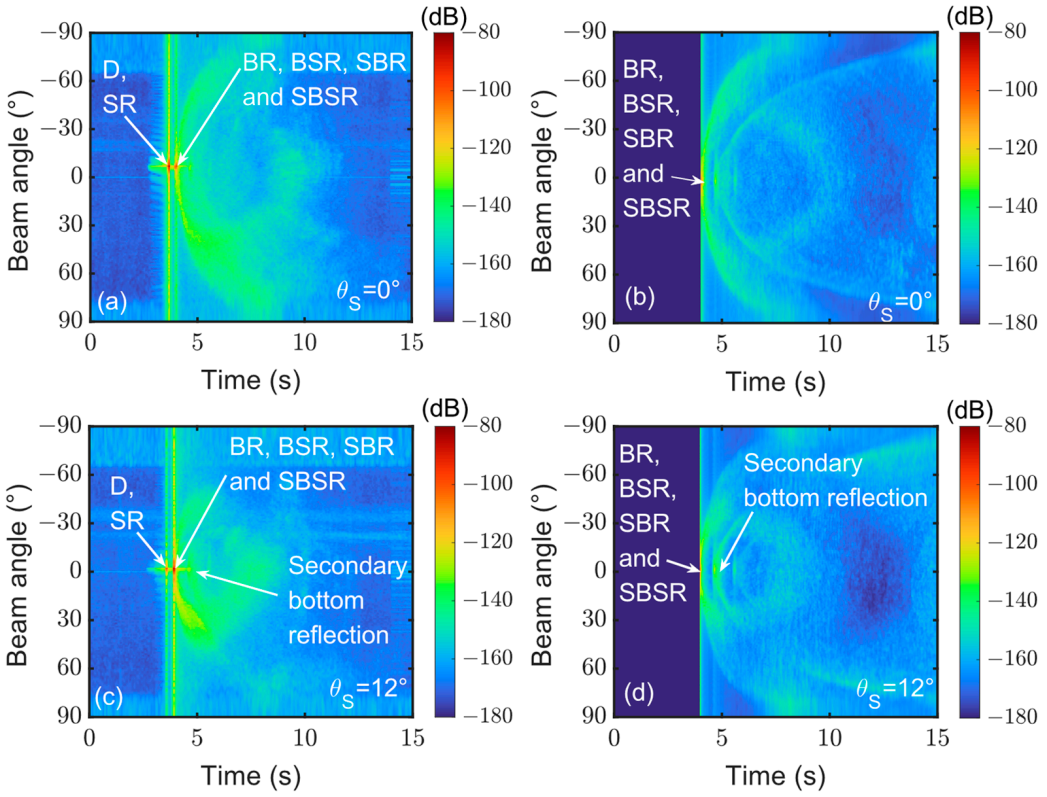


Figure 9. Comparison of experimental results and model outputs in the beam domain at positions A1 and A2. (**a**) Experimental results at position A1. (**b**) Model outputs at position A1. (**c**) Experimental results at position A2. (**d**) Model outputs at position A2.

According to the results illustrated in Figure A4 in Appendix C, which is the spatial distribution of scatterers that contribute to the reverberation at position A1, it is evident that scatterers contributing to the reverberation at a specific time are not completely located on the same ellipse, but are dispersed across multiple ellipses. At a specific time, t = 4 s, the bottom reverberation is mainly from scattered signals of seabed grids in the mirror reflection direction. As time progresses, scattered signals from multiple directions reach the receiving array, resulting in the appearance of the convex stripe in bottom reverberation. As transmission loss increases with distance, the intensity and extent of this stripe are lower than the vertical stripe. At time t = 8 s, an arc is formed by the grids with higher amplitudes in the region between the source and the receiver. These grids are distributed near the mirror reflection direction, corresponding to the span of the concave stripes. Furthermore, the uplifted terrain to the northeast of the source results in relatively strong scattering. Consequently, the reverberation intensity is markedly higher at positive received beam angles compared to negative angles.

The spatial-temporal structures depicted in Figure 9c are analogous to those in Figure 9a. Since the transmitting beam angle is directed toward the sea floor, the extent of certain reflection areas is reduced. Consequently, the convex and concave stripes extend over a span of angles from −45 to 45 degrees, narrower than the range from −85 to 85 degrees in Figure 9a. Similarly, the reverberation intensity at positive angles is significantly higher than at negative angles. As illustrated in Figure 7, the topographic relief exists northeast of the source, and there is likely relatively strong scattering generated in this area.

4.2.2. The Results of Track B and Track C

This section presents the experimental results and simulations for the other two tracks. Figures 10 and 11 illustrate the experimental results and the model outputs for Track B and C, respectively. Similar to the illustration for Track A, the model outputs accord with the experimental results for these two tracks, demonstrating comparable spatial-temporal structures for the bottom reverberation.

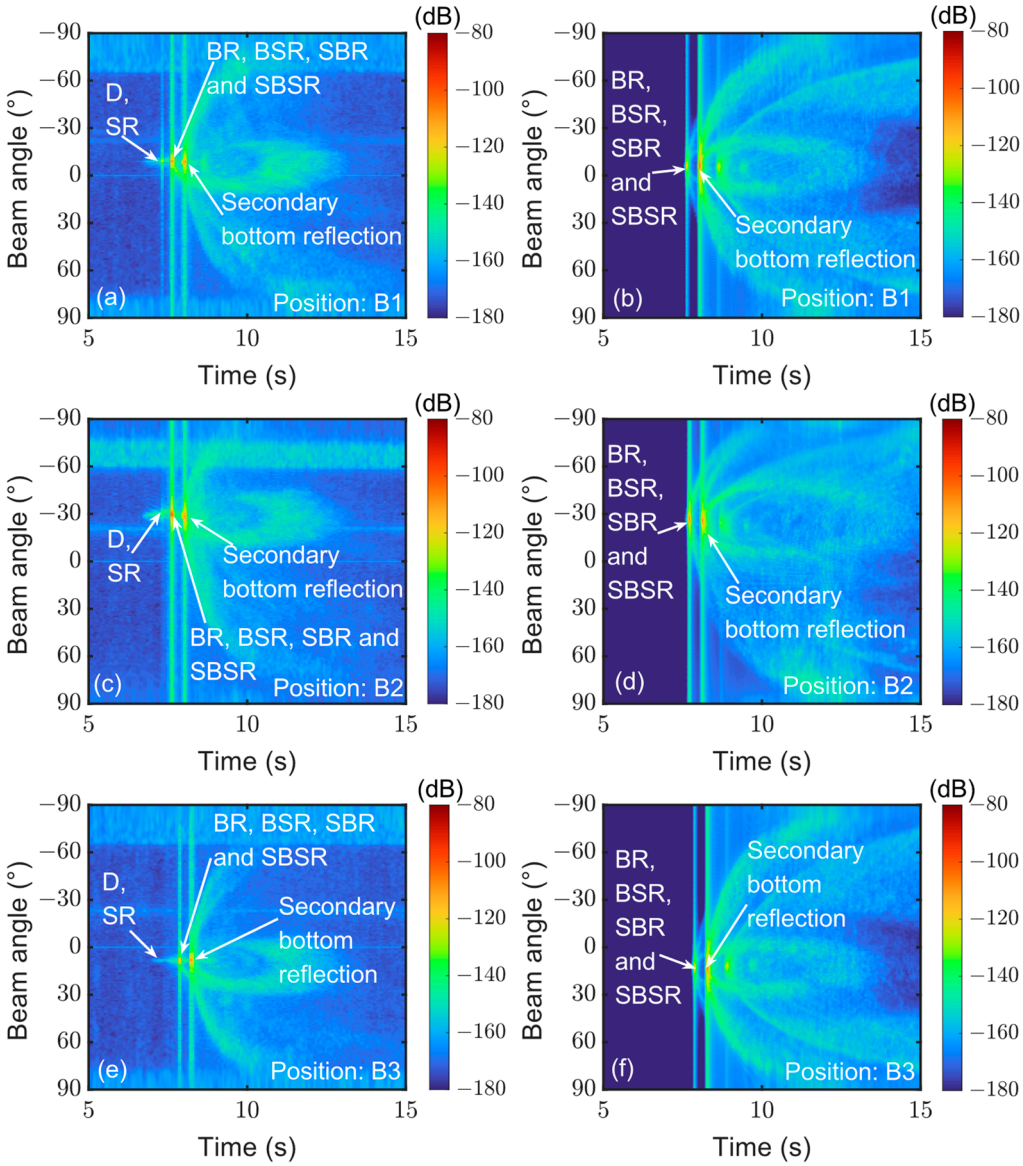


Figure 10. Comparison of experimental results and model outputs in the beam domain at positions B1, B2, and B3. (**a**) Experimental results at position B1. (**b**) Model outputs at position B1. (**c**) Experimental results at position B2. (**d**) Model outputs at position B2. (**e**) Experimental results at position B3. (**f**) Model outputs at position B3.

The receiving range r_{SR} and the heading angle γ at three positions of Track B are proximate, and the spatial-temporal structures presented in Figure 10a,c,e are similar. Meanwhile, the corresponding simulations at the positions are comparable. The receiving vessel moves from position B1 to B3, and the stripe structures shift from negative to positive angles.

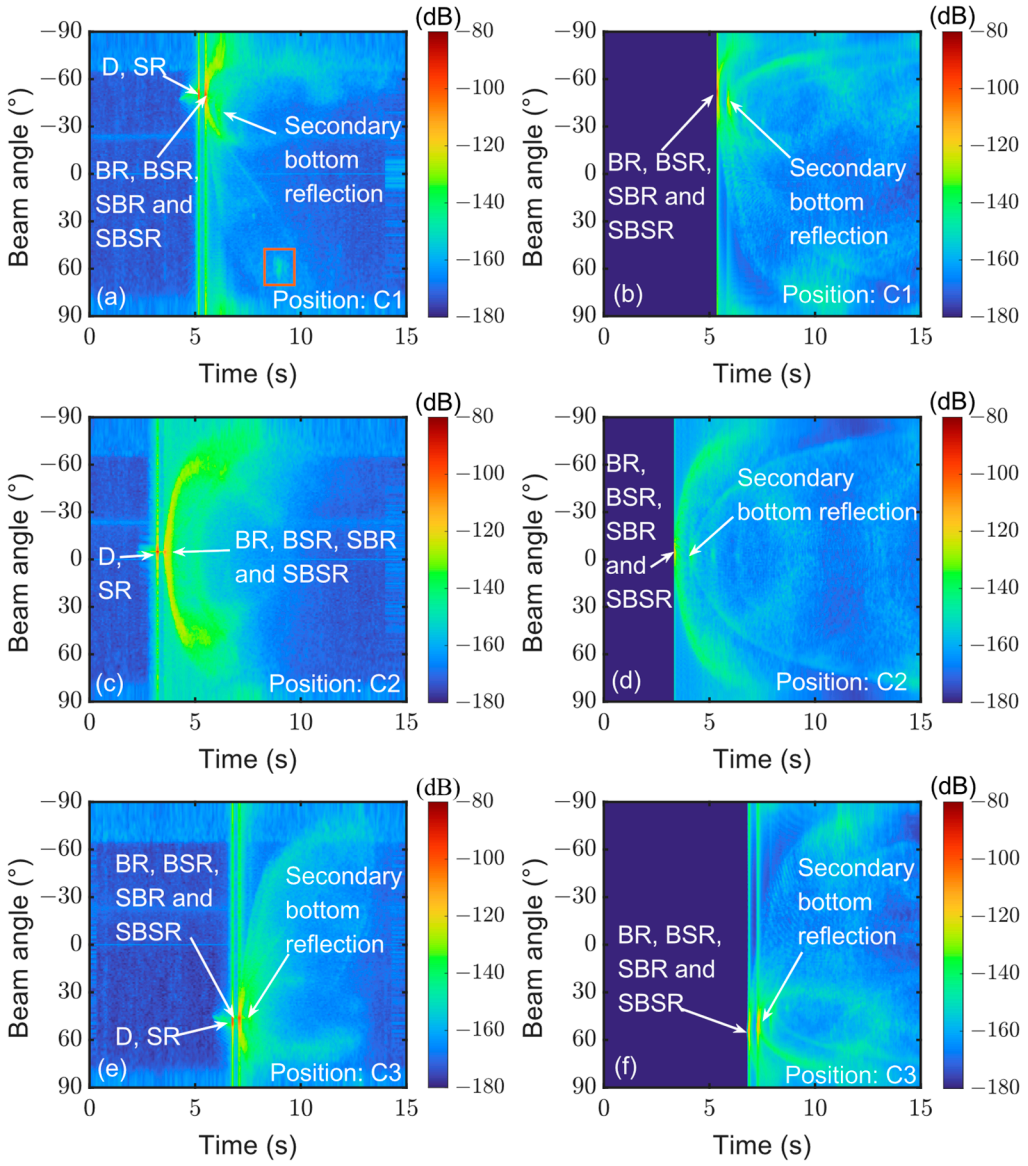


Figure 11. Comparison of experimental results and model outputs in the beam domain at positions C1, C2, and C3. (**a**) Experimental results at position C1. (**b**) Model outputs at position C1. (**c**) Experimental results at position C2. (**d**) Model outputs at position C2. (**e**) Experimental results at position C3. (**f**) Model outputs at position C3.

Two vertical stripes exist within the spatial-temporal structures after D and SR, corresponding to the primary and secondary bottom reflection, respectively. Certain disparities exist between the stripe structures in Track B and Track A. Specifically, as the main lobe of the transmitting beam is directed toward the seafloor, the amplitudes of the secondary bottom reflection increase.

The eigenrays, amplitudes, and arrival times presented in Figure A3 in Appendix B show that, although transmitting beam angles, θ_S, are the same for positions A2 and B1, exit grazing angles of eigenrays at the source are different. At position A2, the exit grazing angle of the secondary bottom reflection lies within the trough between the first and second sidelobes of the transmitting beam, leading to a low amplitude. In contrast, the exit grazing angle of the secondary bottom reflection is nearly aligned with the main lobe of the transmitting beam angle, resulting in a significantly higher amplitude relative to that at position A2. Furthermore, the concave stripe changes into an elliptical stripe following the convex stripe, and this stripe is concentrated in the proximity of the mirror reflection

direction. Based on the simulation illustrated in Figure A5 of Appendix C, it is observed that the scatterers contributing to the bottom reverberation depicted in Figure A5b–d are predominantly distributed along multiple ellipses in the area between the source and the receiver. The spatial distribution of scatterers in these three figures exhibits consistency across different time, and the scattered signals at these grids are the primary contributors to the formation of elliptical stripes.

Regarding the Track C results depicted in Figure 11, the positions ranging from C1 to C3 exhibit significant variations, and the stripe structures at the three positions present distinctions. At position C2, the source is in the broadside beam of the horizontal array. The stripe structures in Figure 11c,d are comparable to those at position A1, with a symmetrical distribution around a receiving beam angle of 0°.

Comparing Figure 11a,e, shows that the two stripes are similar, but their positions are situated at different angles. At positions C1 and C3, the incidence angles of the source are proximate to the negative and positive end-fire directions of the receiving array, respectively. The convex stripe and the subsequent elliptical stripe cover a narrow region of angles, and the region is concentrated near the mirror reflection direction.

At position C1, the source is situated within the negative receiving beam angle, in proximity to the negative end-fire direction of the receiving array. The region of scatterers significantly contributing to reverberation primarily encompasses the flat terrain in the mirror reflection direction, as well as the uplifted terrains to the northeast of the source and the receiver. Since position C1 is close to the boundary on the southwest of the experimental area, accordingly, the intensity of the upper half of the elliptical stripe is higher than that of the lower half. For position C3, situated in the southeast of the experimental area, the source is positioned at a positive receiving beam angle, proximate to the end-fire direction at the tail of the receiving array. The region of scatterers contributing to reverberation primarily encompasses the area in the mirror reflection direction, the uplifted terrain in the northeast of the source, and the uplifted terrain to the west of the receiver. Consequently, the intensity of the upper and lower halves in the elliptical stripe is contrary to that in position C1. Meanwhile, the simulations in Figure 11b,d generate results consistent with the above-mentioned phenomena.

In addition, significant scattering echoes exist within the rectangle in Figure 11a, corresponding to a receiving beam angle of 60° and a time ranging from 8.8 s to 9.3 s. This phenomenon was not discovered in other experimental results or simulations. The experimental records show that a fish shoal was near position C1, and we hypothesize that the echoes resulted from the fish scattering.

4.3. Interpretation of Stripes Observed in the Beam Output

To determine which eigenrays primarily constitute the stripe structures, the proposed model is employed in this section to simulate bottom reverberation under diverse combinations by setting the types of incident and scattered eigenrays.

Figure 12 depicts the types of incident eigenrays from the source to the seafloor and scattered eigenrays from the seafloor to the receiver. Based on the number of eigenrays contacting the seafloor, the types of incident and scattered eigenrays are labeled with numbers. Specifically, the number 0 is employed to mark D and SR, the number 1 is utilized to mark BSR and the SBSR, and so forth. Since the amplitude of secondary and multiple reflection rays is minor in the two planes, four combinations of incident and scattered eigenrays are accounted for in Figure 12. The combinations in Figure 12a–d are as follows: as: (a) D/SR-D/SR (type "00"); (b) D/SR-BSR/SBSR (type "01"); (c) BSR/SBSR-D/SR (type "10"); and (d) BSR/SBSR-BSR/SBSR (type "11"). At a specific scatterer in the mirror reflection direction, $\alpha_{\text{inc}} = \alpha_{\text{scatt}}$, at which point the combination constitutes

eigenrays between the source and the receiver, corresponding to primary and secondary bottom reflections.

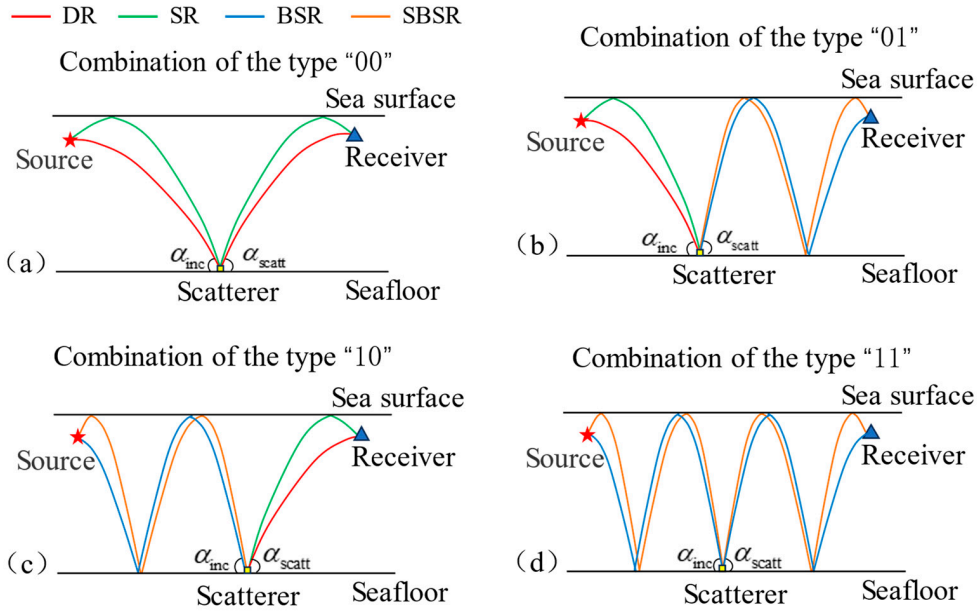


Figure 12. Types of eigenrays from the source to the seafloor scatterer and then to the receiver. (**a**) D/SR-D/SR (type "00"). (**b**) D/SR-BSR/SBSR (type "01"). (**c**) BSR/SBSR-D/SR (type "10"). (**d**) BSR/SBSR-BSR/SBSR (type "11").

Figures 13–15 exhibit reverberation simulations for the four types of combinations at positions A1, B1, and C1, respectively. In the simulation of the three positions, the transmitting beam angles are consistent with those specified in Table 2, which are 0 degrees, 12 degrees, and 0 degrees, respectively. Compared with the outcomes presented in Section 4.2, the four types of combinations exert a significant influence on the spatial-temporal structures of bottom reverberations. The simulations presented in Figures 13–15 demonstrate consistent regularity.

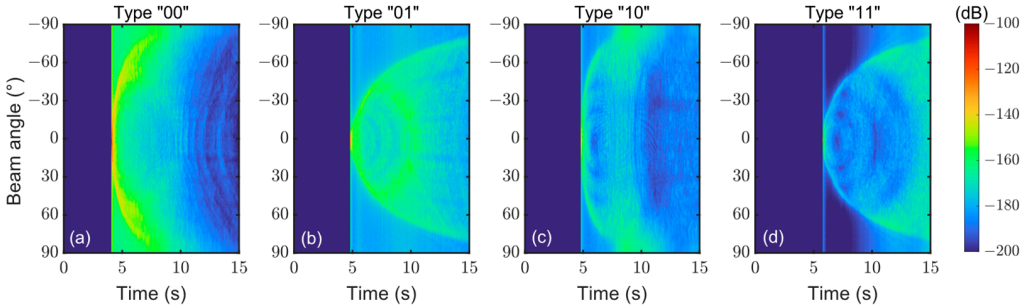


Figure 13. Reverberation simulation outcomes for the combination of the four types at position A1. (**a**) D/SR-D/SR. (**b**) D/SR-BSR/SBSR. (**c**) BSR/SBSR-D/SR. (**d**) BSR/SBSR-BSR/SBSR.

Figures 13–15 show that the "00" type significantly impacts the vertical and convex stripes within the reverberation structures. This type corresponds to the D and SR rays from the source to the seafloor scatterer, as well as those from the seafloor scatterer to the receiver. Similarly, the "01" type predominantly influences the concave or elliptical stripes. Moreover, the "10" and the "11" types exert influences on the above two parts of the stripe structures, respectively. However, the amplitudes of the eigenrays in these two types are lower than those of the initial two types. Figure 14 shows that as the transmitting beam angle is adjusted to 12 degrees, the amplitude of BSR in the incident plane increases,

and the intensity in Figure 14c is higher than in Figure 14a. Accordingly, the vertical and convex stripes are mainly influenced by the "10" type. Nevertheless, similar to the outcomes in Figures 13b and 15b, the "01" type still exerts a predominant influence on the elliptical stripe.

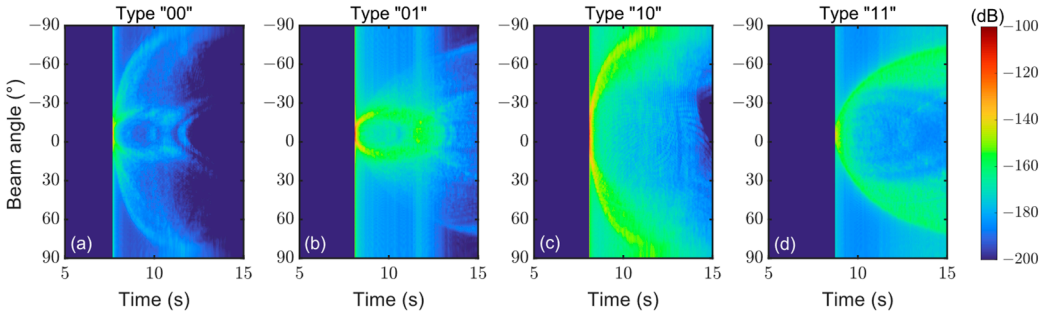


Figure 14. Reverberation simulation outcomes for the combination of the four types at position B1. (**a**) D/SR-D/SR. (**b**) D/SR-BSR/SBSR. (**c**) BSR/SBSR-D/SR. (**d**) BSR/SBSR-BSR/SBSR.

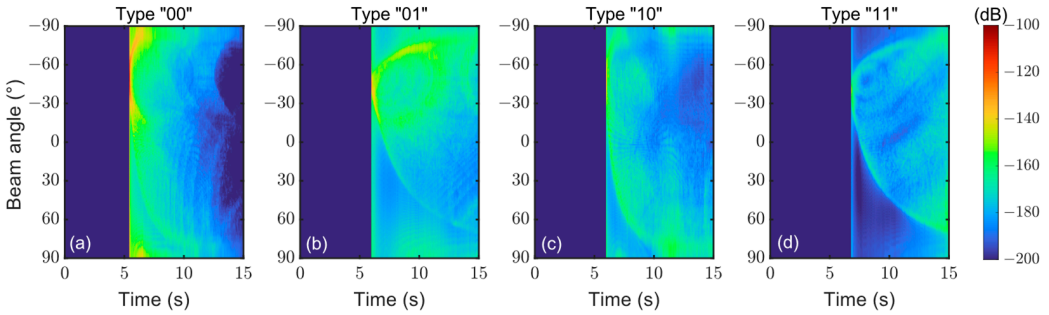


Figure 15. Reverberation simulation outcomes for the combination of the four types at position C1. (**a**) D/SR-D/SR. (**b**) D/SR-BSR/SBSR. (**c**) BSR/SBSR-D/SR. (**d**) BSR/SBSR-BSR/SBSR.

Additionally, the morphological characteristics of the reverberation for the "00" type are analogous to those of the "10" type, while the morphologies of the reverberation corresponding to the "01" type are comparable to those of the "11" type. According to the equation in Section 2 ($\sin\theta'_{i,n} = \sin\theta_i \cos\beta_{Rn}$), the arrival angle $\theta'_{i,n}$ of the nth scattered eigenray is determined by the variables θ_i and β_{Rn}. The former variable corresponds to the angle between the scattering plane and the normal direction of the receiving array, and the latter corresponds to the grazing angle of the scattered eigenray at the receiver. This suggests that, for various combinations of incident and scattered eigenrays, the D and SR rays within the scattered eigenrays primarily affect the convex stripe, while the BSR and the SBSR rays have a more substantial impact on the concave or elliptical stripe.

This comparison of the tracks demonstrates the effectiveness of the model in simulating bottom reverberation. The experimental results and simulations indicate that the vertical stripes within the spatial-temporal structures are formed by the bottom reflected rays in the mirror reflection direction. Furthermore, the curved and elliptical stripes result from the combined effect of the incident and scattered eigenrays from the seafloor scatterers in different areas.

5. Summary and Conclusions

In this paper, a bistatic bottom reverberation model is proposed to model and analyze the spatial-temporal structures of ocean bottom reverberation signals received by a towed horizontal array. Based on ray theory and the SSA scattering model, the model simulates signals in the beams of the array.

Unlike the empirical function scattering model [15,34], the SSA model is applicable to the full span of grazing angles, as the former is merely applicable to small grazing angles. The beam pattern considers the impact of each scatterer on the reverberation intensity, which is more feasible and applicable than dividing the seafloor regions to simulate reverberation at specific receiving beam angles.

The comparison between the simulation and experimental results indicates that the model outputs reproduce the stripe structures of reverberation in the experiment. The stripe structures encompass vertical, curved, and elliptical stripes. Moreover, the bottom reverberation under various combinations is simulated by setting the types of incident and scattered eigenrays. The analysis indicates that the vertical stripes are generated by the bottom reflection eigenrays in the mirror reflection direction. Meanwhile, the convex and concave stripes, as well as the elliptical stripes, are formed through a combination of incident and scattered eigenrays. The convex stripes are mainly influenced by the D and SR rays within the scattered eigenrays, while the concave and elliptical ones are primarily impacted by the BSR and SBSR rays.

The established model has some limitations. Firstly, an $N \times 2D$ ray model is employed to address the varying seabed topography. For environments with abrupt topographic variations such as seamounts, the model does not account for the effects of horizontal refraction [39]. Such changes will affect the eigenray structure, and result in model outputs being inconsistent with the true results. Secondly, the seabed substrate in the actual environment is multi-layered and complex, and the parameters significantly affect the reverberation intensity output of the model. Due to utilizing a single-layer seabed in the model and unknown parameters, certain disparities exist between the model output and the experimental results. Correspondingly, some stripes cannot be observed in the experiment. Considering these problems, we propose the following solutions: For the first case, the three-dimensional sound field model can be employed to address complex topography, providing more accurate calculations such as the propagation delays and amplitudes of the eigenrays. In the second case, accurate environmental parameters are obtained and a multi-layer seabed model can be employed. The extensive prior information enhances the model outputs, ensuring consistency with the experimental results.

There are potential enhancements for future research. For example, considering variations in seabed topography and prevalent oceanic phenomena, it is valuable to improve and extend the model to accurately predict the intensity and spatial-temporal structures of reverberation within complex environments. Furthermore, the parameters of the seabed substrate can be retrieved utilizing the model. In addition, by analyzing the stripe structures presented in both the model output and the measured data, it becomes possible to distinguish the seabed features or potential targets.

Author Contributions: Conceptualization, Z.S. and J.Z.; methodology, Z.S. and J.Z.; software, Z.S.; validation, Z.S., J.Z. and C.S.; formal analysis, Z.S., J.Z. and C.S.; investigation, Z.S.; resources, J.Z. and C.S.; data curation, Z.S.; writing—original draft preparation, Z.S.; writing—review and editing, J.Z. and C.S.; supervision, J.Z. and C.S.; project administration, J.Z. and C.S.; funding acquisition, J.Z. and C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (2016YFC1400104).

Data Availability Statement: The data presented in this study are available from the corresponding author on request.

Acknowledgments: The sea experiment was organized and carried out by Zhejiang University. We sincerely thank all individuals who participated in the acoustic experiment in the South China Sea in August 2020. Their diligent efforts provided reliable experimental data.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

The representation of $R_{ww}(\alpha_{\text{inc},m}, \alpha_{\text{scatt},n}, \varphi_i)$ for Equation (4) in Section 2 varies depending on the type of seafloor substrate utilized. In this study, a standard fluid substrate was employed for the seafloor. The formulas are as follows:

$$R_{ww}(\alpha_{\text{inc},m}, \alpha_{\text{scatt},n}, \varphi_i) = \frac{1}{2}[1 + V_{ww}(\alpha_{\text{inc},m})][1 + V_{ww}(\alpha_{\text{scatt},n})]O_{wp} \tag{A1}$$

$$O_{wp} = \left(1 - \frac{1}{a_\rho}\right)\left(\cos\alpha_{\text{inc},m}\cos\alpha_{\text{scatt},n}\cos\varphi_i - o_{wp}\right) - 1 + \frac{1}{a_{wp}^2 a_\rho} \tag{A2}$$

$$o_{wp} = \frac{\sqrt{1 - a_{wp}^2\cos^2\alpha_{\text{inc},m}}\sqrt{1 - a_{wp}^2\cos^2\alpha_{\text{scatt},n}}}{a_{wp}^2 a_\rho} \tag{A3}$$

$$V_{ww} = \frac{Z_{wp} - 1}{Z_{wp} + 1} \tag{A4}$$

$$Z_{wp} = \frac{a_{wp}a_\rho\sin\alpha}{\sqrt{1 - a_{wp}^2\cos^2\alpha}} \tag{A5}$$

where a_ρ represents the ratio of the density at the bottom to that in the water, and a_{wp} represents the ratio of the speed at the bottom to that in the water. The variable α denotes either $\alpha_{\text{inc},m}$ or $\alpha_{\text{scatt},n}$ depending on the V_{ww} in Equation (A1).

Appendix B

Based on the Bellhop model [38], we simulated the structures, arrival times, and exit grazing angle of eigenrays at the source.

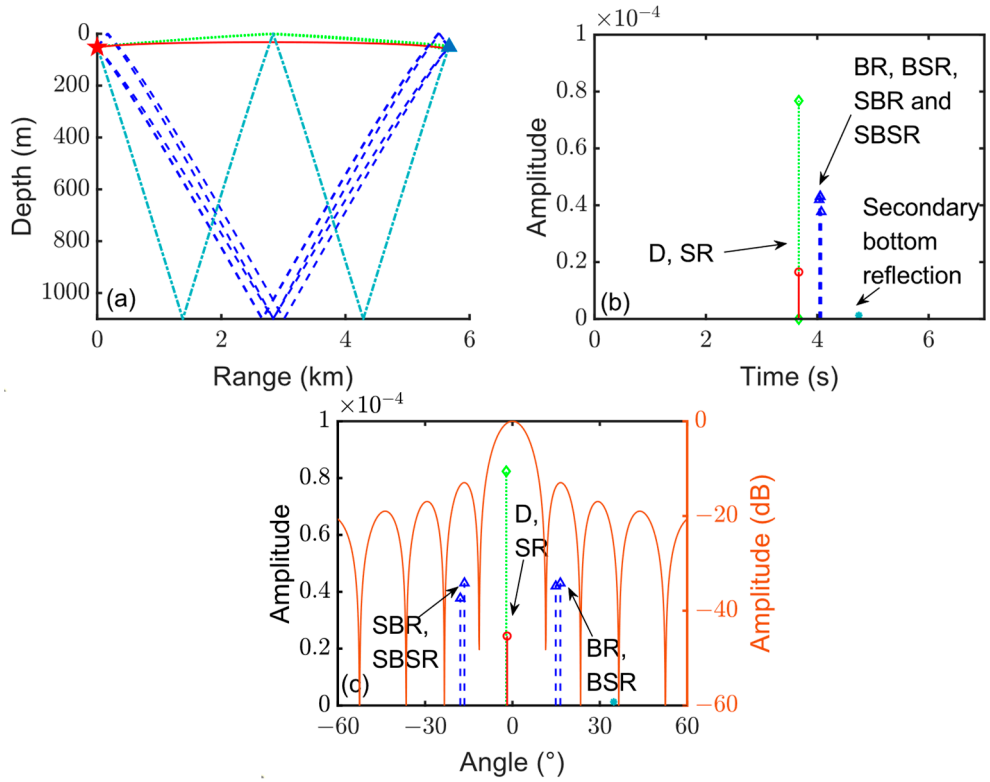


Figure A1. The eigenrays, arrival times, and exit grazing angle at the source at position A1. (**a**) Eigenrays. (**b**) Arrival times. (**c**) Exit grazing angles of eigenrays and the beam pattern of the transmitting array.

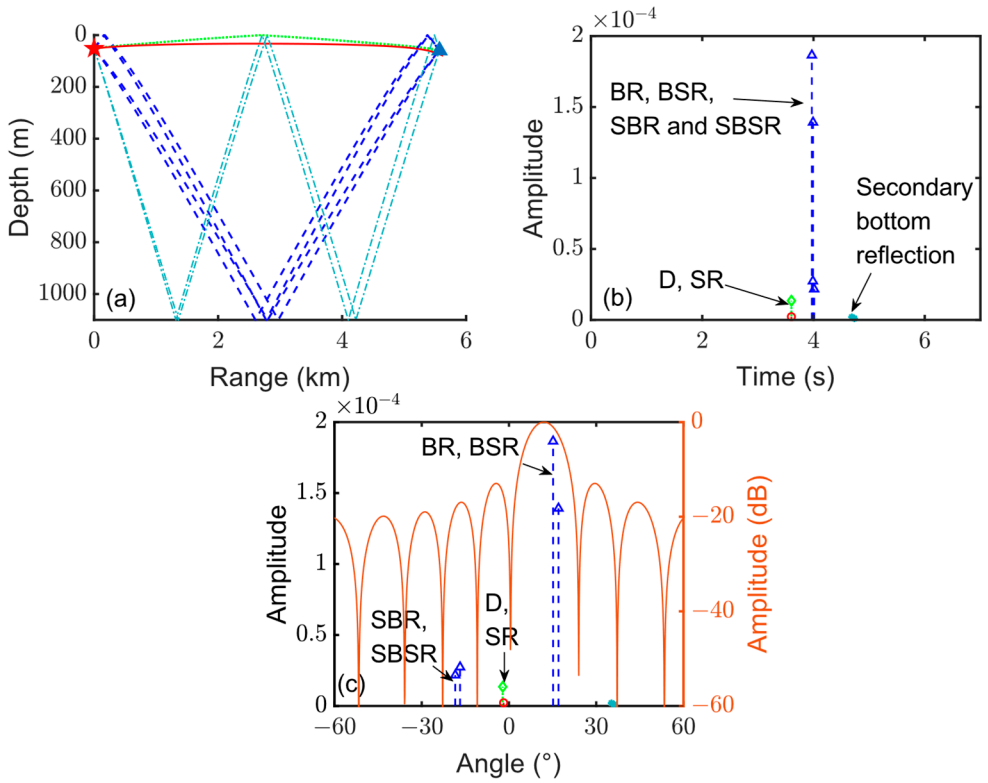


Figure A2. The eigenrays, arrival times, and exit grazing angle at the source at position A2. (**a**) Eigenrays. (**b**) Arrival times. (**c**) Exit grazing angles of eigenrays and the beam pattern of the transmitting array.

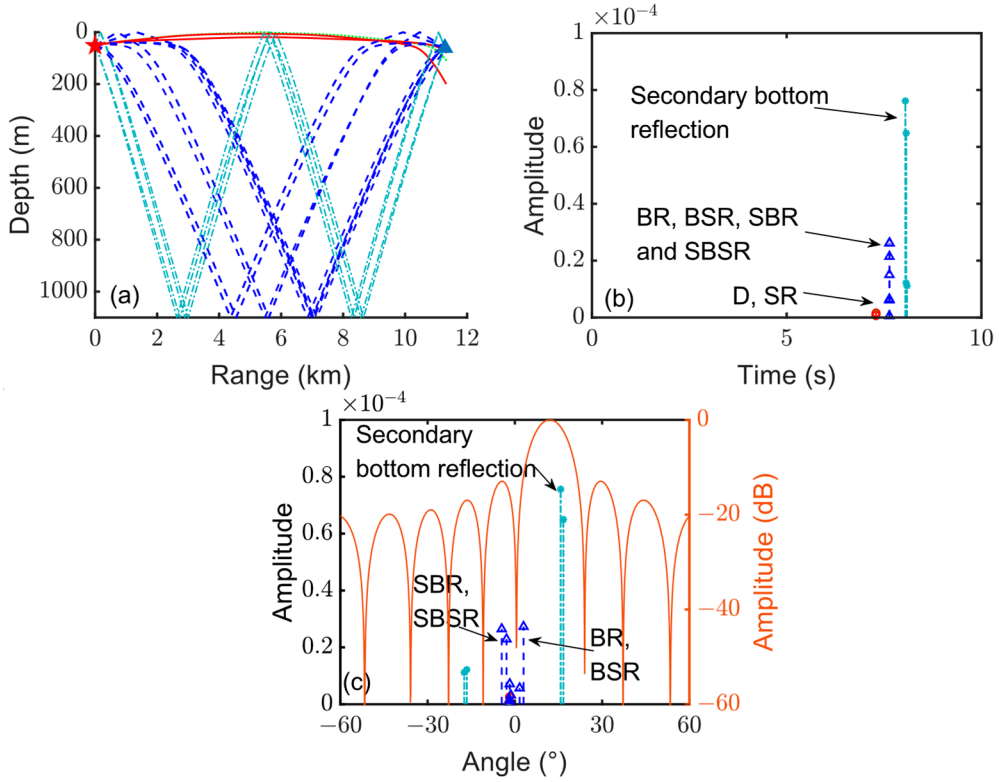


Figure A3. The eigenrays, arrival times, and exit grazing angle at the source at position B1. (**a**) Eigenrays. (**b**) Arrival times. (**c**) Exit grazing angles of eigenrays and the beam pattern of the transmitting array.

Appendix C

Taking positions A1 and B1 as examples, we computed the positions of the scatterers contributing to bottom reverberation and corresponding amplitudes on the grid. The results are shown in Figures A4 and A5, respectively.

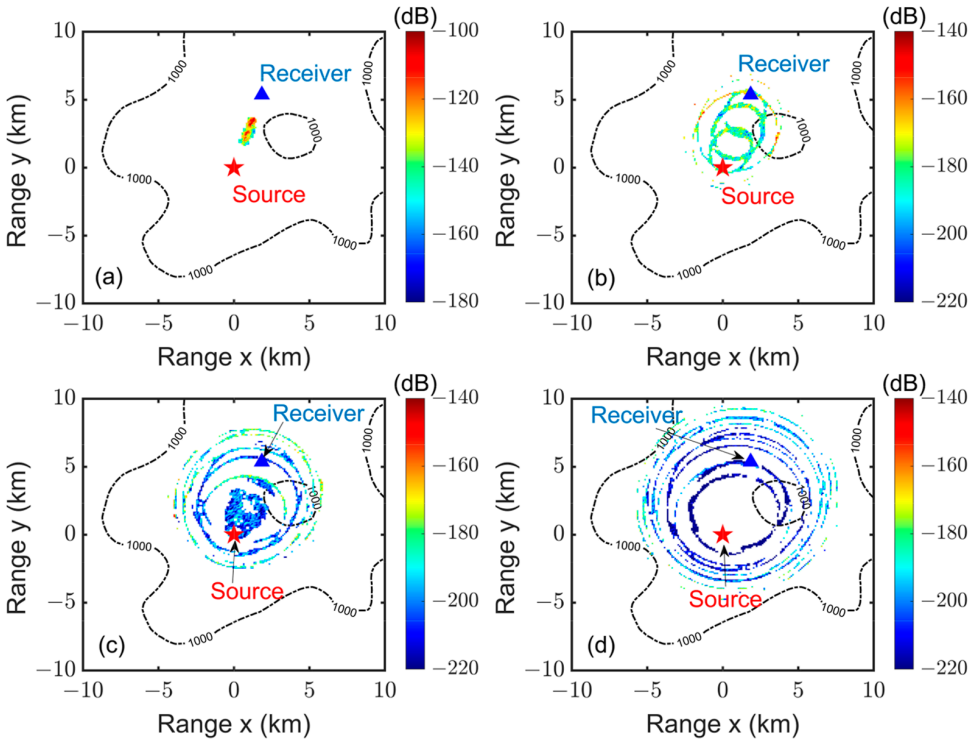


Figure A4. The positions of the scatterers contributing to the reverberation vary over time at position A1. (**a**) t = 4 s. (**b**) t = 6 s. (**c**). t = 8 s. (**d**) t = 10 s.

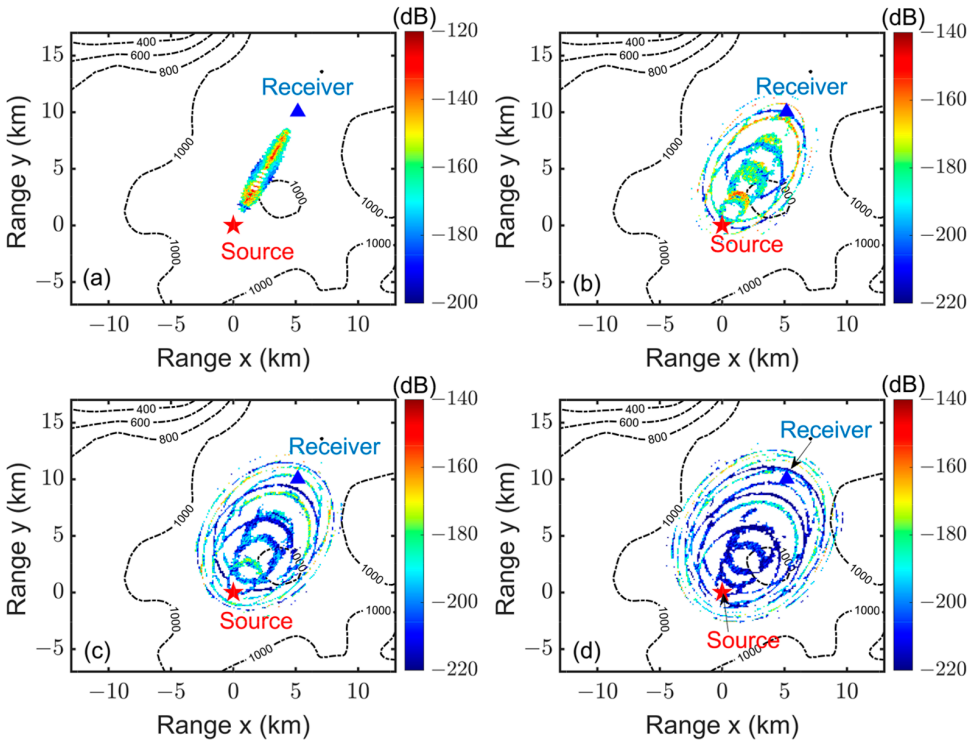


Figure A5. The positions of the scatterers contributing to the reverberation vary over time at position B1. (**a**) t = 7.65 s. (**b**) t = 9.5 s. (**c**). t = 10.5 s. (**d**) t = 11.5 s.

References

1. Zheng, G.; Guo, X.; Zhu, F.; Wang, F.; Bai, L. Analysis of bottom reverberation intensity under beam-controlled emission conditions in deep water. *Arch. Acoust.* **2024**, *49*, 221–231. [CrossRef]
2. Wang, L.; Huang, Y.; Guo, L.; Ren, C. Acoustic scattering modeling and sound field characteristics of rough seafloor in shallow sea. *Acta Phys. Sin.* **2024**, *73*, 034301. [CrossRef]
3. Xu, L.; Yang, K.; Yang, Q. Geoacoustic Inversion Using Physical–Statistical Bottom Reverberation Model in the Deep Ocean. *Acoust. Aust.* **2019**, *47*, 261–269. [CrossRef]
4. Yang, K.; Xu, L.; Yang, Q.; Li, G. Two-Step Inversion of Geoacoustic Parameters with Bottom Reverberation and Transmission Loss in the Deep Ocean. *Acoust. Aust.* **2018**, *46*, 131–142. [CrossRef]
5. Guo, Y.; Duan, R.; Zhang, H.; Yang, K.; Liu, T.; Gong, G. Modeling and interpretation of bistatic bottom reverberation in deep water near seamounts. *J. Acoust. Soc. Am.* **2023**, *154*, 3173–3187. [CrossRef]
6. Ratilal, P.; Lai, Y.; Symonds, D.T.; Ruhlmann, L.A.; Preston, J.R.; Scheer, E.K.; Garr, M.T.; Holland, C.W.; Goff, J.A.; Makris, N.C. Long Range Acoustic Imaging of the Continental Shelf Environment: The Acoustic Clutter Reconnaissance Experiment 2001. *J. Acoust. Soc. Am.* **2005**, *117*, 1977–1998. [CrossRef]
7. Zheng, Y.; Ping, X.; Li, L.; Wang, D. Maximum Likelihood Deconvolution of Beamforming Images with Signal-Dependent Speckle Fluctuations. *Remote Sens.* **2024**, *16*, 1506. [CrossRef]
8. Tang, D.; Jackson, D. Application of small-roughness perturbation theory to reverberation in range-dependent waveguides. *J. Acoust. Soc. Am.* **2012**, *131*, 4428–4441. [CrossRef]
9. Ivakin, A. A full-field perturbation approach to scattering and reverberation in range-dependent environments with rough interfaces. *J. Acoust. Soc. Am.* **2016**, *140*, 657–665. [CrossRef]
10. Tang, D.; Jackson, D. A time-domain model for seafloor scattering. *J. Acoust. Soc. Am.* **2017**, *142*, 2968–2978. [CrossRef]
11. Wu, J.; Gao, T.; Shang, E. Reverberation Intensity Decaying in Range-Dependent Waveguide. *J. Theor. Comput. Acoust.* **2019**, *27*, 1950007. [CrossRef]
12. Bucker, H.; Morris, H. Normal-mode reverberation in channels or ducts. *J. Acoust. Soc. Am.* **1968**, *44*, 827–828. [CrossRef]
13. Ellis, D.D. A shallow-water normal-mode reverberation model. *J. Acoust. Soc. Am.* **1995**, *97*, 2804–2814. [CrossRef]
14. Li, N.; Zhang, M.; Gao, B. Horizontal Correlation of Long-Range Bottom Reverberation in Shallow Sloping Seabed. *J. Mar. Sci. Eng.* **2021**, *9*, 414. [CrossRef]
15. Wang, L.; Qin, J.; Fu, D.; Li, Z.; Liu, J.; Weng, J. Bottom reverberation for large receiving depth in deep water. *Acta Phys. Sin.* **2019**, *68*, 134303. [CrossRef]
16. Guo, R.; Lei, L.; Sheng, Z.; Song, P. A three-dimensional bistatic reverberation model of underwater explosion by interface scattering at variable sound speeds. *J. Low Freq. Noise Vib. Act. Control* **2019**, *38*, 427–440. [CrossRef]
17. Bouchage, G.; Zakharia, M.; LePage, K.D. A Bottom Reverberation Model Based on the Coupling Between Parabolic Equation and Scattering Strengths. *Acta Acust. United Acust.* **2012**, *98*, 216–222. [CrossRef]
18. Lingevitch, J.F.; Fromm, D.M.; LePage, K.D. Parabolic equation technique for reverberation modeling. *J. Acoust. Soc. Am.* **2007**, *122*, 3074. [CrossRef]
19. Makris, N.C.; Avelino, L.Z.; Menis, R. Deterministic Reverberation from Ocean Ridges. *J. Acoust. Soc. Am.* **1995**, *97*, 3547–3574. [CrossRef]
20. Makris, N.C.; Chia, C.S.; Fialkowski, L.T. The Bi-Azimuthal Scattering Distribution of an Abyssal Hill. *J. Acoust. Soc. Am.* **1999**, *106*, 2491–2512. [CrossRef]
21. Chia, C.S.; Makris, N.C.; Fialkowski, L.T. A Comparison of Bistatic Scattering from Two Geologically Distinct Abyssal Hills. *J. Acoust. Soc. Am.* **2000**, *108*, 2053–2070. [CrossRef] [PubMed]
22. Preston, J.R.; Ellis, D.D. Extracting Bottom Information from Towed-Array Reverberation Data: Part I: Measurement Methodology. *J. Mar. Syst.* **2009**, *78*, S359–S371. [CrossRef]
23. Ellis, D.D.; Preston, J.R. Extracting Bottom Information from Towed-Array Reverberation Data. Part II: Extraction Procedure and Modelling Methodology. *J. Mar. Syst.* **2009**, *78*, S372–S381. [CrossRef]
24. Jagannathan, S.; Küsel, E.T.; Ratilal, P.; Makris, N.C. Scattering from extended targets in range-dependent fluctuating ocean-waveguides with clutter from theory and experiments. *J. Acoust. Soc. Am.* **2012**, *132*, 680–693. [CrossRef] [PubMed]
25. Galinde, A.; Donabed, N.; Andrews, M.; Lee, S.; Makris, N.C.; Ratilal, P. Range-dependent waveguide scattering model calibrated for bottom reverberation in a continental shelf environment. *J. Acoust. Soc. Am.* **2008**, *123*, 1270–1281. [CrossRef]
26. Ellis, D.D.; Yang, J.; Preston, J.R.; Pecknold, S. A Normal Mode Reverberation and Target Echo Model to Interpret Towed Array Data in the Target and Reverberation Experiments. *IEEE J. Ocean. Eng.* **2017**, *42*, 344–361. [CrossRef]
27. Yang, J.; Tang, D.; Hefner, B.T.; Williams, K.L.; Preston, J.R. Overview of Midfrequency Reverberation Data Acquired During the Target and Reverberation Experiment 2013. *IEEE J. Ocean. Eng.* **2018**, *43*, 563–585. [CrossRef]
28. Preston, J.R.; Ellis, D.D.; Gauss, R.C. Geoacoustic parameter extraction using reverberation data from the 2000 Boundary Characterization Experiment on the Malta Plateau. *IEEE J. Ocean. Eng.* **2005**, *30*, 709–732. [CrossRef]

29. Prior, M.K. A Scatterer Map for the Malta Plateau. *IEEE J. Ocean. Eng.* **2005**, *30*, 676–690. [CrossRef]
30. Preston, J.R.; Abraham, D.A. Statistical Analysis of Multistatic Echoes from a Shipwreck in the Malta Plateau. *IEEE J. Ocean. Eng.* **2015**, *40*, 643–656. [CrossRef]
31. Holland, C.W.; Preston, J.R.; Abraham, D.A. Long-range acoustic scattering from a shallow water mud-volcano cluster. *J. Acoust. Soc. Am.* **2007**, *122*, 1946–1958. [CrossRef] [PubMed]
32. Jung, Y.; Seong, W.; Lee, K.; Kim, S. A Depth-Bistatic Bottom Reverberation Model and Comparison with Data from an Active Triplet Towed Array Experiment. *Appl. Sci.* **2020**, *10*, 3080. [CrossRef]
33. Jung, Y.; Lee, K. Observation of the Relationship between Ocean Bathymetry and Acoustic Bearing-Time Record Patterns Acquired during a Reverberation Experiment in the Southwestern Continental Margin of the Ulleung Basin, Korea. *J. Mar. Sci. Eng.* **2021**, *9*, 1259. [CrossRef]
34. Ellis, D.D.; Crowe, D.V. Bistatic reverberation calculations using a three-dimensional scattering function. *J. Acoust. Soc. Am.* **1991**, *89*, 2207–2214. [CrossRef]
35. Zhou, C.; Mo, D.; Gao, B. Seamount Clutter Modeling for Long Range Seabed Reverberation in Deep Sea. *J. Phys. Conf. Ser.* **2023**, *2486*, 012099. [CrossRef]
36. Gragg, R.F.; Wurmser, D.; Gauss, R.C. Small-slope scattering from rough elastic ocean floors: General theory and computational algorithm. *J. Acoust. Soc. Am.* **2001**, *110*, 2878–2901. [CrossRef]
37. Neighbors, T.H.; Bradley, D. *Applied Underwater Acoustics (Leif Bjørnø)*, 1st ed.; Elsevier: Amsterdam, The Netherlands, 2017; Chapter 8, pp. 482–486.
38. The Bellhop Model Was Acquired from the Acoustic Toolbox, Version 2020_11_4. Available online: http://oalib.hlsresearch.com/AcousticsToolbox (accessed on 20 November 2024).
39. Li, S.; Li, Z.; Li, W.; Yu, Y. Three-Dimensional Sound Propagation in the South China Sea with the Presence of Seamount. *J. Mar. Sci. Eng.* **2021**, *9*, 1078. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Technical Note

Research on 2-D Direction of Arrival (DOA) Estimation for an L-Shaped Array

Kun Ye [1], Lang Zhou [2], Shaohua Hong [1], Xuebo Zhang [3,*] and Haixin Sun [1]

[1] School of Informatics, Xiamen University, Xiamen 361005, China; yekun@stu.xmu.edu.cn (K.Y.); hongsh@xmu.edu.cn (S.H.); hxsun@xmu.edu.cn (H.S.)
[2] School of Electronic Science and Engineering, Xiamen University, Xiamen 361005, China; zhoulang@stu.xmu.edu.cn
[3] Whale Wave Technology Inc., Kunming 650200, China
* Correspondence: xby_zhang@nwnu.edu.cn

Abstract: Lately, there has been a significant increase in interest in coprime array configurations, as they offer the advantage of generating more extensive array apertures and enhanced degrees of freedom when contrasted with standard linear arrays. This document introduces an innovative two-dimensional direction-of-arrival (2-D DOA) estimation technique founded on the zero-completion principle. In particular, our initial step involves interpolating the synthetic co-array signals to achieve completion, followed by the regeneration of the covariance matrix utilizing the interpolated synthetic signals. Then, the 2-D angle estimation is realized based on the complemented matrix using the auto-pairing method. The computational modeling outcomes indicate that the suggested method demonstrates superior angular discriminability. Moreover, this method excels in estimation accuracy when contrasted with its algorithmic counterparts.

Keywords: 2-D DOA estimation; L-shaped array; virtual array completion; degrees of freedom; matrix reconstruction

Citation: Ye, K.; Zhou, L.; Hong, S.; Zhang, X.; Sun, H. Research on 2-D Direction of Arrival (DOA) Estimation for an L-Shaped Array. *Remote Sens.* **2024**, *16*, 4787. https://doi.org/10.3390/rs16244787

Academic Editor: Andrzej Stateczny

Received: 11 November 2024
Revised: 13 December 2024
Accepted: 20 December 2024
Published: 22 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Two-dimensional (2-D) direction of arrival (DOA) estimation has a wide range of applications in the fields of communication, sonar, hydroacoustic localization, etc. [1–9]. Circular arrays, L-shaped arrays, rectangular arrays, etc., have been proposed in recent decades [10–18]. The L-shaped array configuration boasts superior angular resolution capabilities and is straightforward to construct [19–21], making 2-D direction determination for this type of array the most widely pursued topic in research.

In order to realize 2-D DOA estimation more accurately, many scholars have proposed various 2-D DOA estimation algorithms. In [22] introduces an innovative technique for automatic pairing, which employs the collaborative singular value decomposition technique to handle a pair of cross-correlation matrices, aiming to enhance the accuracy of 2-D DOA estimation. This approach surpasses the traditional CCM-ESPRIT algorithm, particularly in scenarios with a low SNR and minimal angular disparity. In [23], a new sparse L-shaped electromagnetic vector sensor MIMO radar framework for 2-D DOA estimation is proposed. Existing electromagnetic vector sensor—multiple inputs and multiple outputs—radar arrays struggle to strike a good balance between estimation accuracy and computational burden. The sparse L-shaped array topology in this paper has a large sensor spacing, which can effectively reduce the mutual coupling effect and improve the angular resolution. A fast algorithm is proposed in this article to achieve more accurate DOA estimation by combining ambiguous and high-resolution directional cosine estimation with unambiguous but low-resolution estimation. Ultimately, the theoretical derivation is validated by simulation results, showing that the method outperforms existing techniques with better flexibility and accuracy. In [24], a 2-D DOA-assisted framework based on electromagnetic vector

sensors and monostatic multiple-input–multiple-output radar for anonymous unmanned aerial vehicle localization is proposed. The framework considers multipath effects and employs the rearrangement multiple signal classification algorithm, which significantly improves the localization accuracy and robustness for antenna arrays with different geometric configurations. In [14], 2-D DOA estimation using polarized uniform rectangular arrays is proposed to address the challenges of coherent signals under multipath propagation. The rank defect of the source matrix is solved by modeling and rearranging the parallel factors, and the vector cross-product-assisted rotation invariant technique is used to achieve high-accuracy estimation. In [25] presents an improved algorithm for 2-D DOA estimation using electromagnetic vector sensor rectangular arrays under multipath propagation conditions. The method solves the rank deficiency problem by matrix rearrangement and is applicable to single snapshot scenarios, demonstrating superior accuracy and computational efficiency. In [26], a novel parallel synthetic mutual mass array for 2-D DOA and polarization estimation is proposed, aiming to improve the multiparameter estimation performance. By reducing the mutual coupling between antennas and the hardware cost, combined with the compressed sensing method, it demonstrates superior performance compared to existing techniques for target detection and identification in complex electromagnetic environments. In [27] presents a low-computational-complexity 2-D DOA estimation method designed for L-shaped arrays. The technique leverages the symmetrical properties of the array manifold matrix's conjugate to amplify the functional array opening, merging the propagator approach with the ESPRIT technique for spontaneous pairing, thereby eliminating the need for spectral scanning and repetitive cycles. Tests have confirmed that this approach surpasses the performance of traditional algorithms. In [28], a new 2-D DOA estimation method is proposed for non-coherent distributed sources considering gain-phase perturbations in massive multiple-input and multiple-output systems. By establishing a subspace framework and constrained optimization, the authors jointly estimate the nominal DOA, angular expansion, and gain-phase perturbations, providing a solution that is efficient and suitable for real-time applications. In [29] proposed a joint 2-D DOA estimation method for L-shaped arrays based on a compressed perception framework. By introducing a separable observation model, the authors decompose the joint spatial–temporal parameters into three small matrices, which are estimated using L1-paradigm minimization and Bayesian compressed perception algorithms, and numerical simulations verify their effectiveness and superiority. Using sparse L-shaped arrays, in [30] presents a novel technique for 2-D DOA estimation. The technique lowers the computational cost and increases the accuracy of the estimation by utilizing the cross-covariance matrix to create a signal subspace from a combination of a sparse and a uniform linear array. In [31], an innovative scheme for 2-D DOA and polarization detection is introduced for millimeter-wave polarized MIMO arrays, utilizing a compact sampling methodology. The method provides an efficient estimation scheme with low computational complexity and robustness to sensor position errors for a wide range of sensor geometries by reducing the data dimensions, computing the signal subspace, and exploiting the rotational invariance property. In [32] proposed a new method for 2-D DOA estimation based on active hypersurfaces that utilizes non-uniform periodic time modulation to reduce hardware complexity. With amplifier integration, the system achieves high-accuracy (maximum error of 0.31°) multi-source direction estimation for applications such as IoT, wireless communication and radar. In [33] presents a novel polarized smart reflective surface architecture designed for the 2D-DOA estimation of non-line-of-sight signals. The authors develop a normalized vector cross-product estimator that avoids complex data recovery and grid searches, thus improving efficiency. By deploying electromagnetic vector sensor arrays at both the base station and the IRS, the article demonstrates that the method is capable of efficient 2-D-DOA estimation without requiring any a priori knowledge of the base station-IRS channel. In [34] proposed a deep-learning-based 2-D DOA estimation method that uses an L-shaped array to simultaneously estimate the elevation and azimuth angles of a signal. With two classification neural networks, the method performs superiorly in low SNR and highly correlated signal environments,

significantly outperforming existing conventional and deep learning methods. In [35] presents a resource-efficient tensorized deep neural network for the 2-D DOA estimation of uniform rectangular arrays. The trainable parameters are compressed by propagating the covariance tensor to the hidden state tensor via an inverse tucker decomposition. The network significantly reduces the number of parameters, improves training speed, and reduces GPU memory usage while maintaining good estimation accuracy under non-ideal conditions. Simulation results show that the method outperforms traditional matrix-based neural networks and model-based methods in terms of training efficiency and computational performance. In [36] introduced an innovative design of a triple-concurrent compact nested antenna array to enhance the precision in 2-D DOA estimation. This configuration addresses the issue of angular resolution ambiguity through the expansion of the array's effective area and the minimization of element interaction. By integrating the principles of BOMP with the CMT-BCS algorithm, the experimental outcomes indicate a superior performance compared to conventional methodologies.

In order to obtain a more accurate 2-D angle estimation, in this paper, a 2-D DOA algorithm based on interpolation complementation for L-shaped arrays is proposed, which first utilizes larger apertures and higher degrees of freedom. The complementary zeroes are utilized to complement the pairs of matrices. Finally, the DOA estimation of 2-D corners is realized by using the effective array aperture expansion technique. The computational modeling outcomes indicate that the proposed method excels in both the precision of estimation and the quality of angular discrimination. The primary contributions presented in this article can be outlined as follows:

(1) This paper proposes a covariance matrix reconstruction algorithm based on zero completion. The algorithm makes full use of the independent virtual array elements of the L-type array structure based on coprime arrays for DOA estimation.

(2) We utilize the zero-complement reconstructed covariance matrix and then combine the propagator method and ESPRIT method for DOA estimation. The method realizes 2-D angle auto-matching.

(3) Computational modeling results show that the proposed 2-D DOA estimation algorithm performs well in improving angular accuracy. In addition, the algorithm outperforms conventional algorithms in terms of direction-of-arrival estimation performance.

The remainder of this paper is organized as follows: The second part presents the signal model. The third part presents the proposed methodology. The fourth part presents the simulation. The paper concludes with a summary.

2. Signal Model

A uniform linear orthogonal array with a d spacing between two sensors forms an $x-z$ plane L-shaped array configuration, as shown in Figure 1. The e L-shaped coprime array located in the $x-z$ plane consists of two coprime arrays. Each conjugate array S_i has $|S_i| = 2M_i + N_i - 1$ physical sensors and consists of two sparse ULAs. Note that M_i and N_i are mutually prime integers, and the two sparse ULAs contain $2M_i$ and $2N_i$ physical sensors with corresponding spacings of $N_i d$ and $M_i d$, where $d = \lambda/2$ and λ denotes the carrier wavelength.

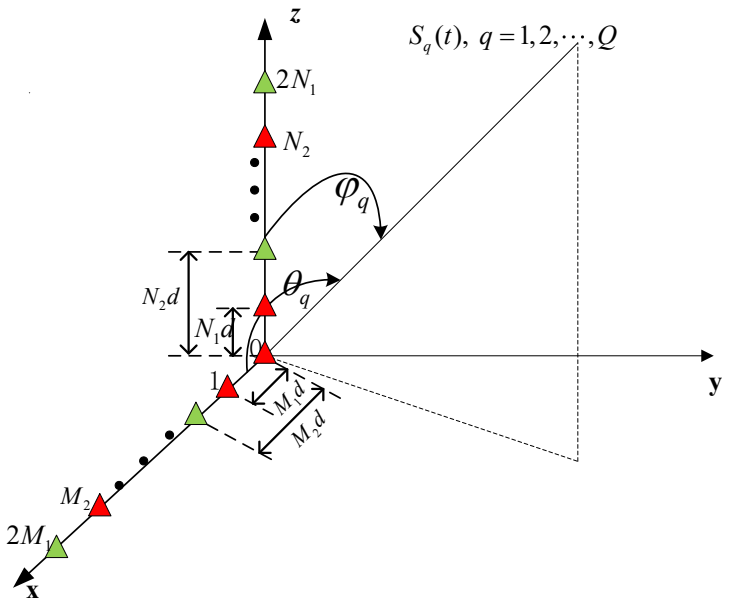


Figure 1. Array structure.

Suppose there are Q unrelated far-field narrowband source signals incident onto an L-shaped array from the $\{(\theta_q, \varphi_q), q \in \langle 1, Q\rangle\}$ directions, where $\theta_q \in [0, \pi]$ and phi $\varphi_q \in [0, \pi]$ represent the azimuth and elevation angles of the qth source signal, respectively. Denote $\mathbf{S}_x(t)$ and $\mathbf{S}_z(t)$ as the information vectors captured by the coprime array $\mathbf{S}_x$ and the coprime array $\mathbf{S}_z$ at the instant t, respectively, and these can be formulated as follows:

$$\mathbf{x}(t) = \mathbf{B}_\mathbf{x}\mathbf{s}(t) + \mathbf{n}_\mathbf{x}(t) \tag{1}$$

$$\mathbf{z}(t) = \mathbf{B}_\mathbf{z}\mathbf{s}(t) + \mathbf{n}_\mathbf{z}(t) \tag{2}$$

where $\mathbf{n}_\mathbf{x}(t)$ and $\mathbf{n}_\mathbf{z}(t)$ are additive Gaussian white noise obeying $\mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I})$,

$$\mathbf{B}_\mathbf{x} = \left[\mathbf{b}_\mathbf{x}(\theta_1), \mathbf{b}_\mathbf{x}(\theta_2), \cdots, \mathbf{b}_\mathbf{x}(\theta_Q)\right], \tag{3}$$

$$\mathbf{b}_\mathbf{x}(\theta_q) = \left[e^{-j\pi\mathbb{S}_x^{(1)}\cos\theta_q}, \cdots, e^{-j\pi\mathbb{S}_x^{(|\mathbb{S}_x|)}\cos\theta_q}\right]^T, \tag{4}$$

$$\mathbf{B}_\mathbf{z} = \left[\mathbf{b}_\mathbf{z}(\varphi_1), \mathbf{b}_\mathbf{z}(\varphi_2), \cdots, \mathbf{b}_\mathbf{z}(\varphi_Q)\right], \tag{5}$$

$$\mathbf{b}_\mathbf{z}(\varphi_q) = \left[e^{-j\pi\mathbb{S}_z^{(1)}\cos\varphi_q}, \cdots, e^{-j\pi\mathbb{S}_z^{(|\mathbb{S}_z|)}\cos\varphi_q}\right]^T. \tag{6}$$

Next, $x(t)$ and $z(t)$ can be represented as

$$\boldsymbol{R}_{xz} = \mathrm{E}\left[\boldsymbol{x}(t)\boldsymbol{z}^{\mathrm{H}}(t)\right] = \mathbf{A}_\mathbf{x}\mathbf{R}_\mathbf{ss}\mathbf{A}_\mathbf{z}^\mathbf{H}, \tag{7}$$

where $\mathbf{R}_\mathbf{ss} = E\left[\mathbf{s}(t)\mathbf{s}^H(t)\right] = diag(\mathbf{p})$, and $\mathbf{p} = [\sigma_1^2, \sigma_2^2, \cdots, \sigma_Q^2]^T$ is the source signal power vector.

Considering that there is no precise covariance matrix R_x, in practice, its sample version can be used to approximate it

$$\hat{\boldsymbol{R}}_{xz} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}(t)\boldsymbol{z}^{\mathrm{H}}(t), \tag{8}$$

where T is the number of snapshots.

3. Proposed Method

In this section, we first use the mutual prime array arrival method of virtual array interpolation to achieve completion based on the virtual common array signal y. Then, the effective array aperture expansion technique is used for 2-D angle estimation.

3.1. Virtual Array Completion

Vectorizing $\boldsymbol{R_{xz}}$ yields the following virtual co-array signal

$$\mathbf{y} = vec(\boldsymbol{R_{xz}}) = (\mathbf{A}_\mathbf{z}^* \odot \mathbf{A}_\mathbf{x})\mathbf{p}. \tag{9}$$

Due to the contamination of model noise affecting the recurring components within the estimated y_v, as detailed subsequently, an aggregation process is employed that involves averaging all duplicate elements [37]. Consequently, the conversion from y_v to y_c is realized through the utilization of the transformation matrix $\mathbf{D}$, that is,

$$\mathbf{y}_{xz} = \mathbf{D}\mathbf{y}, \tag{10}$$

where each entry in the successive line of matrix $\mathbf{D}$ represents the mean multiplier pertaining to the respective element in vector y, with matrix $\mathbf{D}$ adhering to the following condition

$$\mathbf{D}_{i,j} = \frac{\delta_{[\mu_c]_i,[\mu_v]_j}}{\sum_{j=1}^{(2M+N-1)^2} \delta_{[\mu_c]_i,[\mu_v]_{f'}}}, \tag{11}$$

$i = 1, 2, \ldots, 3MN + M - N$, $j = 1, 2, \ldots, (2M + N - 1)^2$, and the expression $\delta_{p,q}$ represents the Kronecker delta. Equation (11) is derived from the principle that the shift from y to y_{xz} is analogous to the shift from μ to μ_{xz}, implying that the matrix $\mathbf{D}$ adheres to the relation $\boldsymbol{\mu} = \mathbf{D}\boldsymbol{\mu}_{xz}$.

The interpolated virtual array signal y can be initialized as

$$[\boldsymbol{y}]_i = \begin{cases} [\tilde{\boldsymbol{y}}_w]_i, & i \in \mathcal{S}_\mathcal{W}, \\ 0, & i \in \mathcal{S}_\mathcal{I} - \mathcal{S}_\mathcal{W}, \end{cases} \tag{12}$$

where $[\cdot]_i$ is the virtual signal of the virtual sensor at position i. $\mathcal{S}_\mathcal{W}$ is the virtual sensor of the difference array and $\mathcal{S}_\mathcal{I}$ is all the virtual sensors containing the holes.

According to the Vandermonde decomposition lemma [38], there exists a variable vector z satisfying

$$\mathbf{T}_z = \sum_{k=1}^{K} p_k \boldsymbol{r}(\theta_k)\boldsymbol{r}^{\mathrm{H}}(\theta_k), \tag{13}$$

where p_k denotes the power of the k-th source signal. $\|\boldsymbol{r}(\theta_k)\|_2 = \sqrt{L}$, where L is the overlapping subarray. We can denote $\mathbf{T}_z$ as

$$[\mathbf{T}_z]^2 = \left[\sum_{k=1}^{K} p_k \boldsymbol{r}(\theta_k)\boldsymbol{r}^{\mathrm{H}}(\theta_k)\right]^2. \tag{14}$$

Since the Toeplitz matrix $\mathbf{T}_z$ is the covariance matrix of the signals received by the reference virtual array, it can be derived from the principal square root of $\mathbf{Y}\mathbf{Y}^{\mathrm{H}}$. We can obtain

$$\tilde{\boldsymbol{R}}_v^2 = \sum_{\ell=1}^{L} \tilde{\boldsymbol{y}}_\ell \tilde{\boldsymbol{y}}_\ell^{\mathrm{H}} = \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^{\mathrm{H}}, \tag{15}$$

where $\tilde{R}_v \in \mathbb{C}^{L\times L}$, $\tilde{\boldsymbol{Y}} = [\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \ldots, \tilde{\boldsymbol{y}}_L]$.

The problem of minimizing the atomic paradigm through [39] of the Toeplitz covariance matrix reconstruction can be expressed as follows:

$$\begin{aligned} \min_{z\in\mathbb{C}^L} \quad & \frac{1}{2}\left\|\mathbf{T}_z \circ \boldsymbol{G} - \tilde{\boldsymbol{R}}_v\right\|_F^2 + \tau\|z\|_{\mathcal{A}_r} \\ \text{subject to} \quad & \mathbf{T}_z \succeq \mathbf{0}. \end{aligned} \tag{16}$$

where $G \in \mathbb{R}^{L\times L}$,$\tau$ is a regularization parameter and $\mathcal{A}_r$ is an atom set.

3.2. Two-Dimensional DOA Estimation

We divide the inter-correlation matrix R_{xz} into two matrices of size $M \times (M-1)$ by virtual array complement optimization to obtain [27]

$$\mathbf{Y}_1 = \mathbf{T}_z(:, 1 : M-1), \tag{17}$$

$$\mathbf{Y}_2 = \mathbf{T}_z(:, 2 : M), \tag{18}$$

after which construct the following two new interrelated submatrices are obtained as follows:

$$\mathbf{Y}_3 = \mathbf{J}_M \mathbf{Y}_2^* \mathbf{J}_{M-1}, \tag{19}$$

$$\mathbf{Y}_4 = \mathbf{J}_M \mathbf{Y}_1^* \mathbf{J}_{M-1}, \tag{20}$$

where $\mathbf{J}_M$ is a Vandermonde matrix. Thus, we obtain the large matrix of received data

$$\mathbf{Y} = [\mathbf{Y}_1^T, \mathbf{Y}_2^T, \mathbf{Y}_3^T, \mathbf{Y}_4^T]^T. \tag{21}$$

The autocorrelation matrix of the signal is obtained by normalizing the product of the signal matrix with its conjugate transpose

$$\mathbf{R_{yy}} = Y * Y^T / T, \tag{22}$$

and the first K columns and columns K+1 to 4M of the matrix $\mathbf{H}$ are extracted to form a new matrix

$$\mathbf{R}\mathrm{yy}_1 = \mathbf{H}(:, 1 : \mathrm{K})^T \tag{23}$$

$$\mathbf{R}\mathrm{yy}_2 = \mathbf{H}(:, \mathrm{K}+1 : 4 * \mathrm{M})^T, \tag{24}$$

where $\mathbf{H} = \mathbf{R}yy^T$.

We define two orientation selection matrices

$$\mathbf{G}_{x,1} = \mathbf{I}_4 \otimes [\mathbf{I}_{M-1}, \mathbf{0}_{(M-1)\times 1}] \tag{25}$$

$$\mathbf{G}_{x,2} = \mathbf{I}_4 \otimes [\mathbf{0}_{(M-1)\times 1}, \mathbf{I}_{M-1}]. \tag{26}$$

After that, we can obtain

$$[Ux, \Phi_x] = eig((Gx_1P_e)^{\dagger}Gx_2P_e), \tag{27}$$

where $\Phi_x = \mathrm{diag}(\Phi_x)^T$.

Define two elevation selection matrices for automatic pairing

$$\mathbf{G}_{z,1} = \mathbf{I}_2 \otimes [\mathbf{I}_M, \mathbf{0}_{M\times M}] \tag{28}$$

$$\mathbf{G}_{z,2} = \mathbf{I}_2 \otimes [\mathbf{0}_{M\times M}, \mathbf{I}_M]. \tag{29}$$

We have access to

$$\Phi_z = (G_{z_2}P_e\hat{A}_{xz_1})^{\dagger}G_{z_1}P_e\hat{A}_{xz_1}, \tag{30}$$

where $\hat{A}_{xz_1} = Ux$, $\Phi_z = \mathrm{diag}(\Phi_z)^T$.

Finally, we can obtain an estimate of the azimuth and pitch angles

$$\hat{\theta}_k = \arccos\left[\mathrm{angle}\left(\tilde{\Phi}_x(k,k)\right)\lambda/(2\pi d)\right] \tag{31}$$

$$\hat{\phi}_k = \arccos\left[\mathrm{angle}\left(\tilde{\Phi}_z(k,k)\right)\lambda/(2\pi d)\right]. \tag{32}$$

4. Numerical Simulations

Initially, the efficacy of the 2-D direction detection is assessed utilizing the root mean square error (RMSE), and its formulation is specified as

$$RMSE = \sqrt{\frac{1}{KQ}\sum_{k=1}^{K}\sum_{q=1}^{Q}\left(\left(\hat{\theta}_q^k - \theta_q\right)^2 + \left(\hat{\varphi}_q^k - \varphi_q\right)^2\right)}, \tag{33}$$

where $K = 500$ represents the number of Monte Carlo trials and $\hat{\theta}_q^k$ and $\hat{\varphi}_q^k$ represent the qth estimated azimuth and elevation angles for the kth trial, respectively. The operating system used is Windows, running on an Intel(R) Core(TM) i7-8700 CPU at 3.20 GHz. The programming language was developed under Matlab 2022 (a).

We evaluated the various algorithms' resolutions in the first simulation. Ten sensors were used, and T = 500 snapshots were taken. The signal-to-noise ratio (SNR) is fixed to 0 dB, and the 10 sources are uniformly distributed in the range [30°–150°], i.e., $(30^\circ, 30^\circ)$, $(43.33^\circ, 43.33^\circ)$, $(56.67^\circ, 56.67^\circ)$, $(70^\circ, 70^\circ)$, $(83.33^\circ, 83.33^\circ)$, $(96.67^\circ, 96.67^\circ)$, $(110^\circ, 110^\circ)$, $(123.33^\circ, 123.33^\circ)$, $(136.67^\circ, 136.67^\circ)$, $(150^\circ, 150^\circ)$.

The outcomes of DOA estimation utilizing the discrete inverse Fourier transform (DFT) [4] technique are depicted in Figure 2a, where two out of ten sources are not recognized and three sources are recognized with large errors. The recognition results of the joint singular value decomposition (JSVD) [22] algorithm are shown in Figure 2b. For the JSVD recognition results, there are five signal sources with large errors. the recognition results of the low-computation-complexity method (LCCM) [27] algorithm are shown in Figure 2c. It can be seen that there are three signal sources with large recognition errors. The recognition angles of the proposed method are shown in Figure 2d, and only one estimated angle has a slight error. It is evident that the suggested approach surpasses current techniques in terms of a more precise estimation of the DOA.

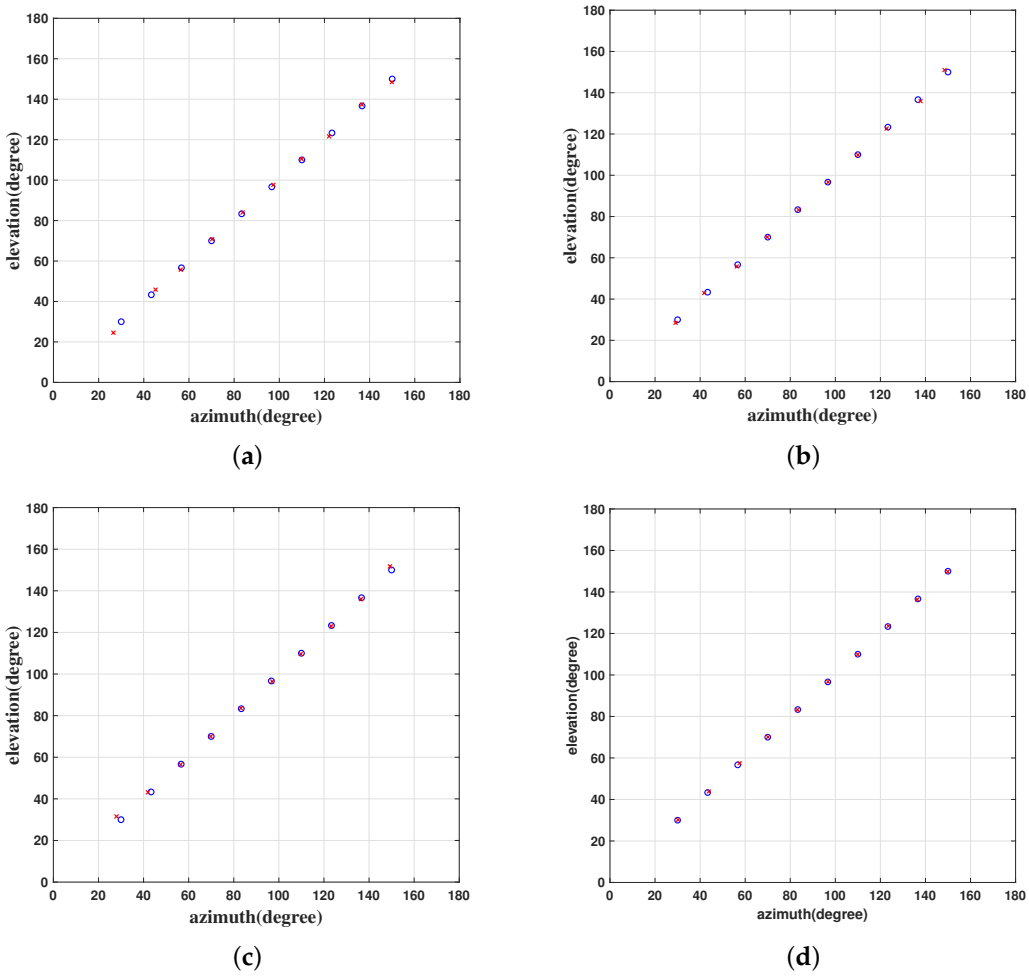


Figure 2. DOA estimation results obtained by different algorithms. (**a**) DFT. (**b**) JSVD. (**c**) LCCM. (**d**) Proposed. (Red crosses indicate estimated results, while blue circles indicate actual values).

In the experiments in the spatial spectrum, we consider three uncorrelated source signals [25°,15°], [45°,35°] and [70°,55°]. The SNR and snapshots are configured as 5 dB and 1000, respectively. Figure 3 shows that the DFT and JSVD algorithms have large sidelobes, and the sidelobes of the proposed algorithm are minimized. In summary, the proposed algorithm achieves better DOA estimation performance.

The DOA performance is compared with the known algorithms, including DFT, JSVD, and LCCM, to validate the DOA accuracy of the proposed algorithm. Assuming 10 physical arrays are used, it is estimated that 10 independent sources are $(30°, 30°)$, $(43.33°, 43.33°)$, $(56.67°, 56.67°)$, $(70°, 70°)$, $(83.33°, 83.33°)$, $(96.67°, 96.67°)$, $(110°, 110°)$, $(123.33°, 123.33°)$, $(136.67°, 136.67°)$, $(150°, 150°)$. A consistent total of 500 snapshots is maintained, while the SNR is adjusted across a range from -10 dB up to 30 dB to assess the system's performance under diverse SNR conditions. In parallel, a variation of sample sizes from 100 to 1000 is tested, with the SNR held steady at 5 dB, to evaluate the system's effectiveness with varying snapshot counts. Here, we also introduce CRB as a reference curve [40].

The illustrations in Figure 4a,b depict that the RMSE for DOA estimation diminishes as the SNR or the count of snapshots rises, confirming the efficacy of the suggested DOA estimation technique. Additionally, the RMSE metrics for the suggested approach are consistently lower compared to those of the DFT, JSVD, and LCCM algorithms in both scenarios, suggesting that the proposed technique outperforms the other three in terms of DOA estimation accuracy across varying SNR levels and snapshot counts.

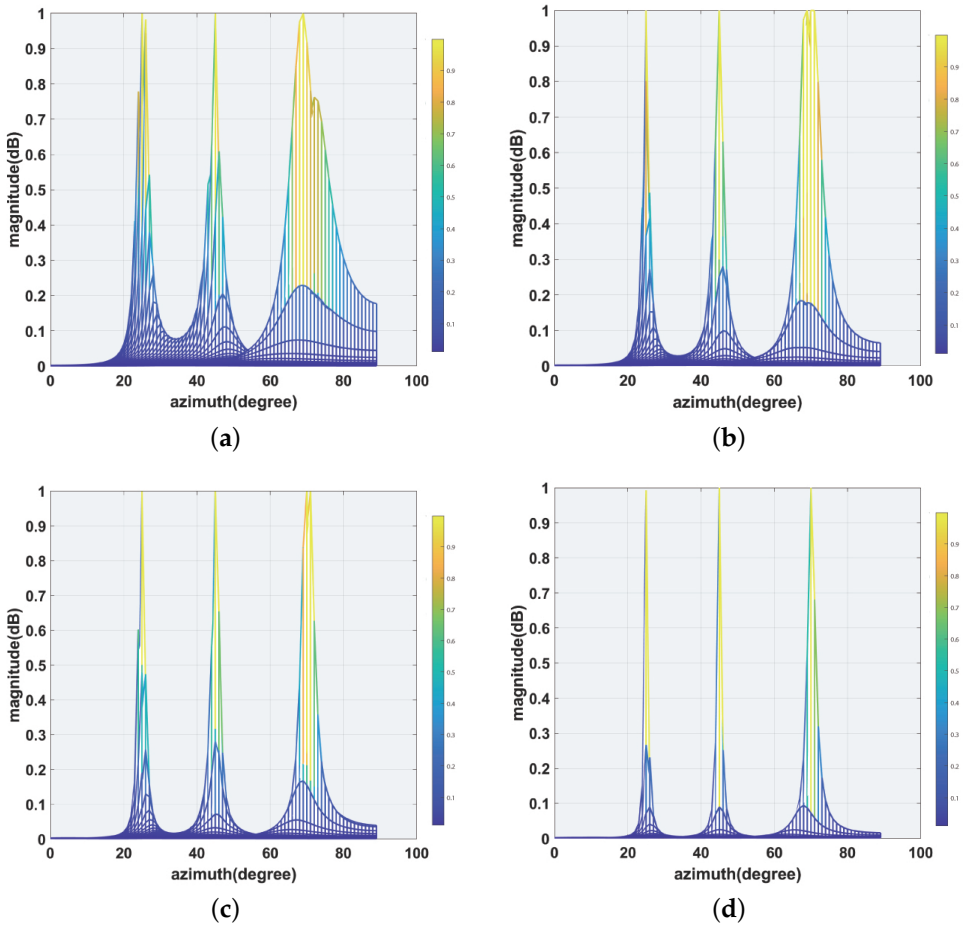


Figure 3. Spatial spectrum of DOA estimation by different algorithms. (**a**) DFT. (**b**) JSVD. (**c**) LCCM. (**d**) Proposed.

In the simulation of the effect of the number of arrays on the algorithm, 8, 10, 12, and 14 physical sensors were used. The parameters were the same as in the previous simulation. As can be seen from Figure 5a,b, in both the SNR and the number of snapshots, a smaller RMSE is obtained with the increase in the number of physical arrays.

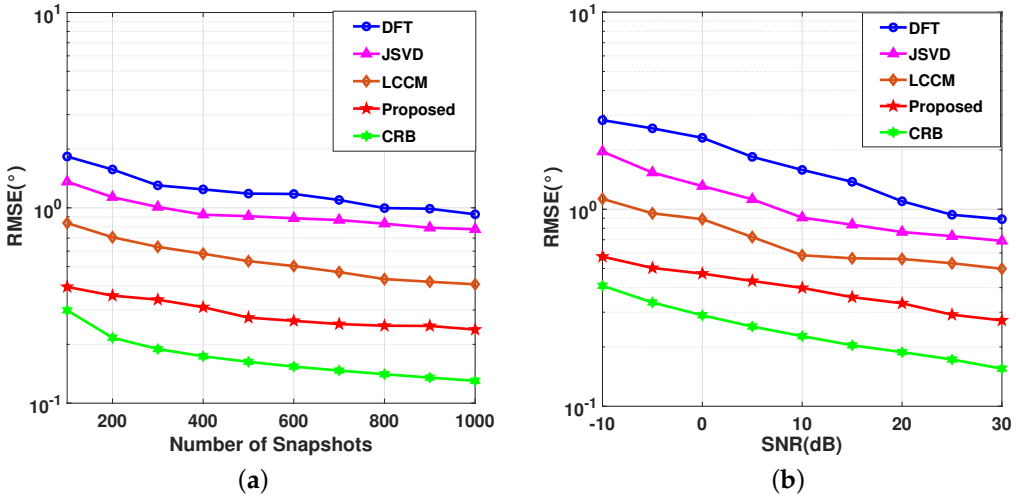


Figure 4. Comparison of the RMSE performance of DFT, JSVD, and LCCM and the proposed (**a**) number of snapshots. (**b**) SNR.

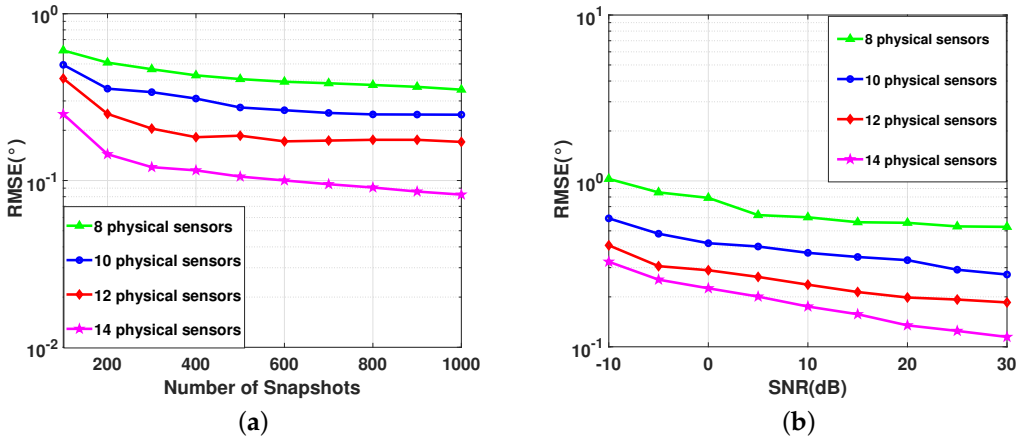


Figure 5. Comparison of RMSE performance of the proposed algorithm under different physical sensors: (**a**) the number of snapshots. (**b**) SNR.

In the RMSE simulation with a CA structure and an ACA structure, for 10 physical sensors, CA takes M = 5, N = 6, while ACA takes M = 3, N = 5. Other simulation parameters are consistent with the above simulation. As can be seen from Figure 6a,b, the RMSE performance of the proposed zero-complement structure is better than the traditional array structure, indicating that the complement structure can increase the number of available virtual arrays and thus obtain better performance.

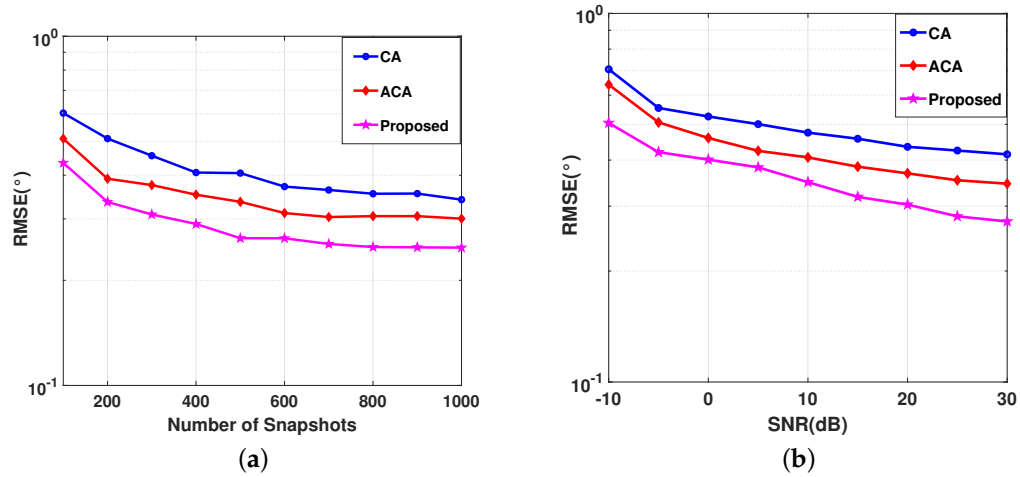


Figure 6. RMSE performance comparison of array structures: (**a**) number of snapshots. (**b**) SNR.

In the RMSE performance experiments at 1/Sqrt(T) and 1/Sqrt(SNR), we consider three uncorrelated source signals, [25°,15°], [45°,35°], and [70°,55°]. The snap count is set to 100 to 800 intervals of 100, and SNR is set to 1 to 20 intervals of 4. Figures 7 and 8 show that the DFT algorithm has the largest RMSE followed by the JSVD algorithm, and the proposed algorithm has the smallest RMSE. The RMSE increases and tends to 1 as 1/Sqrt(T) and 1/Sqrt(SNR) increase, which is in line with the theoretical principles.

Finally, we compared the computation times of various algorithms. In this example, we considered five uncorrelated source signals impacting the array structure. The SNR and snapshots were configured to 5 dB and 500, respectively. Table 1 shows the computation times of various algorithms after executing 1000 Monte Carlo trials. It can be observed that the DFT algorithm has the highest computational complexity, followed by the JSVD algorithm. The proposed algorithm is close to the LCCM. In summary, the proposed algorithm demonstrates the best overall performance.

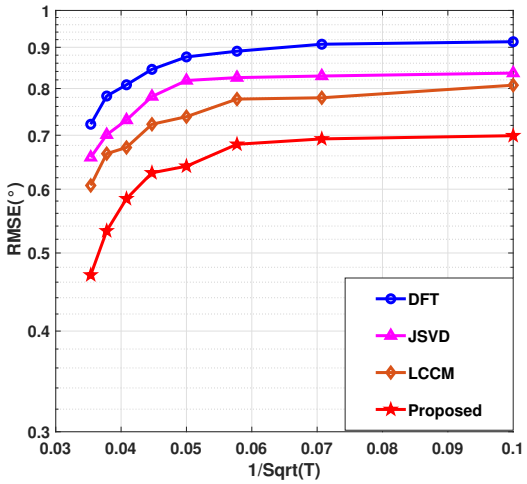


Figure 7. Comparison of RMSE performance at 1/Sqrt(T).

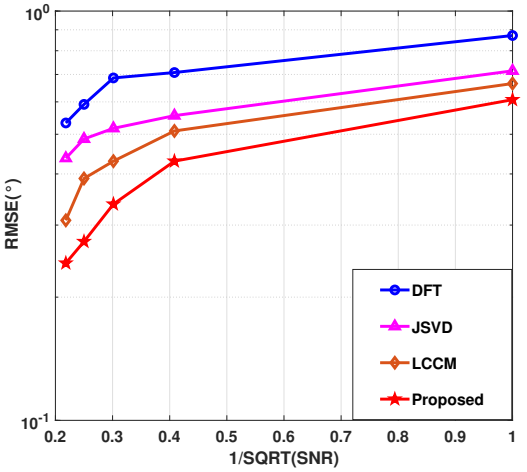


Figure 8. Comparison of RMSE performance at 1/Sqrt(SNR).

Table 1. Processing time for the algorithm.

Algorithm	Processing Time
DFT	1.6308 s
JSVD	0.6292 s
LCCM	0.0339 s
The proposed method	0.0509 s

5. Conclusions

This study presents an innovative 2-D DOA estimation technique based on the zero-complement method. The method spontaneously pairs 2-D angles by reconstructing the covariance matrix. The proposed algorithm is compared with DFT, JSVD, and LCCM algorithms. The results of computational simulations show that the proposed technique has excellent angular resolution capability. Moreover, in terms of DOA estimation accuracy, the proposed technique surpasses the performance of conventional methods.

Author Contributions: Conceptualization, K.Y. and X.Z.; methodology, K.Y. and X.Z.; software, K.Y.; validation, K.Y., S.H. and L.Z.; formal analysis, K.Y.; investigation, K.Y.; resources, K.Y.; data curation, L.Z.; writing—original draft preparation, K.Y.; writing—review and editing, S.H. and L.Z.;

visualization, K.Y.; supervision, H.S.; project administration, L.Z.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Yunnan Xing-dian Yingcai Support Program Industrial Innovative Talents project and the Natural Resources Science and Technology Innovation Project of Fujian KY-080000-04-2022-025.

Conflicts of Interest: Author Xuebo Zhang was employed by the company Whale Wave Technology Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Nomenclature

The symbols used and the meanings attached to them.

A, **a**	Matrices, vectors
I	Identity matrix
$\mathbb{A}$	Set
$(\cdot)^T$ and $(\cdot)^H$	transpose and Hermitian transpose
$\circ$	Hadamard product
$(\cdot)^*$	complex conjugate transpose
$E(\cdot)$	Expectation operator
$vec(\cdot)$	Vectorization operator
$\odot$	Khatri-Rao product
$(\cdot)^\dagger$	pseudoinverse

References

1. Gu, Y.; Goodman, N.A.; Ashok, A. Radar target profiling and recognition based on TSI-optimized compressive sensing kernel. *IEEE Trans. Signal Process.* **2014**, *62*, 3194–3207. [CrossRef]
2. Ye, K.; Zhou, L.; Chen, Z.; Huang, Y.; Hong, S.; Zhang, X.; Sun, H. DOA estimation based on a novel shifted coprime array structure. *AEU-Int. J. Electron. Commun.* **2024**, *179*, 155308. [CrossRef]
3. Zhang, X.; Yang, P.; Wang, Y.; Shen, W.; Yang, J.; Ye, K.; Zhou, M.; Sun, H. LBF-based CS algorithm for multireceiver SAS. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1502505. [CrossRef]
4. Zhang, Z.; Zhou, C.; Gu, Y.; Zhou, J.; Shi, Z. An IDFT approach for coprime array direction-of-arrival estimation. *Digit. Signal Process.* **2019**, *94*, 45–55. [CrossRef]
5. Zhou, C.; Gu, Y.; Shi, Z.; Zhang, Y.D. Off-grid direction-of-arrival estimation using coprime array interpolation. *IEEE Signal Process. Lett.* **2018**, *25*, 1710–1714. [CrossRef]
6. Ye, K.; Wu, S.; Cai, Y.; Zhou, L.; Xiao, L.; Zhang, X.; Zheng, Z.; Lin, J. Transfer-Learning-Based Human Activity Recognition Using Antenna Array. *Remote Sens.* **2024**, *16*, 845. [CrossRef]
7. Zhang, X.; Yang, P.; Wang, Y.; Shen, W.; Yang, J.; Wang, J.; Ye, K.; Zhou, M.; Sun, H. A novel multireceiver sas rd processor. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4203611. [CrossRef]
8. Zhou, L.; Ye, K.; Qi, J.; Sun, H. DOA estimation based on pseudo-noise subspace for relocating enhanced nested array. *IEEE Signal Process. Lett.* **2022**, *29*, 1858–1862. [CrossRef]
9. Ye, K.; Cai, Y.; Hong, S.; Sun, H. Direction-of-arrival estimation based on difference-sum co-array of a special coprime array. *Electron. Lett.* **2023**, *59*, e12701. [CrossRef]
10. Zheng, Z.; Mu, S. 2-D Direction FindingWith Pair-Matching Operation for L-Shaped Nested Array. *IEEE Commun. Lett.* **2020**, *25*, 975–979. [CrossRef]
11. Tayem, N.; Kwon, H.M. L-shape 2-dimensional arrival angle estimation with propagator method. *IEEE Trans. Antennas Propag.* **2005**, *53*, 1622–1630. [CrossRef]
12. Tayem, N.; Majeed, K.; Hussain, A.A. Two-dimensional DOA estimation using cross-correlation matrix with L-shaped array. *IEEE Antennas Wirel. Propag. Lett.* **2015**, *15*, 1077–1080. [CrossRef]
13. Zoltowski, M.D.; Haardt, M.; Mathews, C.P. Closed-form 2-D angle estimation with rectangular arrays in element space or beamspace via unitary ESPRIT. *IEEE Trans. Signal Process.* **1996**, *44*, 316–328. [CrossRef]
14. Zhang, Z.; Wen, F.; Shi, J.; He, J.; Truong, T.K. 2D-DOA Estimation for Coherent Signals via a Polarized Uniform Rectangular Array. *IEEE Signal Process. Lett.* **2023**, *30*, 893–897. [CrossRef]
15. Heidenreich, P.; Zoubir, A.M.; Rubsamen, M. Joint 2-D DOA estimation and phase calibration for uniform rectangular arrays. *IEEE Trans. Signal Process.* **2012**, *60*, 4683–4693. [CrossRef]
16. Zhang, T.; Lu, Y.; Hui, H. Compensation for the mutual coupling effect in uniform circular arrays for 2D DOA estimations employing the maximum likelihood technique. *IEEE Trans. Aerosp. Electron. Syst.* **2008**, *44*, 1215–1221. [CrossRef]

17. Mathews, C.P.; Zoltowski, M.D. Eigenstructure techniques for 2-D angle estimation with uniform circular arrays. *IEEE Trans. Signal Process.* **1994**, *42*, 2395–2407. [CrossRef]
18. Zhang, Y.; Xu, X.; Sheikh, Y.A.; Ye, Z. A rank-reduction based 2-D DOA estimation algorithm for three parallel uniform linear arrays. *Signal Process.* **2016**, *120*, 305–310. [CrossRef]
19. Nie, X.; Li, L. A computationally efficient subspace algorithm for 2-D DOA estimation with L-shaped array. *IEEE Signal Process. Lett.* **2014**, *21*, 971–974.
20. Liang, J.; Liu, D. Joint elevation and azimuth direction finding using L-shaped array. *IEEE Trans. Antennas Propag.* **2010**, *58*, 2136–2141. [CrossRef]
21. Wang, G.; Xin, J.; Zheng, N.; Sano, A. Computationally efficient subspace-based method for two-dimensional direction estimation with L-shaped array. *IEEE Trans. Signal Process.* **2011**, *59*, 3197–3212. [CrossRef]
22. Gu, J.F.; Wei, P. Joint SVD of Two Cross-Correlation Matrices to Achieve Automatic Pairing in 2-D Angle Estimation Problems. *IEEE Antennas Wirel. Propag. Lett.* **2007**, *6*, 553–556. [CrossRef]
23. Wen, F.; Shi, J.; He, J.; Truong, T.K. 2D-DOD and 2D-DOA estimation using sparse L-shaped EMVS-MIMO radar. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *59*, 2077–2084. [CrossRef]
24. Wen, F.; Zhang, Z.; Sun, H.; Gui, G.; Sari, H.; Adachi, F. 2D-DOA estimation auxiliary localization of anonymous UAV using EMVS-MIMO radar. *IEEE Internet Things J.* **2024**, *11*, 16255–16266. [CrossRef]
25. Zhang, Z.; Zhang, L.; Wang, H.; Shi, J. 2D-DOA Estimation in Multipath Using EMVS Rectangle Array. *Remote Sens.* **2023**, *15*, 3308. [CrossRef]
26. Yang, Y.; Shan, M.; Jiang, G. 2D DOA and Polarization Estimation Using Parallel Synthetic Coprime Array of Non-Collocated EMVSs. *Remote Sens.* **2024**, *16*, 3004. [CrossRef]
27. Dong, Y.Y.; Dong, C.X.; Xu, J.; Zhao, G.Q. Computationally efficient 2-D DOA estimation for L-shaped array with automatic pairing. *IEEE Antennas Wirel. Propag. Lett.* **2016**, *15*, 1669–1672. [CrossRef]
28. Tian, Y.; Liu, W.; Xu, H.; Liu, S.; Dong, Z. 2-D DOA estimation of incoherently distributed sources considering gain-phase perturbations in massive MIMO systems. *IEEE Trans. Wirel. Commun.* **2021**, *21*, 1143–1155. [CrossRef]
29. Shixin, W.; Yuan, Z.; Ibrahim, L.; Ying, X.; Jun, W.; Bin, T. Joint 2D DOA and Doppler frequency estimation for L-shaped array using compressive sensing. *J. Syst. Eng. Electron.* **2020**, *31*, 28–36.
30. Gu, J.F.; Zhu, W.P.; Swamy, M. Joint 2-D DOA estimation via sparse L-shaped array. *IEEE Trans. Signal Process.* **2015**, *63*, 1171–1182. [CrossRef]
31. Wen, F.; Gui, G.; Gacanin, H.; Sari, H. Compressive sampling framework for 2D-DOA and polarization estimation in mmWave polarized massive MIMO systems. *IEEE Trans. Wirel. Commun.* **2022**, *22*, 3071–3083. [CrossRef]
32. Xia, D.; Wang, X.; Han, J.; Xue, H.; Liu, G.; Shi, Y.; Li, L.; Cui, T.J. Accurate 2-D DOA estimation based on active metasurface with nonuniformly periodic time modulation. *IEEE Trans. Microw. Theory Tech.* **2022**, *71*, 3424–3435. [CrossRef]
33. Wen, F.; Wang, H.; Gui, G.; Sari, H.; Adachi, F. Polarized intelligent reflecting surface aided 2d-doa estimation for nlos sources. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 8085–8098. [CrossRef]
34. Fadakar, A.; Jafari, A.; Tavana, P.; Jahani, R.; Akhavan, S. Deep learning based 2D-DOA estimation using L-shaped arrays. *J. Frankl. Inst.* **2024**, *361*, 106743. [CrossRef]
35. Zheng, H.; Shi, Z.; Zhou, C.; Vorobyov, S.A.; Gu, Y. Deep Tensor 2-D DOA Estimation for URA. *IEEE Trans. Signal Process.* **2024**, *72*, 4065–4080. [CrossRef]
36. Jiang, G.; Huang, J.; Yang, Y. High-accuracy 2D DOA estimation with three parallel sparse nested array. *AEU-Int. J. Electron. Commun.* **2024**, *179*, 155319. [CrossRef]
37. Pan, Y.; Luo, G.Q. Efficient direction-of-arrival estimation via annihilating-based denoising with coprime array. *Signal Process.* **2021**, *184*, 108061. [CrossRef]
38. Stoica, P.; Moses, R. *Spectral Analysis of Signals*; Prentice-Hall: Upper Saddle River, NJ, USA, 2005.
39. Zhou, C.; Gu, Y.; Fan, X.; Shi, Z.; Mao, G.; Zhang, Y.D. Direction-of-arrival estimation for coprime array via virtual array interpolation. *IEEE Trans. Signal Process.* **2018**, *66*, 5956–5971. [CrossRef]
40. Liu, C.L.; Vaidyanathan, P. Cramér–Rao bounds for coprime and other sparse arrays, which find more sources than sensors. *Digit. Signal Process.* **2017**, *61*, 43–61. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

A Fast Two-Dimensional Direction-of-Arrival Estimator Using Array Manifold Matrix Learning

Jieyi Lu [1,2], Long Yang [1,2,*], Yixin Yang [1,2] and Lu Wang [1,2]

[1] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China; lujieyi@mail.nwpu.edu.cn (J.L.); yxyang@nwpu.edu.cn (Y.Y.); wangluna@nwpu.edu.cn (L.W.)
[2] Shaanxi Key Laboratory of Underwater Information Technology, Xi'an 710072, China
* Correspondence: lyang@nwpu.edu.cn

Abstract: Sparsity-based methods for two-dimensional (2D) direction-of-arrival (DOA) estimation often suffer from high computational complexity due to the large array manifold dictionaries. This paper proposes a fast 2D DOA estimator using array manifold matrix learning, where source-associated grid points are progressively selected from the set of predefined angular grids based on marginal likelihood maximization in the sparse Bayesian learning framework. This grid selection reduces the size of the manifold dictionary matrix, avoiding large-scale matrix inversion and resulting in reduced complexity. To overcome grid mismatch errors, grid optimization is established based on the marginal likelihood, with a dichotomizing-based solver provided that is applicable to arbitrary planar arrays. For uniform rectangular arrays, we present a 2D zoom fast Fourier transform as an alternative to the dichotomizing-based solver by transforming the manifold vector in a specific form, thus accelerating the computation without compromising accuracy. Simulation results verify the superior performance of the proposed methods in terms of estimation accuracy, computational efficiency, and angle resolution compared to state-of-the-art methods for 2D DOA estimation.

Keywords: DOA estimation; array manifold matrix learning; marginal likelihood maximization; Bayesian model; planar arrays; 2D fast Fourier transform

Citation: Lu, J.; Yang, L.; Yang, Y.; Wang, L. A Fast Two-Dimensional Direction-of-Arrival Estimator Using Array Manifold Matrix Learning. *Remote Sens.* **2024**, *16*, 4654. https://doi.org/10.3390/rs16244654

Academic Editors: Xuebo Zhang, Lei Zhang, Jingyue Lu, Haiyong Xu, Marc Pinto and Farman Ali

Received: 10 October 2024
Revised: 1 December 2024
Accepted: 10 December 2024
Published: 12 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Direction-of-arrival (DOA) estimation plays a crucial role in various fields such as sonar, radar, and wireless communications [1]. Classical subspace-based methods, including multiple signal classification (MUSIC) [2] and estimation parameter via rotational invariance technique (ESPRIT) [3], face performance degradation in scenarios with a low signal-to-noise ratio (SNR) or limited snapshots [4]. To address this issue, sparse representation (SR) methods [5–12] have emerged as a promising solution. By exploiting the spatial sparsity of the source signals, SR methods transform the DOA estimation problem into the sparse signal recovery problem. The locations of the peaks of recovered signals indicate the source DOAs.

On-grid SR methods [5–8] employ an overcomplete array manifold dictionary across a set of predefined angular grids in sparse parameter estimation. However, grid mismatch, arising from the discrepancies between source DOAs and predefined grids, can degrade the estimation accuracy of on-grid SR methods. To mitigate this problem, off-grid SR methods [8–12] that integrate a grid refinement process have been proposed. Off-grid SR methods iteratively adjust the grids, enabling DOA estimates to approach the source DOAs, thereby improving the estimation accuracy. One major limitation of many grid-based methods is the high computational complexity due to the large size of the manifold dictionary matrix. Most existing grid-based SR methods focus on the one-dimensional (1D) DOA estimation.

For two-dimensional (2D) DOA estimation, where both azimuth and elevation angles are estimated, the 2D counterparts of 1D grid-based SR methods typically have significantly

increased computational demand due to the larger manifold dictionary. Therefore, fast SR methods with low complexity have become a focal point in 2D DOA estimation. An efficient SR-based combination method, called SAMV-RELAX, is proposed in [13] by utilizing the sparse asymptotic minimum variance (SAMV) method [8] for initial sparse parameter estimation, followed by the RELAX approach [14] for refinement. SAMV-RELAX achieves good computational efficiency by employing coarse 2D angular grids. In contrast, another fast gridless maximum likelihood (FGML) method [15] effectively retrieves source DOAs by processing the reconstructed covariance matrix using the 2D MUSIC or 2D ESPRIT algorithms [16]. By deriving a closed-form formula for covariance matrix reconstruction, FGML exhibits favorable computational efficiency. However, both SAMV-RELAX and FGML methods require knowledge of the source number for parameter estimation, which can limit their performance in practical situations.

A recently proposed method, marginal likelihood maximization-based fast array manifold matrix learning (FAMML) [17], achieves low-complexity 1D DOA estimation within the sparse Bayesian learning (SBL) framework [18,19]. This method incrementally learns source-associated angles and forms a small-size manifold matrix during estimation, avoiding large-scale matrix inversion. Each learned angle is identified from predefined grids and is subsequently optimized to overcome grid mismatch, yielding high estimation accuracy. Given these advantages, the 2D extension of FAMML holds significant research value. However, the extension faces challenges as grid optimization becomes more complicated due to azimuth–elevation coupling.

Inspired by the FAMML method [17], a fast 2D DOA estimator for planar arrays is developed in this study, where predefined 2D angular grids are selected one by one for marginal likelihood maximization, rather than being used simultaneously. This grid selection reduces the size of the manifold dictionary matrix, avoiding the need for large-scale matrix inversion. To achieve a more accurate value of the selected grid point, an efficient dichotomizing-based grid optimization is established, iteratively refining the grid within a bounded area that shrinks with each iteration. This method is applicable to arbitrary planar arrays. For uniform rectangular arrays (URAs), we further express the grid optimization objective function in a 2D fast Fourier transform (FFT) form by exploiting the Kronecker structure of the manifold vector. This enables employing a 2D zoom-FFT [20] for grid optimization, accelerating the computation without compromising estimation accuracy.

The proposed DOA estimator extends the FAMML method in [17] from 1D to 2D DOA estimation, preserving low complexity while achieving high estimation accuracy. The primary contributions of this work are as follows:

(1) The proposal of a 2D angular grid optimization for arbitrary planar arrays, where the selected grid is iteratively refined using a dichotomizing method, improving estimation accuracy.
(2) The proposal of an enhanced-efficiency grid optimization for URAs: The utilization of the 2D zoom-FFT, exploiting the Kronecker structure of the manifold vector, accelerates the computation in grid optimization.

The remainder of this paper is organized as follows. Section 2 presents the signal model for 2D DOA estimation. In Section 3, the proposed 2D fast DOA estimator is introduced in detail. The numerical simulations are conducted in Section 4, and conclusions are drawn in Section 5.

Notation: Bold symbols in small and capital fonts represent the matrices and vectors, respectively. $\mathbb{C}$ stands for the set of complex numbers. $(\cdot)^T$, $(\cdot)^H$, and $(\cdot)^*$ are the transpose, conjugate transpose, and conjugate operators, respectively. $(\cdot)^{-1}$ and $(\cdot)^\dagger$ are the matrix inverse and the Moore–Penrose pseudo-inverse, respectively. $\mathrm{diag}(\cdot)$ denotes the diagonal operator to create a diagonal matrix with a vector being the diagonal elements. $\|\cdot\|$, $\|\cdot\|_2$, and $\|\cdot\|_F$ denote the absolute value, $\updownarrow_2$ norm, and Frobenius norm, respectively. $\otimes$ denotes the Kronecker product. $\mathrm{Re}(\cdot)$ represents the real part of a complex vector. $\mathrm{size}(\cdot)$ represents the matrix size. $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. $\mathrm{vec}^{-1}(\cdot)$ denotes the inverse operator of

vectorization that returns a matrix. $\cos^{-1}(\cdot)$ and $\sin^{-1}(\cdot)$ denote the inverse cosine function and inverse sine function, respectively. The notation ∂ denotes the partial derivative operator. $\max(\cdot)$ represents the maximum value of one or more numbers or variables. $\gg$ is used to denote that one quantity is much greater than another. $\mathcal{CN}(\cdot)$ represents the complex Gaussian distribution. $\mathcal{O}(\cdot)$ is used to describe the order of computational complexity. $\triangleq$ is used to indicate an exact definition or description of a term. The imaginary unit is denoted by $j = \sqrt{-1}$.

2. Signal Model

Consider a planar array composed of M sensors lying in the y–z plane. The position of array sensors in a Cartesian coordinate system is given by $\mathbf{r} = [\mathbf{r}_1, \cdots, \mathbf{r}_M]$, where $\mathbf{r}_m = [0, y_m, z_m]^T$, $m = 1, \cdots, M$. The configuration of the planar array is depicted in Figure 1. Let φ and θ denote the azimuth and elevation angles, respectively. The K narrowband far-field source signals arrival at the planar array from directions $\mathbf{\Theta}_S \triangleq \{\mathbf{\Theta}_{Sk}\}_{k=1}^K = \{(\theta_{Sk}, \varphi_{Sk})\}_{k=1}^K$, where $\mathbf{\Theta}_{Sk} = (\theta_{Sk}, \varphi_{Sk})$ is the kth source signal's DOA. The array manifold vector of the kth signal is:

$$\mathbf{a}(\mathbf{\Theta}_{Sk}) = [\exp(j\omega_{1,k}), \cdots, \exp(j\omega_{M,k})]^T, \tag{1}$$

where $\omega_{m,k} = 2\pi(y_m \sin\theta_{Sk} \sin\varphi_{Sk} + z_m \cos\theta_{Sk})/\lambda$, $m = 1, \cdots, M$, $k = 1, \cdots, K$, and λ is the signal wavelength.

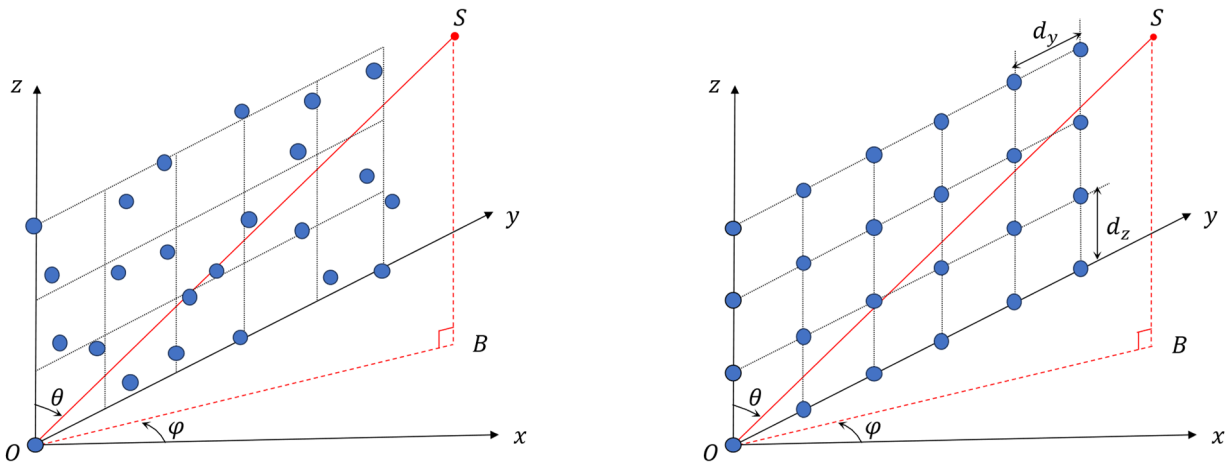


Figure 1. Configuration of the planar array: (**a**) arbitrary planar array; (**b**) URA. Blue dots represent array sensors; red dots represent the source position; B represents the projection of the source position onto the x–y plane.

The array output matrix $\mathbf{Y} = [\mathbf{y}(1), \cdots, \mathbf{y}(L)] \in \mathbb{C}^{M \times L}$ can be modeled as:

$$\mathbf{Y} = \mathbf{A}(\mathbf{\Theta}_S)\mathbf{X} + \mathbf{N}, \tag{2}$$

where L is the number of snapshots, $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(L)] \in \mathbb{C}^{K \times L}$ is the source signal matrix, $\mathbf{N} = [\mathbf{n}(1), \cdots, \mathbf{n}(L)] \in \mathbb{C}^{M \times L}$ is the additional white Gaussian noise matrix, and $\mathbf{A}(\mathbf{\Theta}_S) = [\mathbf{a}(\mathbf{\Theta}_{S1}), \cdots, \mathbf{a}(\mathbf{\Theta}_{SK})] \in \mathbb{C}^{M \times K}$ is the array manifold matrix.

To estimate 2D DOAs using sparse representation, the 2D spatial space is discretized into a set of 2D angular grids. One can define azimuth grids $\{\overline{\varphi}_i\}_{i=1}^{Q_a}$ and elevation grids $\{\overline{\theta}_j\}_{j=1}^{Q_e}$ over the angle range, where Q_a and Q_e are the numbers of grid points. The set of 2D angular grids, containing $Q = Q_a Q_e$ elements, can be expressed as $\mathbf{\Theta}_G = \left\{(\overline{\theta}_1, \overline{\varphi}_1), \cdots, (\overline{\theta}_1, \overline{\varphi}_{Q_a}), (\overline{\theta}_2, \overline{\varphi}_1), \cdots, (\overline{\theta}_2, \overline{\varphi}_{Q_a}), \cdots, (\overline{\theta}_{Q_e}, \overline{\varphi}_1), \cdots, (\overline{\theta}_{Q_e}, \overline{\varphi}_{Q_a})\right\}$. For

convenience, the qth predefined grid is represented by $\mathbf{\Theta}_{Gq} = (\theta_{Gq}, \varphi_{Gq})$, $q = 1, \cdots, Q$. Then, the array output $\mathbf{Y}$ can be sparsely represented over $\mathbf{\Theta}_G$ as:

$$\mathbf{Y} = \mathbf{A}(\mathbf{\Theta}_G)\overline{\mathbf{X}} + \mathbf{N}, \tag{3}$$

where $\mathbf{A}(\mathbf{\Theta}_G) = [\mathbf{a}(\mathbf{\Theta}_{G1}), \cdots, \mathbf{a}(\mathbf{\Theta}_{GQ})] \in \mathbb{C}^{M \times Q}$ is the overcomplete manifold dictionary, and $\overline{\mathbf{X}} = [\overline{\mathbf{x}}(1), \ldots, \overline{\mathbf{x}}(L)] \in \mathbb{C}^{Q \times L}$ is the sparse signal matrix, whose qth row represents a possible signal at $(\theta_{Gq}, \varphi_{Gq})$, $q \in \{1, \cdots, Q\}$. As the number of predefined grids is usually considerably larger than that of source signals, i.e., $Q \gg K$, only a few elements in $\overline{\mathbf{x}}(t)$ are non-zero, and their locations indicate the DOA estimates.

3. The Proposed Fast 2D Estimation Using Array Manifold Matrix Learning

In this section, we address the 2D DOA estimation problem in the SBL framework, using an array manifold matrix learning process to achieve accurate, efficient estimates.

3.1. Bayesian Model

Assume that every column of the sparse signal matrix $\overline{\mathbf{X}}$ adheres to a zero-mean complex Gaussian distribution with precision matrix (inverse covariance matrix) $\mathbf{\Lambda}$, leading to:

$$P(\overline{\mathbf{X}}|\boldsymbol{\alpha}) = \prod_{t=1}^{L} \mathcal{CN}(\overline{\mathbf{x}}(t)|\mathbf{0}, \mathbf{\Lambda}^{-1}), \tag{4}$$

where $\mathbf{\Lambda} = \mathrm{diag}(\boldsymbol{\alpha})$, and $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_Q]^T$ with α_q being the precision parameter (inverse variance) of the qth element in $\overline{\mathbf{x}}(t)$, $q = 1, \cdots, Q$. Obviously, $\alpha_q \to \infty$ when no signal is located on the qth grid point $\mathbf{\Theta}_{Gq}$.

Under the assumption of while Gaussian noise, the likelihood distribution of array output $\mathbf{Y}$ given the signal matrix $\overline{\mathbf{X}}$ can be expressed as follows:

$$P(\mathbf{Y}|\overline{\mathbf{X}}|\mathbf{\Theta}_G, \beta) = \prod_{t=1}^{L} \mathcal{CN}\Big(\mathbf{y}(t)|\mathbf{A}(\mathbf{\Theta}_G)\overline{\mathbf{x}}(t), \beta^{-1}\mathbf{I}_M\Big), \tag{5}$$

where β is the noise precision and $\mathbf{I}_M$ is the identity matrix of order M.

Based on the Bayes theory, the Gaussian posterior of distribution of $\overline{\mathbf{X}}$ can be derived, expressed as:

$$P\big(\overline{\mathbf{X}}|\mathbf{Y}, \boldsymbol{\alpha}, \mathbf{\Theta}_G, \beta\big) = \prod_{t=1}^{L} \mathcal{CN}(\overline{\mathbf{x}}(t)|\boldsymbol{\mu}(t), \mathbf{\Sigma}), \tag{6}$$

where $\boldsymbol{\mu}(t)$ and $\mathbf{\Sigma}$ are the mean vector and covariance matrix, respectively, given by:

$$\begin{cases} \boldsymbol{\mu}(t) = \beta\mathbf{\Sigma}\mathbf{A}^H(\mathbf{\Theta}_G)\mathbf{y}(t), t = 1, \cdots, L \\ \mathbf{\Sigma} = \Big(\mathbf{\Lambda} + \beta\mathbf{A}^H(\mathbf{\Theta}_G)\mathbf{A}(\mathbf{\Theta}_G)\Big)^{-1} \end{cases}. \tag{7}$$

As shown in (7), the computation of the posterior covariance $\mathbf{\Sigma}$ involves a matrix inversion. However, due to the commonly large number of angular grids, the array manifold dictionary $\mathbf{A}(\mathbf{\Theta}_G)$ is high-dimensional with the consequence that the matrix inversion is computationally expensive. Thus, the conventional SBL method often has a high computational complexity in the iterative parameter estimation.

3.2. Parameter Estimation Using Array Manifold Matrix Learning

Inspired by the array manifold matrix learning process in [17], the proposed DOA estimator progressively activates predefined grids and uses them within the parameter estimation process, avoiding the need for large-scale matrix inversion.

3.2.1. Criterion on Active Grid Selection

The proposed DOA estimator activates predefined grids in $\mathbf{\Theta}_G$ one by one during iterations. These activated grids are referred to as "active grids", whose grid indexes compose a set, denoted by $\mathcal{I}_{active}$. The number of active grids is denoted by $\widetilde{K}$. All active grids construct an "active set", denoted by $\mathbf{\Theta}_{active}$ and given by:

$$\mathbf{\Theta}_{active} = \{(\theta_{Gi}, \varphi_{Gi}) : \ i \in \mathcal{I}_{active}\} \triangleq \left\{(\widetilde{\theta}_k, \widetilde{\varphi}_k)\right\}_{k=1}^{\widetilde{K}}, \tag{8}$$

where $(\widetilde{\theta}_k, \widetilde{\varphi}_k)$ represents the kth active grid. The posterior signal mean and covariance, as represented in (7), are computed using $\mathbf{\Theta}_{active}$, i.e.,

$$\begin{cases} \boldsymbol{\mu}(t) = \beta \mathbf{\Sigma} \mathbf{A}^H(\mathbf{\Theta}_{active}) \mathbf{y}(t), t = 1, \cdots, L \\ \mathbf{\Sigma} = \left(\widetilde{\mathbf{\Lambda}} + \beta \mathbf{A}^H(\mathbf{\Theta}_{active}) \mathbf{A}(\mathbf{\Theta}_{active})\right)^{-1} \end{cases}, \tag{9}$$

where $\widetilde{\mathbf{\Lambda}} = \mathrm{diag}(\widetilde{\boldsymbol{\alpha}})$, and $\widetilde{\boldsymbol{\alpha}} = [\widetilde{\alpha}_1, \cdots, \widetilde{\alpha}_{\widetilde{K}}]^T$ is a vector containing signal precisions corresponding to active grids, i.e., $\{\widetilde{\alpha}_k\}_{k=1}^{\widetilde{K}} = \{\alpha_i : i \in \mathcal{I}_{active}\}$.

In each iteration, a predefined grid is selected as a potential addition to $\mathbf{\Theta}_{active}$, pending determination in subsequent processing. This active grid selection is performed by maximizing the increment in log-marginal likelihood from a single grid point.

The log-marginal likelihood function is computed by:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}, \mathbf{\Theta}_{active}, \beta) &= \log p(\overline{\mathbf{X}} | \mathbf{Y}, \boldsymbol{\alpha}, \mathbf{\Theta}_{active}, \beta) \\ &= \log \textstyle\int p(\mathbf{Y} | \overline{\mathbf{X}}, \mathbf{\Theta}_{active}, \beta) p(\overline{\mathbf{X}} | \boldsymbol{\alpha}) d\overline{\mathbf{X}} \\ &= -L\left(M \log \pi + \log|\mathbf{C}| + \frac{1}{L} \sum_{t=1}^{L} \mathbf{y}^H(t) \mathbf{C}^{-1} \mathbf{y}(t)\right), \end{aligned} \tag{10}$$

where $\overline{\mathbf{X}}$ and $\boldsymbol{\alpha}$ contain elements corresponding only to active grids in $\mathbf{\Theta}_{active}$, and $\mathbf{C} = \beta^{-1}\mathbf{I}_M + \mathbf{A}(\mathbf{\Theta}_{active}) \widetilde{\mathbf{\Lambda}}^{-1} \mathbf{A}^H(\mathbf{\Theta}_{active})$.

Given the noise precision β, the contribution to $\mathcal{L}(\boldsymbol{\alpha}, \mathbf{\Theta}_{active}, \beta)$ arising from a single parameter pair $\{\alpha_i, \theta_{Gi}, \varphi_{Gi}\}$, $i \notin \mathcal{I}_{active}$ can be expressed as follows [17,19]:

$$\updownarrow(\alpha_i, \theta_{Gi}, \varphi_{Gi}) = \log \alpha_i - \log(\alpha_i + s_{\theta_{Gi}, \varphi_{Gi}}) + \frac{p_{\theta_{Gi}, \varphi_{Gi}}}{\alpha_i + s_{\theta_{Gi}, \varphi_{Gi}}}, \tag{11}$$

with

$$s_{\theta_{Gi}, \varphi_{Gi}} = \mathbf{a}_i^H \mathbf{C}_{-i}^{-1} \mathbf{a}_i, \tag{12}$$

and

$$p_{\theta_{Gi}, \varphi_{Gi}} = \frac{1}{L} \sum_{t=1}^{L} \left| \mathbf{a}_i^H \mathbf{C}_{-i}^{-1} \mathbf{y}(t) \right|^2, \tag{13}$$

where $\mathbf{a}_i = \mathbf{a}(\theta_{Gi}, \varphi_{Gi})$, $\mathbf{C}_{-i} = \beta^{-1}\mathbf{I}_M + \mathbf{A}\big((\mathbf{\Theta}_{active})_{-i}\big) \widetilde{\mathbf{\Lambda}}_{-i}^{-1} \mathbf{A}^H\big((\mathbf{\Theta}_{active})_{-i}\big)$, and $(\cdot)_{-i}$ represents the bracketed term with the factor related to $(\theta_{Gi}, \varphi_{Gi})$ removed.

For a $(\theta_{Gi}, \varphi_{Gi})$ not currently in $\mathbf{\Theta}_{active}$, i.e., $i \notin \mathcal{I}_{active}$, $\updownarrow(\alpha_i, \theta_{Gi}, \varphi_{Gi})$ measures the increment in log-marginal likelihood when $(\theta_{Gi}, \varphi_{Gi})$ is added to $\mathbf{\Theta}_{active}$. By taking the partial derivative of $\updownarrow(\alpha_i, \theta_{Gi}, \varphi_{Gi})$ with respect to α_i and setting the derivative to zero, the α_i estimate is obtained:

$$\alpha_i = \begin{cases} \frac{s_{\theta_{Gi}, \varphi_{Gi}}^2}{p_{\theta_{Gi}, \varphi_{Gi}} - s_{\theta_{Gi}, \varphi_{Gi}}}, & \text{if } p_{\theta_{Gi}, \varphi_{Gi}} > s_{\theta_{Gi}, \varphi_{Gi}} \\ +\infty, & \textit{otherwise} \end{cases}. \tag{14}$$

By substituting (14) into $\updownarrow(\alpha_i, \theta_{Gi}, \varphi_{Gi})$ and maximizing it with respect to $(\theta_{Gi}, \varphi_{Gi})$, the optimal solution for active grid selection can be obtained as follows [17]:

$$\{\theta_{Gi^{opt}}, \varphi_{Gi^{opt}}\} = \underset{\mathbf{\Theta}_{Gi} \in \mathbf{\Theta}_G, i \notin \mathcal{I}_{active}}{\operatorname{argmax}} f(\mathbf{\Theta}_{Gi}),\ f(\mathbf{\Theta}_{Gi}) = \frac{p_{\theta_{Gi}, \varphi_{Gi}}}{s_{\theta_{Gi}, \varphi_{Gi}}} = \frac{\beta \mathbf{a}_i^H \mathbf{H} \mathbf{Y} \mathbf{Y}^H \mathbf{H}^H \mathbf{a}_i}{L \mathbf{a}_i^H \mathbf{H} \mathbf{a}_i} = \frac{\beta \mathbf{a}_i^H \mathbf{H} \hat{\mathbf{R}}_{\mathbf{Y}} \mathbf{H}^H \mathbf{a}_i}{\mathbf{a}_i^H \mathbf{H} \mathbf{a}_i} \tag{15}$$

where i^{opt} represents the grid index of the optimal solution, $\mathbf{\Theta}_{Gi} = (\theta_{Gi}, \varphi_{Gi})$, $\hat{\mathbf{R}}_{\mathbf{Y}} = \mathbf{Y}\mathbf{Y}^H / L$, and $\mathbf{H} = \mathbf{I}_M - \beta \mathbf{A}(\mathbf{\Theta}_{active}) \mathbf{\Sigma} \mathbf{A}^H(\mathbf{\Theta}_{active})$. Once $\mathbf{\Theta}_{Gi^{opt}} = (\theta_{Gi^{opt}}, \varphi_{Gi^{opt}})$ is determined, the corresponding $\alpha_{i^{opt}}$ can be obtained from (14).

3.2.2. Grid Optimization

The optimal problem (15) is performed on a coarse grid to decrease computational workload. However, it easily causes basis mismatch. To solve this problem, a grid optimization method is proposed.

A conventional approach to optimizing $(\theta_{Gi^{opt}}, \varphi_{Gi^{opt}})$ is to take the partial derivative of the cost function $f(\mathbf{\Theta}_{Gi})$, as expressed in (15), with respect to $\theta_{Gi^{opt}}$ and $\varphi_{Gi^{opt}}$, respectively, and set the results to zero. However, this approach results in a complex system of multivariate equations. Conventional Newton's method requires calculating second partial derivatives to obtain the root, which will bring much additional computational complexity. In contrast, we introduce an efficient dichotomizing method to directly maximize $f(\mathbf{\Theta}_{Gi})$, obtaining a more accurate value of $(\theta_{Gi^{opt}}, \varphi_{Gi^{opt}})$.

First, based on (15), the grid optimization problem of $(\theta_{Gi^{opt}}, \varphi_{Gi^{opt}})$ is expressed as:

$$\{\theta_o, \varphi_o\} = \underset{\mathbf{\Theta}_{Gi^{opt}}}{\operatorname{argmax}} f(\mathbf{\Theta}_{Gi^{opt}}) \tag{16}$$

A dichotomizing method is then presented to solve (16) by iteratively refining $(\theta_{Gi^{opt}}, \varphi_{Gi^{opt}})$ within a bounded area that shrinks with each iteration. More specifically, at each iteration, the range of $(\theta_{Gi^{opt}}, \varphi_{Gi^{opt}})$ in (16) is limited to an area centered around the previous estimate. As the iteration proceeds, this area is continuously zoomed in, and more accurate estimates are obtained within this area. The processing is as follows:

(1) *Initialization:* Set the solution of (15) as the initial estimate for (16), i.e.,

$$(\theta_o^{(0)}, \varphi_o^{(0)}) = (\theta_{Gi^{opt}}, \varphi_{Gi^{opt}}), \tag{17}$$

where the superscript (n) represents the nth iteration. The initial search ranges for elevation and azimuth angles are $[\dot{\theta}_l^{(0)}, \dot{\theta}_r^{(0)}]$ and $[\dot{\varphi}_l^{(0)}, \dot{\varphi}_r^{(0)}]$, respectively, where $\dot{\theta}_l^{(0)} = \left(\theta_{Gi^{opt}} + \theta_{G(i^{opt}-1)}\right)/2$, $\dot{\theta}_r^{(0)} = \left(\theta_{Gi^{opt}} + \theta_{G(i^{opt}+1)}\right)/2$, $\dot{\varphi}_l^{(0)} = \left(\varphi_{Gi^{opt}} + \varphi_{G(i^{opt}-1)}\right)/2$, and $\dot{\varphi}_r^{(0)} = \left(\varphi_{Gi^{opt}} + \varphi_{G(i^{opt}+1)}\right)/2$. Based on this, a candidate grid set is defined below, as shown in Figure 2.

$$\begin{aligned} \mathbf{\Theta}_c^{(0)} = \Big\{ & (\theta_o^{(0)}, \varphi_o^{(0)}), (\theta_o^{(0)}, \dot{\varphi}_l^{(0)}), (\theta_o^{(0)}, \dot{\varphi}_r^{(0)}), (\dot{\theta}_l^{(0)}, \varphi_o^{(0)}), (\dot{\theta}_r^{(0)}, \varphi_o^{(0)}), \\ & (\dot{\theta}_l^{(0)}, \dot{\varphi}_l^{(0)}), (\dot{\theta}_r^{(0)}, \dot{\varphi}_l^{(0)}), (\dot{\theta}_l^{(0)}, \dot{\varphi}_r^{(0)}), (\dot{\theta}_r^{(0)}, \dot{\varphi}_r^{(0)}) \Big\}. \end{aligned} \tag{18}$$

(2) *The nth iteration:* Obtain the updated estimate by identifying the grid point that maximizes the cost function in (16) among $\mathbf{\Theta}_c^{(n-1)}$, as given by:

$$\left\{\theta_o^{(n)}, \varphi_o^{(n)}\right\} = \underset{\mathbf{\Theta}_{Gi^{opt}} \in \mathbf{\Theta}_c^{(n-1)}}{\operatorname{argmax}} f(\mathbf{\Theta}_{Gi^{opt}}). \tag{19}$$

After obtaining $(\theta_o^{(n)}, \varphi_o^{(n)})$, the elevation range and azimuth range are updated. The new elevation bounds are given by:

$$\begin{array}{lll} \text{if } \theta_o^{(n)} = \dot{\theta}_l^{(n-1)}: & \dot{\theta}_l^{(n)} = \dot{\theta}_l^{(n-1)} + \left(\dot{\theta}_l^{(n-1)} - \theta_o^{(n-1)}\right)/2, & \dot{\theta}_r^{(n)} = \left(\dot{\theta}_l^{(n-1)} + \theta_o^{(n-1)}\right)/2 \\ \text{if } \theta_o^{(n)} = \dot{\theta}_o^{(n-1)}: & \dot{\theta}_l^{(n)} = \left(\dot{\theta}_l^{(n-1)} + \theta_o^{(n-1)}\right)/2, & \dot{\theta}_r^{(n)} = \left(\dot{\theta}_r^{(n-1)} + \theta_o^{(n-1)}\right)/2 \\ \text{if } \theta_o^{(n)} = \dot{\theta}_r^{(n-1)}: & \dot{\theta}_l^{(n)} = \left(\dot{\theta}_r^{(n-1)} + \theta_o^{(n-1)}\right)/2, & \dot{\theta}_r^{(n)} = \left(\dot{\theta}_r^{(n-1)} + \theta_o^{(n-1)}\right)/2 \end{array} \tag{20}$$

and the new azimuth bounds are given by:

$$\begin{array}{lll} \text{if } \varphi_o^{(n)} = \dot{\varphi}_l^{(n-1)}: & \dot{\varphi}_l^{(n)} = \dot{\varphi}_l^{(n-1)} + \left(\dot{\varphi}_l^{(n-1)} - \varphi_o^{(n-1)}\right)/2, & \dot{\varphi}_r^{(n)} = \left(\dot{\varphi}_l^{(n-1)} + \varphi_o^{(n-1)}\right)/2 \\ \text{if } \varphi_o^{(n)} = \varphi_o^{(n-1)}: & \dot{\varphi}_l^{(n)} = \left(\dot{\varphi}_l^{(n-1)} + \varphi_o^{(n-1)}\right)/2, & \dot{\varphi}_r^{(n)} = \left(\dot{\varphi}_r^{(n-1)} + \varphi_o^{(n-1)}\right)/2 \\ \text{if } \varphi_o^{(n)} = \dot{\varphi}_r^{(n-1)}: & \dot{\varphi}_l^{(n)} = \left(\dot{\varphi}_r^{(n-1)} + \varphi_o^{(n-1)}\right)/2, & \dot{\varphi}_r^{(n)} = \dot{\varphi}_r^{(n-1)} + \left(\dot{\varphi}_r^{(n-1)} - \varphi_o^{(n-1)}\right)/2 \end{array} \tag{21}$$

Then, a new candidate set $\boldsymbol{\Theta}_c^{(n)}$ is obtained according to (18) for the next iteration.

The number of iterations required by the dichotomizing method depends on the initial angular grid interval (GI) and the desired computational accuracy of the (θ_o, φ_o) estimate. Assuming a GI of 2° is used for grid discretization in both azimuth and elevation ranges, 15 iterations are sufficient to achieve a computational accuracy of $2^\circ/2^{15} < 0.0001^\circ$, which is sufficient for the (θ_o, φ_o) estimate. In the numerical simulations, we set 15 iterations for the dichotomizing method.

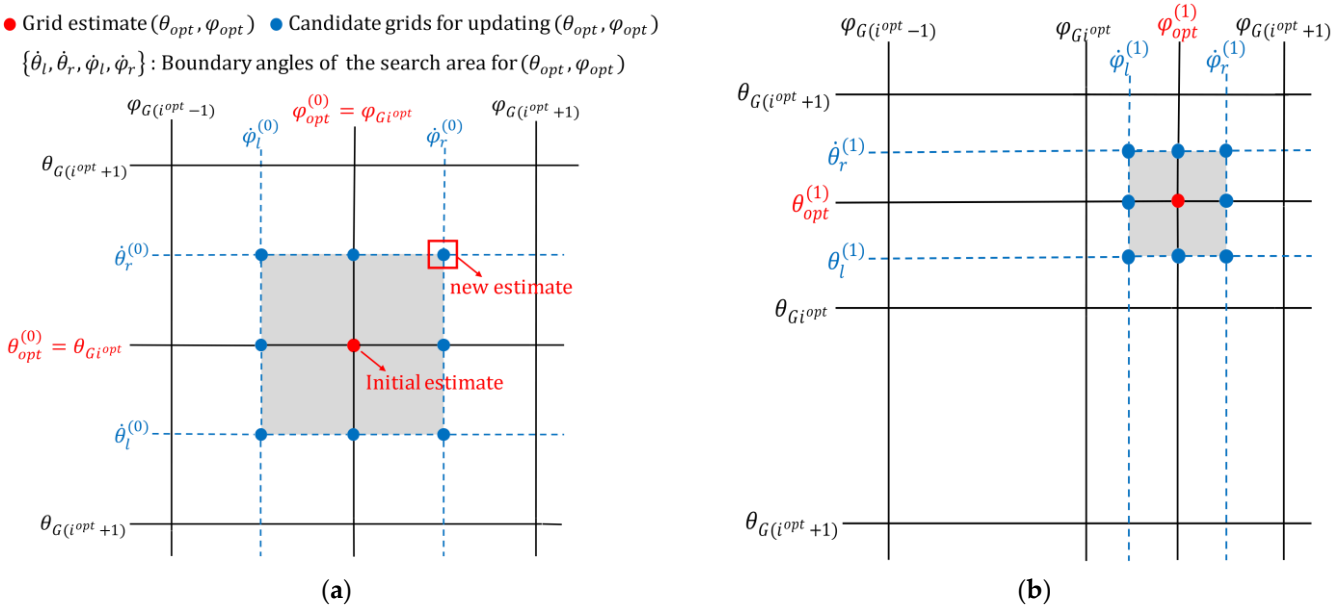


Figure 2. Diagram of the dichotomizing method: (**a**) initialization; (**b**) the first iteration.

The overall process of estimating a potential active grid is outlined in Table 1. In practice, the estimated (θ_o, φ_o) is used to replace the initial predefined grid $(\theta_{Gi^{opt}}, \varphi_{Gi^{opt}})$, when (θ_o, φ_o) is added to the active set $\boldsymbol{\Theta}_{active}$.

Table 1. Active grid selection and optimization processes.

Input: $\mathbf{\Theta}_G$ and $\mathcal{I}_{active}$.
1 Perform active grid selection: Obtain $(\theta_{Gi^{opt}}, \varphi_{Gi^{opt}})$, $i^{opt} \notin \mathcal{I}_{active}$, by solving (15) using $\mathbf{\Theta}_G$. **2 Perform grid optimization using the dichotomizing method:** (1) set $(\theta_o^{(0)}, \varphi_o^{(0)}) = (\theta_{Gi^{opt}}, \varphi_{Gi^{opt}})$, and obtain $\mathbf{\Theta}_c^{(0)}$ according to (18); (2) for n = 1 to 15: update $(\theta_o^{(n)}, \varphi_o^{(n)})$ by solving (19) using $\mathbf{\Theta}_c^{(n-1)}$; update $\mathbf{\Theta}_c^{(n)}$ using (18), (20), and (21). end
Output: $(\theta_o^{(n)}, \varphi_o^{(n)})$ and i^{opt}.

3.2.3. Estimation of Noise Precision β

The previous process assumes that the noise precision β is known, which may not be true in many practical situations. Typically, β can be estimated by differentiating the log-marginal likelihood function $\mathcal{L}(\boldsymbol{\alpha}, \mathbf{\Theta}_{active}, \beta)$ with respect to β and equating the derivative to zero [18,19,21]. Here, we use the maximum likelihood noise update rule [22,23] to obtain a more accurate β estimate:

$$\beta = \frac{M - \widetilde{K}}{\mathrm{tr}\left[\left(\mathbf{I}_M - \mathbf{A}(\mathbf{\Theta}_{active})\mathbf{A}^{\dagger}(\mathbf{\Theta}_{active})\right)\hat{\mathbf{R}}_{\mathbf{Y}}\right]}, \tag{22}$$

where $\hat{\mathbf{R}}_{\mathbf{Y}} = \mathbf{Y}\mathbf{Y}^H/L$, and $\widetilde{K}$ is the number of active grids.

3.2.4. The Iterative Procedure of Array Manifold Matrix Learning

This subsection details the iterative procedure of the array manifold matrix learning-based parameter estimation. In this procedure, signal parameters (i.e., the active set, the signal precision) and the noise precision are updated alternatingly and iteratively. The fundamental content of the array manifold matrix learning process is provided below.

Remember, the grid indexes of active grids are collected in $\mathcal{I}_{active}$; the active set is $\mathbf{\Theta}_{active} = \left\{(\widetilde{\theta}_k, \widetilde{\varphi}_k)\right\}_{k=1}^{\widetilde{K}}$, as expressed in (8); the corresponding signal precisions are $\{\widetilde{\alpha}_k\}_{k=1}^{\widetilde{K}}$; and $(\cdot)_{-k}$ represents the bracketed term with the factor related to $(\widetilde{\theta}_k, \widetilde{\varphi}_k)$ removed.

(1) *Initialization:* The noise precision β is set to 10^3. The active set $\mathbf{\Theta}_{active}$ starts empty. From this, the matrix $\mathbf{C} = \beta^{-1}\mathbf{I}_M$ in (10). Following Table 1, a predefined grid of index $i^{opt(0)}$ is selected, optimized to $(\theta_o^{(0)}, \varphi_o^{(0)})$, and serves as the initial active set, i.e., $\mathbf{\Theta}_{active}^{(0)} = (\theta_o^{(0)}, \varphi_o^{(0)})$. In other words, the first element in the active set is $(\widetilde{\theta}_1^{(0)}, \widetilde{\varphi}_1^{(0)}) = (\theta_o^{(0)}, \varphi_o^{(0)})$, and $\widetilde{K} = 1$. It is worth mentioning that the superscript (n) in this subsection means the nth iteration. The identified active grid index $i^{opt(0)}$ is added to the set $\mathcal{I}_{active}$. The corresponding signal precision $\alpha_{i^{opt(0)}}$ is obtained from (14), yielding the initial precision vector $\widetilde{\boldsymbol{\alpha}}^{(0)} = \alpha_{i^{opt(0)}}$. The posterior signal mean $\left\{\boldsymbol{\mu}^{(0)}(t)\right\}_{t=1}^{L}$ and covariance $\mathbf{\Sigma}^{(0)}$ are then computed using (9). After this, $\beta^{(0)}$ is estimated using (22).

(2) *The nth iteration mainly involves six steps:*

Step 1: Determine a new $(\theta_o^{(n)}, \varphi_o^{(n)})$ outside of $\mathbf{\Theta}_{active}^{(n-1)}$.

A new $(\theta_o^{(n)}, \varphi_o^{(n)})$ of grid index $i^{opt(n)}$ is estimated following Table 1, and the corresponding precision $\alpha_{i^{opt(n)}}$ is computed using (14).

Step 2: Evaluate the log-marginal likelihood increments associated with the new $(\theta_o^{(n)}, \varphi_o^{(n)})$ and the active grids in $\mathbf{\Theta}_{active}^{(n-1)}$, respectively.

For the new $(\theta_o^{(n)}, \varphi_o^{(n)})$, the log-marginal likelihood increment is computed by:

$$\Delta\updownarrow_o^{(n)} = \updownarrow(\alpha_{i^{opt(n)}}, \theta_o^{(n)}, \varphi_o^{(n)}) \tag{23}$$

For each active grid in $\boldsymbol{\Theta}_{active}^{(n-1)}$, update parameters $p^{(n)}_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}}$ and $s^{(n)}_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}}$ by [17]:

$$\begin{cases} p^{(n)}_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}} = \dfrac{\sum_{t=1}^{L}\left|\boldsymbol{\mu}_k^{(n-1)}(t)\right|^2}{L\left(\boldsymbol{\Sigma}_{k,k}^{(n-1)}\right)^2} \\ s^{(n)}_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}} = \dfrac{1}{\boldsymbol{\Sigma}_{k,k}^{(n-1)}} - \widetilde{\alpha}_k^{(n-1)} \end{cases}, \; k = 1, \cdots, \widetilde{K} \tag{24}$$

where $(\widetilde{\theta}_k^{(n-1)}, \widetilde{\varphi}_k^{(n-1)})$ is the kth active grid, $\widetilde{\alpha}_k^{(n-1)}$ is the corresponding precision, $\boldsymbol{\mu}_k^{(n-1)}(t)$ is the kth element in $\boldsymbol{\mu}^{(n-1)}(t)$, and $\boldsymbol{\Sigma}_{k,k}^{(n-1)}$ is the (k,k)th element in $\boldsymbol{\Sigma}^{(n-1)}$. With the updated $p^{(n)}_{\widetilde{\theta}_k,\widetilde{\varphi}_k}$ and $s^{(n)}_{\widetilde{\theta}_k,\widetilde{\varphi}_k}$, the new precision parameter can be obtained using (14), recorded as $\widetilde{\alpha}_k^{new}$. The log-marginal likelihood increment caused by $\widetilde{\alpha}_k^{new}$ can be computed as:

$$\Delta\updownarrow_k^{(n)} = \begin{cases} \Delta\updownarrow_k^{\prime(n)} = -\updownarrow(\widetilde{\alpha}_k^{(n-1)}, \widetilde{\theta}_k^{(n-1)}, \widetilde{\varphi}_k^{(n-1)}), & \text{if } p^{(n)}_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}} \le s^{(n)}_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}} \\ \Delta\updownarrow_r^{\prime\prime(n)} \begin{array}{l} = \updownarrow(\widetilde{\alpha}_k^{new}, \widetilde{\theta}_k^{(n-1)}, \widetilde{\varphi}_k^{(n-1)}) \\ \quad -\updownarrow(\widetilde{\alpha}_k^{(n-1)}, \widetilde{\theta}_k^{(n-1)}, \widetilde{\varphi}_k^{(n-1)}), \end{array} & \text{if } p^{(n)}_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}} > s^{(n)}_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}} \end{cases} \tag{25}$$

Step 3: Update $\boldsymbol{\Theta}_{active}^{(n)}$ with the grid bringing the maximum log-marginal likelihood increment.

Compare the log-marginal likelihood increments $\Delta\updownarrow_o^{(n)}$ and $\left\{\Delta\updownarrow_k^{(n)}\right\}_{k=1}^{\widetilde{K}}$. If the maximum one is $\Delta\updownarrow_o^{(n)}$, then $\widetilde{K} = \widetilde{K} + 1$, and execute the addition operation. Specifically, the estimated $(\theta_o^{(n)}, \varphi_o^{(n)})$ is added to $\boldsymbol{\Theta}_{active}^{(n-1)}$ and becomes the $\widetilde{K}$th active grid, yielding the new $\boldsymbol{\Theta}_{active}^{(n)} = \left\{\boldsymbol{\Theta}_{active}^{(n-1)}, (\theta_o^{(n)}, \varphi_o^{(n)})\right\}$. Similarly, the precision vector is updated as $\widetilde{\boldsymbol{\alpha}}^{(n)} = \left[(\widetilde{\boldsymbol{\alpha}}^{(n-1)})^T, \alpha_{i^{opt(n)}}\right]^T$. The set of active grid indexes $\mathcal{I}_{active}$ is updated by adding $i^{opt(n)}$ to it.

If the maximum log-marginal likelihood increment is $\Delta\updownarrow_k^{(n)} = \Delta\updownarrow_k^{\prime\prime(n)}$, then $\widetilde{K} = \widetilde{K} - 1$, and execute the deletion operation. Specifically, the corresponding $(\widetilde{\theta}_k^{(n-1)}, \widetilde{\varphi}_k^{(n-1)})$ and $\widetilde{\alpha}_k^{(n-1)}$ are deleted from $\boldsymbol{\Theta}_{ative}^{(n-1)}$ and $\widetilde{\boldsymbol{\alpha}}^{(n-1)}$, respectively, yielding the new $\boldsymbol{\Theta}_{active}^{(n)} = (\boldsymbol{\Theta}_{active}^{(n-1)})_{-k}$ and $\widetilde{\boldsymbol{\alpha}}^{(n)} = (\widetilde{\boldsymbol{\alpha}}^{(n-1)})_{-k}$. Then, the grid index of $(\widetilde{\theta}_k^{(n-1)}, \widetilde{\varphi}_k^{(n-1)})$ is removed from $\mathcal{I}_{active}$.

If the maximum log-marginal likelihood increment is $\Delta\updownarrow_k^{(n)} = \Delta\updownarrow_k^{\prime(n)}$, the corresponding $\widetilde{\alpha}_k^{(n-1)}$ is replaced by the re-estimated $\widetilde{\alpha}_k^{new}$ without any grid change, i.e., $\boldsymbol{\Theta}_{active}^{(n)} = \boldsymbol{\Theta}_{active}^{(n-1)}$, and $\widetilde{\boldsymbol{\alpha}}^{(n)} = \widetilde{\boldsymbol{\alpha}}^{(n-1)}$ with $\widetilde{\alpha}_k^{(n-1)} = \widetilde{\alpha}_k^{new}$.

Step 4: Sequentially update the active grids in the new $\boldsymbol{\Theta}_{active}^{(n)}$ to improve the estimation accuracy.

Let $K' = \widetilde{K}$ record the total number of active grids to be updated, as $\widetilde{K}$ may change during this processing. The re-estimated value of the kth active grid $(\widetilde{\theta}_k^{(n)}, \widetilde{\varphi}_k^{(n)})$, denoted by $(\widetilde{\theta}_k^{new}, \widetilde{\varphi}_k^{new})$, $k = 1, \cdots, K'$, is obtained though the grid optimization in Table 1. In this process, $(\widetilde{\theta}_k^{(n)}, \widetilde{\varphi}_k^{(n)})$ is temporarily removed from $\boldsymbol{\Theta}_{active}^{(n)}$ with $\widetilde{K} = \widetilde{K} - 1$, serving as the initial estimate in grid optimization. After re-estimation, if $p_{\widetilde{\theta}_k^{new},\widetilde{\varphi}_k^{new}} > s_{\widetilde{\theta}_k^{new},\widetilde{\varphi}_k^{new}}$, $(\widetilde{\theta}_k^{new}, \widetilde{\varphi}_k^{new})$ is accepted, replacing $(\widetilde{\theta}_k^{(n)}, \widetilde{\varphi}_k^{(n)})$, and is added back to $\boldsymbol{\Theta}_{active}^{(n)}$ with $\widetilde{K} = \widetilde{K} + 1$.

Otherwise, if $p_{\widetilde{\theta}_k^{new},\widetilde{\varphi}_k^{new}} \leq s_{\widetilde{\theta}_k^{new},\widetilde{\varphi}_k^{new}}$, $(\widetilde{\theta}_k^{new}, \widetilde{\varphi}_k^{new})$ is discarded, and $(\widetilde{\theta}_k^{(n)}, \widetilde{\varphi}_k^{(n)})$ is permanently removed from $\boldsymbol{\Theta}_{active}^{(n)}$. Then, the grid index of $(\widetilde{\theta}_k^{(n)}, \widetilde{\varphi}_k^{(n)})$ is removed from $\mathcal{I}_{active}$.

Step 5: Calculate the posterior signal statistics $\boldsymbol{\mu}^{(n)}(t)$ and $\boldsymbol{\Sigma}^{(n)}$ using (9). The signal estimate is obtained by $\widetilde{\mathbf{X}}^{(n)} = [\boldsymbol{\mu}^{(n)}(1), \cdots, \boldsymbol{\mu}^{(n)}(L)]$.

Step 6: Update the noise precision $\beta^{(n)}$ using (22).

(3) *Iteration termination conditions:* The iteration will terminate when the iteration number n reaches the maximum number, which is defined as $MaxIter$, or when the following convergence criterion is met:

$$\text{size}(\widetilde{\mathbf{X}}^{(n)}) = \text{size}(\widetilde{\mathbf{X}}^{(n-1)}) \text{ and } \frac{\left\| \widetilde{\mathbf{X}}^{(n)} - \widetilde{\mathbf{X}}^{(n-1)} \right\|_F^2}{\left\| \widetilde{\mathbf{X}}^{(n-1)} \right\|_F^2} \leq \tau, \tag{26}$$

where τ is the convergence threshold. The summary of the proposed DOA estimator using manifold matrix learning is given in Table 2.

Table 2. Proposed fast 2D DOA estimator using array manifold matrix learning.

Input: $\mathbf{Y}$, $MaxIter$ and τ.
Initialization: Set $\beta = 10^3$.
Obtain $(\theta_o^{(0)}, \varphi_o^{(0)})$ and $i^{opt(0)}$ following Table 1, and then compute $\alpha_{i^{opt(0)}}$ using (14).
Obtain $\boldsymbol{\Theta}_{active}^{(0)} = (\theta_o^{(0)}, \varphi_o^{(0)})$, $\widetilde{\boldsymbol{\alpha}}^{(0)} = \alpha_{i^{opt(0)}}$, $\widetilde{K} = 1$, and add $i^{opt(0)}$ to $\mathcal{I}_{ative}$.
Compute $\beta^{(0)}$ using (22), and set $n = 1$.

While $n \leq MaxIter$ and the convergence condition (26) not satisfied **do**
(a) Determine a new $(\theta_o^{(n)}, \varphi_o^{(n)})$ **outside of** $\boldsymbol{\Theta}_{active}^{(n-1)}$**:**
(a1) obtain $(\theta_o^{(n)}, \varphi_o^{(n)})$ and its grid index $i^{opt(n)}$, according to Table 1;
(a2) compute $\widetilde{\alpha}_{i^{opt(n)}}$ using (14).
(b) Evaluate $\Delta_o^{\updownarrow(n)}$ **and** $\Delta_k^{\updownarrow(n)}$**,** $k = 1, \cdots, \widetilde{K}$**:**
(b1) for $(\theta_o^{(n)}, \varphi_o^{(n)})$, compute $\Delta_o^{\updownarrow(n)}$ using (23);
(b2) for $(\widetilde{\theta}_k^{(n-1)}, \widetilde{\varphi}_k^{(n-1)})$ in $\boldsymbol{\Theta}_{active}^{(n-1)}$, update $s_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}}^{(n)}$ and $p_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}}^{(n)}$ using (24);
(b3) compute $\widetilde{\alpha}_k^{new}$ using (14), if $p_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}}^{(n)} > s_{\widetilde{\theta}_k^{(n-1)},\widetilde{\varphi}_k^{(n-1)}}^{(n)}$;
(b4) compute $\Delta_k^{\updownarrow(n)}$ using (25).
(c) Update $\boldsymbol{\Theta}_{active}^{(n)}$**:**
(c1) compare $\left\{\Delta_o^{\updownarrow(n)}, \Delta_k^{\updownarrow(n)}\right\}_{k=1}^{\widetilde{K}}$ and find the maximum one;
(c2) if it is $\Delta_o^{\updownarrow(n)}$, $\boldsymbol{\Theta}_{active}^{(n)} = \left\{\boldsymbol{\Theta}_{active}^{(n-1)}, (\theta_o^{(n)}, \varphi_o^{(n)})\right\}$, $\widetilde{\boldsymbol{\alpha}}^{(n)} = [(\widetilde{\boldsymbol{\alpha}}^{(n-1)})^T, \widetilde{\alpha}_{i^{opt(n)}}]^T$,
add $i^{opt(n)}$ to $\mathcal{I}_{active}$, and $\widetilde{K} = \widetilde{K} + 1$;
(c3) if it is $\Delta_k^{\updownarrow(n)} = \Delta_k^{\updownarrow\prime(n)}$, $\boldsymbol{\Theta}_{active}^{(n)} = (\boldsymbol{\Theta}_{active}^{(n-1)})_{-k}$, $\widetilde{\boldsymbol{\alpha}}^{(n)} = (\widetilde{\boldsymbol{\alpha}}^{(n-1)})_{-k}$,
remove the grid index of $(\widetilde{\theta}_k^{(n-1)}, \widetilde{\varphi}_k^{(n-1)})$ from $\mathcal{I}_{active}$, and $\widetilde{K} = \widetilde{K} - 1$;
(c4) if it is $\Delta_k^{\updownarrow(n)} = \Delta_k^{\updownarrow''(n)}$, $\boldsymbol{\Theta}_{active}^{(n)} = \boldsymbol{\Theta}_{active}^{(n-1)}$, and $\widetilde{\boldsymbol{\alpha}}^{(n)} = \widetilde{\boldsymbol{\alpha}}^{(n-1)}$ with $\widetilde{\boldsymbol{\alpha}}_k^{(n-1)} = \widetilde{\alpha}_k^{new}$.
(d) Sequentially update the active grids in $\boldsymbol{\Theta}_{active}^{(n)}$**:**
(d1) $K' = \widetilde{K}$;
for k = 1 to $\widetilde{K}$
(d2) remove $(\widetilde{\theta}_k^{(n)}, \widetilde{\varphi}_k^{(n)})$ from $\boldsymbol{\Theta}_{active}^{(n)}$, and $\widetilde{K} = \widetilde{K} - 1$;
(d3) obtain $(\widetilde{\theta}_k^{new}, \widetilde{\varphi}_k^{new})$ by grid optimization in Table 1, using $(\widetilde{\theta}_k^{(n)}, \widetilde{\varphi}_k^{(n)})$ as the initial estimate;
(d4) if $p_{\widetilde{\theta}_k^{new},\widetilde{\varphi}_k^{new}} > s_{\widetilde{\theta}_k^{new},\widetilde{\varphi}_k^{new}}$, add $(\widetilde{\theta}_k^{(n)}, \widetilde{\varphi}_k^{(n)}) = (\widetilde{\theta}_k^{new}, \widetilde{\varphi}_k^{new})$ back to $\boldsymbol{\Theta}_{active}^{(n)}$, and $\widetilde{K} = \widetilde{K} + 1$;

Table 2. *Cont.*

(d5) if $p_{\widetilde{\theta}_k^{new},\widetilde{\varphi}_k^{new}} \leq s_{\widetilde{\theta}_k^{new},\widetilde{\varphi}_k^{new}}$, discard $(\widetilde{\theta}_k^{new}, \widetilde{\varphi}_k^{new})$, and remove the grid index of $(\widetilde{\theta}_k^{(n)}, \widetilde{\varphi}_k^{(n)})$ from $\mathcal{I}_{active}$.
end
(e) Update $\widetilde{\mathbf{X}}^{(n)}$:
(e1) compute $\left\{\boldsymbol{\mu}^{(n)}(t)\right\}_{t=1}^{L}$ and $\boldsymbol{\Sigma}^{(n)}$ using (9);
(e2) obtain $\widetilde{\mathbf{X}}^{(n)} = [\boldsymbol{\mu}^{(n)}(1), \cdots, \boldsymbol{\mu}^{(n)}(L)]$.
(f) Update $\beta^{(n)}$ using (22).
(g) $n = n + 1$.
End while

Output: $\boldsymbol{\Theta}_{active}^{(n)}$ and $\widetilde{\mathbf{X}}^{(n)}$.

3.2.5. Enhanced-Efficiency Implementation for URAs

When dealing with a URA, the objective function $f(\boldsymbol{\Theta}_{Gi})$ in the active grid selection Problem (15) can be reformulated in the 2D FFT form, enabling direct application of the 2D zoom-FFT technique [20] to efficiently estimate the active grid.

Consider a URA that comprises $M = M_y \times M_z$ sensors equispaced by d_y in the y direction and d_z in the z direction, as depicted in Figure 1b. The manifold vector of a $(\theta_{Gi}, \varphi_{Gi})$ can be expressed as a Kronecker product of two components, given by:

$$\mathbf{a}(\theta_{Gi}, \varphi_{Gi}) = \mathbf{a}(u_i) \otimes \mathbf{a}(v_i), \tag{27}$$

where

$$\begin{cases} \mathbf{a}(u_i) = [1, \exp(j2\pi u_i), \cdots, \exp(j2\pi(M_y - 1)u_i)]^T \\ \mathbf{a}(v_i) = [1, \exp(j2\pi v_i), \cdots, \exp(j2\pi(M_z - 1)v_i)]^T \end{cases}, \tag{28}$$

with $u_i = d_y \sin\theta_{Gi} \sin\varphi_{Gi}/\lambda$ and $v_i = d_z \cos\theta_{Gi}/\lambda$, regarded as spatial frequencies.

For a $(\theta_{Gi}, \varphi_{Gi})$, $i \notin \mathcal{I}_{active}$, not in the active set $\boldsymbol{\Theta}_{active}$, the corresponding $s_{\theta_{Gi},\varphi_{Gi}}$, as expressed in (12), is $s_{\theta_{Gi},\varphi_{Gi}} = \mathbf{a}_i^H \mathbf{C}_{-i}^{-1} \mathbf{a}_i = \mathbf{a}^H(\theta_{Gi}, \varphi_{Gi}) \mathbf{C}^{-1} \mathbf{a}(\theta_{Gi}, \varphi_{Gi})$, where $\mathbf{a}_i$ is the compact representation of $\mathbf{a}(\theta_{Gi}, \varphi_{Gi})$, and $\mathbf{C}_{-i} = \mathbf{C} = \beta^{-1}\mathbf{I}_M + \mathbf{A}(\boldsymbol{\Theta}_{active})\widetilde{\boldsymbol{\Lambda}}^{-1}\mathbf{A}^H(\boldsymbol{\Theta}_{active})$. Using (27), $s_{\theta_{Gi},\varphi_{Gi}}$ can be reformulated as:

$$\begin{aligned} s_{\theta_i,\varphi_i} &= \mathbf{a}^H(\theta_{Gi}, \varphi_{Gi}) \mathbf{C}^{-1} \mathbf{a}(\theta_{Gi}, \varphi_{Gi}) \\ &= \beta \mathbf{a}^H(\theta_i, \varphi_i) \Big(\mathbf{I}_M - \beta \mathbf{A}(\boldsymbol{\Theta}_{active}) \boldsymbol{\Sigma} \mathbf{A}^H(\boldsymbol{\Theta}_{active}) \Big) \mathbf{a}(\theta_i, \varphi_i) \\ &= \beta M - \beta^2 \left\| \mathbf{a}^H(\theta_{Gi}, \varphi_{Gi}) \mathbf{A}(\boldsymbol{\Theta}_{active}) \boldsymbol{\Sigma}^{1/2} \right\|_2^2 \\ &= \beta M - \beta^2 \sum_{k=1}^{\widetilde{K}} \left| \mathbf{a}^H(v_i) \mathrm{vec}^{-1}(\mathbf{Z}(:,k)) \mathbf{a}^*(u_i) \right|^2, \end{aligned} \tag{29}$$

where $\mathbf{C}^{-1} = \beta\Big(\mathbf{I}_M - \beta \mathbf{A}(\boldsymbol{\Theta}_{active}) \boldsymbol{\Sigma} \mathbf{A}^H(\boldsymbol{\Theta}_{active})\Big)$, $\mathbf{Z} = \mathbf{A}\boldsymbol{\Sigma}^{1/2}$, and $\mathbf{Z}(:,k)$ is the kth column vector of $\mathbf{Z}$. The similar reformulation is applied to $p_{\theta_{Gi},\varphi_{Gi}}$ in (13), yielding:

$$\begin{aligned} p_{\theta_i,\varphi_i} &= (1/L) \sum_{t=1}^{L} \left| \mathbf{a}^H(\theta_i, \varphi_i) \mathbf{C}^{-1} \mathbf{y}(t) \right|^2 \\ &= (\beta^2/L) \sum_{t=1}^{L} \left| \mathbf{a}^H(\theta_i, \varphi_i) \Big(\mathbf{I}_M - \beta \mathbf{A}(\boldsymbol{\Theta}_{active}) \boldsymbol{\Sigma} \mathbf{A}^H(\boldsymbol{\Theta}_{active}) \Big) \mathbf{y}(t) \right|^2 \\ &= (\beta^2/L) \sum_{t=1}^{L} \left| \mathbf{a}^H(\theta_i, \varphi_i) (\mathbf{y}(t) - \mathbf{A}(\boldsymbol{\Theta}_{active}) \boldsymbol{\mu}(t)) \right|^2 \\ &= (\beta^2/L) \sum_{t=1}^{L} \left| \mathbf{a}^H(v_i) \mathrm{vec}^{-1}(\mathbf{y}(t) - \mathbf{A}(\boldsymbol{\Theta}_{active}) \boldsymbol{\mu}(t)) \mathbf{a}^*(u_i) \right|^2. \end{aligned} \tag{30}$$

Note that the manifold vector components, $\mathbf{a}(v_i)$ and $\mathbf{a}(u_i)$, as expressed in (28), have a structural similarity to the FFT basis function. This suggests that both $s_{\theta_{Gi},\varphi_{Gi}}$ and $p_{\theta_{Gi},\varphi_{Gi}}$ manifest a 2D FFT structure. Accordingly, the cost function of the active grid selection Problem (15) can be reformulated in the 2D FFT domain:

$$f(\mathbf{\Theta}_{Gi}) = \frac{p_{\theta_{Gi},\varphi_{Gi}}}{s_{\theta_{Gi},\varphi_{Gi}}} = \frac{\beta\sum_{t=1}^{L}\left|\mathbf{a}^H(v_i)\mathrm{vec}^{-1}(\mathbf{y}(t)-\mathbf{A}\boldsymbol{\mu}(t))\mathbf{a}^*(u_i)\right|^2}{L\left(M-\beta\sum_{k=1}^{\widetilde{K}}\left|\mathbf{a}^H(v_i)\mathrm{vec}^{-1}(\mathbf{Z}(:,k))\mathbf{a}^*(u_i)\right|^2\right)} \tag{31}$$

From this, the 2D zoom-FFT method [20] can be utilized to maximize (31) and obtain the optimal grid solution. In 2D zoom-FFT, the coarse optimal estimate is located and then refined to obtain a more accurate estimate. Compared to the conventional zero-padding 2D FFT, the 2D zoom-FFT requires fewer FFT points to achieve the same resolution, which results in increased computational efficiency.

Assume that an $(N_1 \times N_2)$-point 2D zoom-FFT is performed to estimate the optimal solution to the maximization of $f(\mathbf{\Theta}_{Gi})$. First, $p_{\theta_{Gi},\varphi_{Gi}}$ and $s_{\theta_{Gi},\varphi_{Gi}}$ are computed using an $(N_1 \times N_2)$-point 2D FFT, and $f(\mathbf{\Theta}_{Gi})$ is obtained as an $(N_1 \times N_2)$ matrix. The 2D spatial frequency pair $(v_{i^{opt}}, u_{i^{opt}})$ of the maximum value of $f(\mathbf{\Theta}_{Gi})$ is the coarse optimal solution. Subsequently, $(v_{i^{opt}}, u_{i^{opt}})$ is iteratively refined via the dichotomizing method introduced in the previous section. Put simply, each refinement iteration re-locates the maximum values of $f(\mathbf{\Theta}_{Gi})$ among the last updated estimate and its neighbor frequence points, which are newly interpolated in the search area. Defining the finally tuned 2D frequency pair by (v_o, u_o), following the previous notation, the optimal grid estimate (θ_o, φ_o) can be computed by:

$$\begin{cases} \theta_o = \cos^{-1}(v_o\lambda/d_z) \\ \varphi_o = \sin^{-1}\left(u_o\lambda/\left(d_y\sqrt{1-\cos^2\theta_o}\right)\right) \end{cases}, \tag{32}$$

where (θ_o, φ_o) is accepted if the constraint $p_{\theta_o,\varphi_o} > s_{\theta_o,\varphi_o}$ is satisfied; otherwise, (θ_o, φ_o) is discarded if $p_{\theta_o,\varphi_o} \leq s_{\theta_o,\varphi_o}$. It is noteworthy that the active grid selection and optimization processes are simultaneously achieved by maximizing (31) with the 2D zoom-FFT.

Recall the outline of the proposed fast 2D DOA estimator in Table 2. When dealing with a URA, the initialization, step a, and step d need to be modified. Specifically, the active grid estimation based on the criteria presented in Table 1 is replaced by maximizing (31) by the 2D zoom-FFT method, where the optimal grid estimate is obtained using (32).

3.3. Complexity Analysis

The computational complexity of the proposed fast 2D DOA estimator is measured by the number of multiplication operations in a single iteration. The estimator consists of two processes: the grid selection process and the grid optimization process, which are performed in an alternating manner. The grid selection process dominates the complexity due to its grid searching over predefined grids covering a large space, whereas the grid optimization process has negligible complexity as it is confined to a small bounded area around the selected grid. Recall the grid selection problem, as represented in (15). At each iteration, the parameters $p_{\theta_{Gi},\varphi_{Gi}}$ and $s_{\theta_{Gi},\varphi_{Gi}}$ are computed for $Q-\widetilde{K}$ candidate grid points $\{(\theta_{Gi},\varphi_{Gi})\}_{i=1,i\notin\mathcal{I}_{active}}^{Q}$, requiring $\mathcal{O}(\max\left\{ML(Q-\widetilde{K}), M^2(Q-\widetilde{K})\right\})$ complexity, where M, L, Q, and $\widetilde{K}$ are the numbers of sensors, snapshots, predefined grid points, and active grids, respectively. Additionally, the posterior estimates of source signals are computed at each iteration, resulting in a complexity of $\mathcal{O}(\max(M\widetilde{K}^2, M\widetilde{K}L))$. Therefore, the computational complexity of the proposed estimator is approximately $\mathcal{O}(\max\left\{ML(Q-\widetilde{K}), M^2(Q-\widetilde{K}), M\widetilde{K}^2, M\widetilde{K}L\right\})$ per iteration.

Regarding the grid-based SR method, whose computational complexity often concentrates on calculating the signal covariance matrix. The computational complexity of the conventional SBL method is $\mathcal{O}(\max(MQ^2, MQL))$ per iteration [10]. For the fast SR method SAMV-RELAX [13], its computational complexity depends on the SAMV step,

which is $\mathcal{O}(M^2Q)$ [12]. Since $\widetilde{K}$ increases from zero and retains a small value during estimation, we have $Q \gg \widetilde{K}$. Based on this, it can be observed that the proposed DOA estimator has a significantly lower complexity than the conventional SBL method, and a slight decrease in complexity compared to SAMV-RELAX, indicating improved efficiency.

For URAs, the efficiency of the proposed DOA estimator is improved by using the 2D zoom-FFT in active grid estimation compared to the dichotomizing-based solver. This is because the 2D zoom-FFT can require fewer FFT points than the number of predefined grids used by the dichotomizing-based solver, with similar accuracy.

3.4. Correlation to the FAMML Method

This work extends the 1D FAMML method in [17] to 2D DOA estimation and focuses on reducing computational complexity. The crucial task in the extension is to devise an efficient 2D grid refinement to reduce the off-grid errors and thereby achieve high accuracy. A dichotomizing-based grid refinement method, applicable to arbitrary planar arrays, is proposed, which iteratively modifies the coarse grid within a bounded area that shrinks with each iteration. This method directly maximizes the objective function in (16) without relying on derivatives on the objective function, as done in FAMML. FAMML will yield a multivariate equation system, while its root cannot be solved with the bisection method used by FAMML. Moreover, Newton's method, mentioned in FAMML, requires calculation of partial derivatives of (16), leading to increased computational complexity compared with the dichotomizing method. For 2D DOA estimation with URAs, a 2D zoom-FFT-based grid refinement is developed. This method exploits the Kronecker structure of the manifold vector to formulate the cost function in the 2D FFT domain, as expressed in (31). Based on this, the refined grid can be obtained by the 2D zoom-FFT technique, providing computational efficiency improvements over the dichotomizing-based solver.

4. Simulation Results

In this section, we compare the performance of the proposed fast 2D DOA estimator with two fast SR methods: SAMV-RELAX [13] and FGML [15]. The version of the proposed estimator that uses the dichotomizing method for grid optimization is referred to as Method 1. The URA-tailored variant that uses the 2D zoom-FFT is referred to as Method 2. The simulations are considered as being implemented in an underwater free field. The source velocity is 1500 m/s. Narrowband signals with frequency of 6 kHz and equal power of 0 dB impinge on the planar array. These source signals are located in the far field relative to the planar array. The sample frequency is 30 kHz.

During simulations, two array configurations are considered, a URA and an arbitrary planar array, as shown in Figure 1. The URA, consisting of $M = M_y \times M_z = 6 \times 4$ sensors with $d_y = d_z = \lambda/2$, is utilized for all four methods to assess their performance. The arbitrary planar array of M sensors is utilized to demonstrate Method 1's adaptability to different array configurations. Unless otherwise stated, simulation results are obtained from 200 Monte Carlo trials, where $K = 4$ uncorrelated narrowband signals with equal power arrival from directions $(\theta_{S1}, \varphi_{S1}) = (80.5^\circ, -37.5^\circ)$, $(\theta_{S2}, \varphi_{S2}) = (90.5^\circ, -12.6^\circ)$, $(\theta_{S3}, \varphi_{S3}) = (100.5^\circ, 8.2^\circ)$ and $(\theta_{S4}, \varphi_{S4}) = (110.6^\circ, 32.6^\circ)$, respectively. The root-mean-square error (RMSE) of DOA estimates is used to evaluate the estimation accuracy of the methods, with the Cramér–Rao bound (CRB) [24] serving as a benchmark. Additionally, the average runtime and resolution performance for both azimuth and elevation angles are evaluated to provide a comprehensive assessment of the methods. All simulations are carried out on a PC with 2.9 GHz CPU and 8 GB RAM.

The parameter settings for the evaluated methods are as follows. A uniform discretization of 2° is applied to both the azimuth range $[-90^\circ, 90^\circ]$ and the elevation range $[30^\circ, 150^\circ]$. For both proposed methods, the iteration parameters are set to $MaxIter = 50$ and $\tau = 10^{-5}$. In method 1, the iteration number for the dichotomizing-based solver of grid optimization is set to 15. In method 2, a (32×16)-point 2D zoom-FFT is employed, where the number of iterations for the refinement process is set to 8. For SAMV-RELAX,

the parameters are set to $MaxIter = 50$ and $\tau = 10^{-3}$ in the SAMV-step, and $\tau = 10^{-2}$ in the SAMV-step. Additionally, knowledge of the source number K is required by the SAMV-RELAX and FGML methods. While FGML assumes that K is known, SAMV-RELAX estimates it during the SAMV-step by identifying peaks in the estimated normalized spatial spectrum that surpass a threshold of -5 dB.

Simulation 1 assesses methods' estimation performance versus SNR in terms of RMSE and average runtime. The SNR varies from -10 dB to 20 dB, and the snapshots is $L = 50$. The URA is utilized for 2D DOA estimation. Figure 3 shows the RMSE results. For each evaluated method, the RMSE performance has a consistent trend across the azimuth and elevation angles. It can be observed that both Method 1 and Method 2 achieve RMSEs close to the CRB across all SNRs, while FGML deviates from the CRB due to insufficient snapshots. Regarding SAMV-RELAX, the RMSEs of this method closely approximate the CRB when $SNR > 5$ dB. However, it becomes considerably larger at lower SNRs. This phenomenon can be attributed to SAMV-RELAX's tendency to inaccurately estimate the source number K at relatively low SNRs, which negatively impacts the angle estimation accuracy during the RELAX-step. Figure 4 shows the average runtime of the four methods. One can see that SAMV-RELAX has the longest average runtime, while FGML has the shortest. The average runtime of the two proposed methods is positioned between those of FGML and SAMV-RELAX, with Method 2 showing improved efficiency by having a shorter average runtime than Method 1.

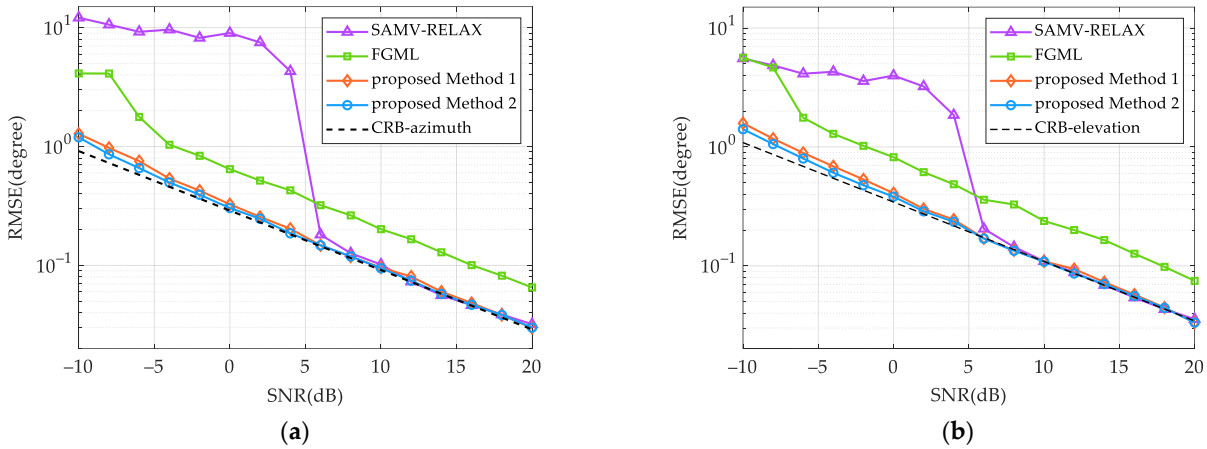


Figure 3. RMSE performance versus SNR, where a URA is utilized, and L is 50: (**a**) azimuth; (**b**) elevation.

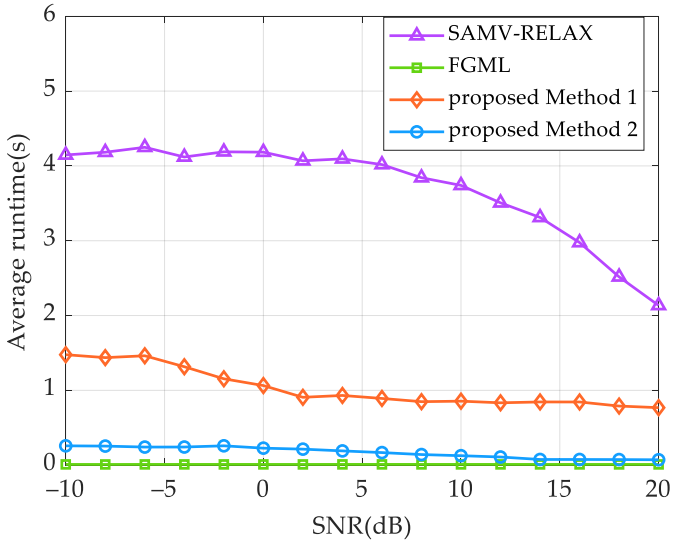


Figure 4. Average runtime versus SNR, where a URA is utilized, and L is 50.

Simulation 2 assesses the estimation performance versus the number of snapshots L, which is swept from 10 to 300. The SNR is fixed at 10 dB with other parameters unchanged. The RMSEs of the four methods are shown in Figure 5. Each method's RMSE performance has a consistent trend across the azimuth and elevation angles. From Figure 5, the two proposed methods and SAMV-RELAX exhibit similar RMSE performance, closely approaching the CRB across all L. FGML displays a decreasing RMSE trend as L increases, nearing the CRB at relatively large snapshots ($L \geq 200$). The corresponding average runtimes of the four methods are shown in Figure 6. It is seen that FGML is the fastest method, and Method 2 ranks second, followed by Method 1, whereas SAMV-RELAX is the slowest. Although both proposed methods take more time as the number of snapshots grows, Method 2 manages to maintain a runtime of less than one second, demonstrating its enhanced computational efficiency compared to Method 1.

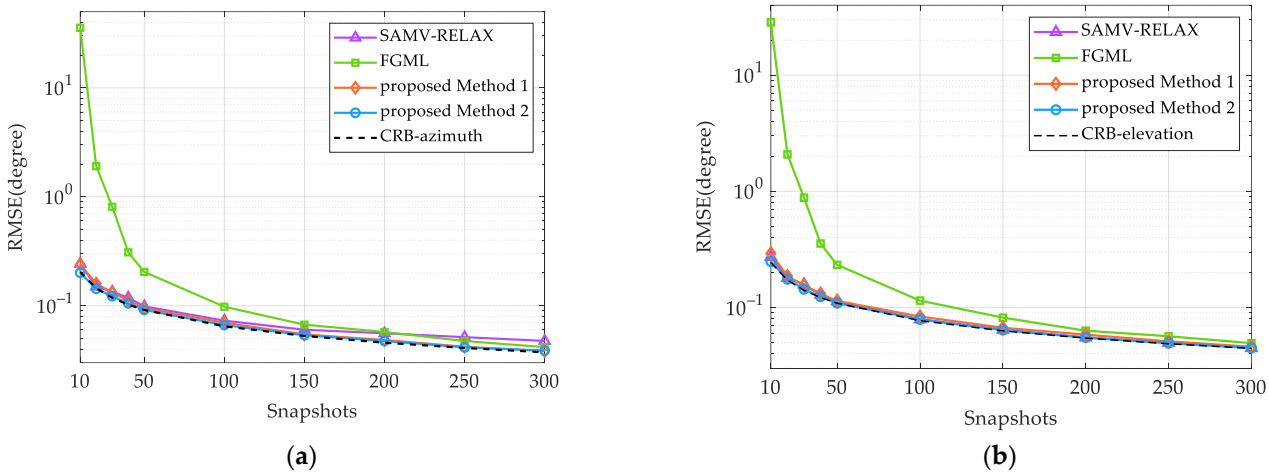


Figure 5. RMSE performance versus snapshots, where a URA is utilized, and SNR is 10 dB: (**a**) azimuth; (**b**) elevation.

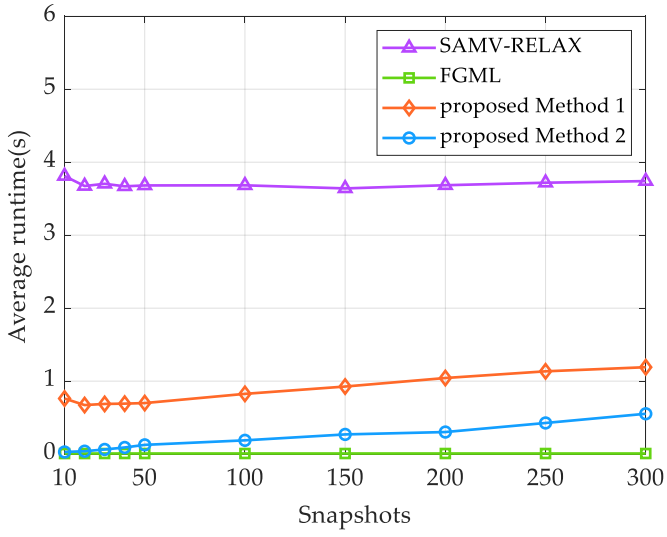


Figure 6. Average runtime versus snapshots, where a URA is utilized, and SNR is 10 dB.

Simulation 3 assesses the angle resolution performance by examining the probability of source resolution [25]. Two equal-power uncorrelated signals are analyzed with the SNR of 10 dB and 50 snapshots. Both the azimuth and elevation resolution of the four methods are compared in this simulation. The angular separation in terms of azimuth and elevation is defined as $\Delta\varphi$ and $\Delta\theta$, respectively, with values ranging from 2° to 10°. To model off-grid DOAs, we define an angular deviation as δ, which is randomly selected with a uniform

distribution within $[-0.5°, 0.5°]$. In assessing the azimuth resolution capability, the two signals have a common elevation angle of 90° and distinct azimuth angles of $-\Delta\varphi/2 + \delta_\varphi$ and $\Delta\varphi/2 + \delta_\varphi$, where δ_φ is the off-grid error of the azimuth. The two signals are said to be resolved in the lth run if [25]:

$$\sum_{i=1}^{2} |\hat{\varphi}_i(l) - \hat{\varphi}_i| < |\varphi_1 - \varphi_2|, \tag{33}$$

where $\hat{\varphi}_i(l)$ is the estimate of the true azimuth φ_i in the lth run. When assessing elevation resolution capability, the two signals have the same azimuth angle of 0° and distinct elevation angles of $-\Delta\theta/2 + 90° + \delta_\theta$ and $\Delta\theta/2 + 90° + \delta_\theta$. Similarly, the two signals are said to be resolved if the elevation estimates satisfy (33), in which $\hat{\varphi}_i(l)$ and φ_i are replaced by $\hat{\theta}_i(l)$ and θ_i, respectively.

Figure 7 shows the simulation results of the probability of source resolution. There is a consistent performance trend for both azimuth and elevation resolution. It is observed that Method 2 exhibits the best resolution performance even in small angular separation scenarios. Although Method 1 has slightly inferior resolution performance compared to Method 2, it outperforms both FGML and SAMV in cases of large angular separation.

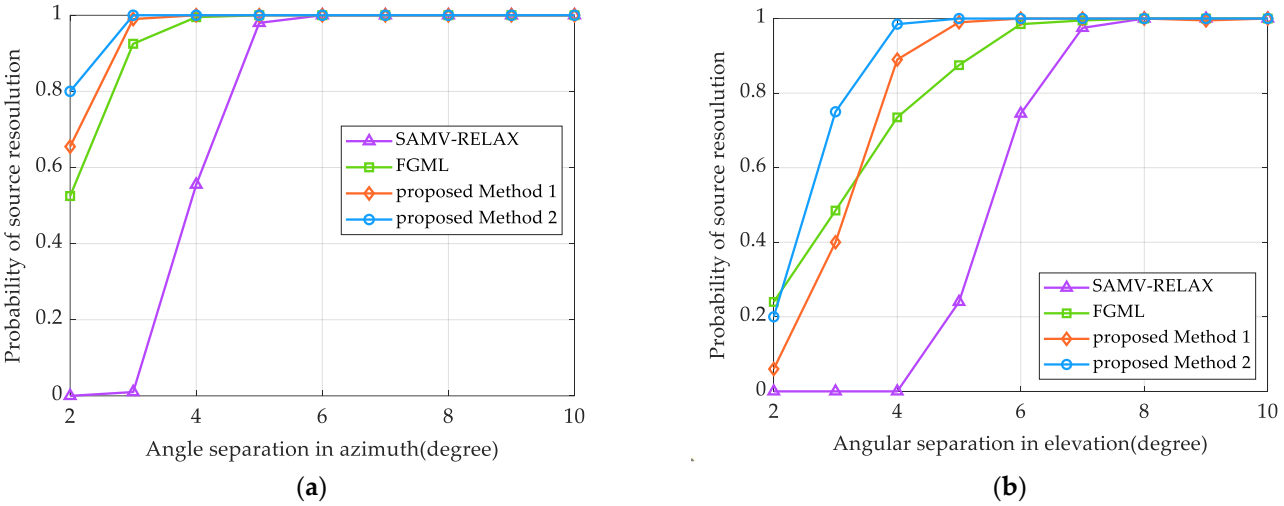


Figure 7. Probability of source resolution versus angular separation, where a URA is utilized, SNR is 10 dB, and L is 50: (**a**) azimuth; (**b**) elevation.

Simulation 4 examines the impact of angular grid interval (GI) on the RMSE performance and average runtime of the methods. In this simulation, GI varies from 1° to 4° for grid discretization in both azimuth range and elevation range, with the SNR of 10 dB and 50 snapshots. Set $K = 4$ signals with DOAs the same as in Simulation 1. Since Method 2 and FGML do not involve grid discretization, neither are considered.

Figure 8 shows the RMSE performance of Method 1 and SAMV-RELAX with respect to GI. Each method exhibits a consistent performance trend for both azimuth and elevation angles. Method 1 has stable RMSE values regardless of GI values, showing robustness to GI. In contrast, SAMV-RELAX exhibits comparable RMSE performance in the GI range of 2° to 3.5°, but it has higher RMSE values at other GI values, indicating reduced accuracy. This accuracy degradation in SAMV-RELAX can be attributed to overestimation of source number, leading to large errors in DOA estimates derived from the RELAX process. Considering the average runtime, Figure 9 shows that both Method 1 and SAMV-RELAX require less time as GI increases. This is because larger GI values lead to fewer grid points, resulting in a reduction in computational workload. When GI $< 3.5°$, Method 1 spends shorter time than SAMV-RELAX, showing better computational efficiency.

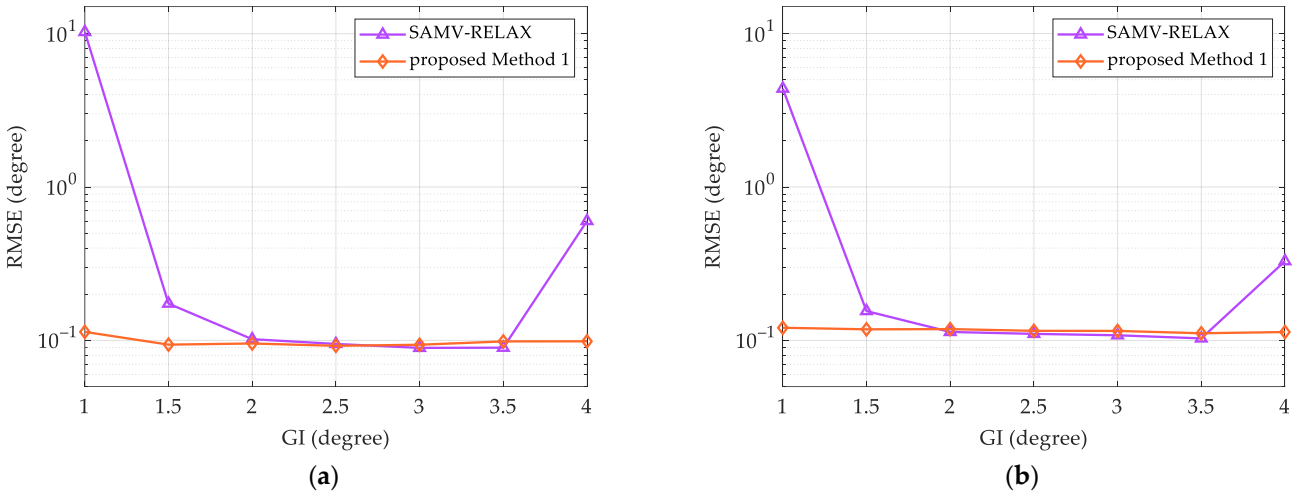


Figure 8. RMSE performance versus GI, where a URA is utilized, SNR is 10 dB, and L is 50: (**a**) azimuth; (**b**) elevation.

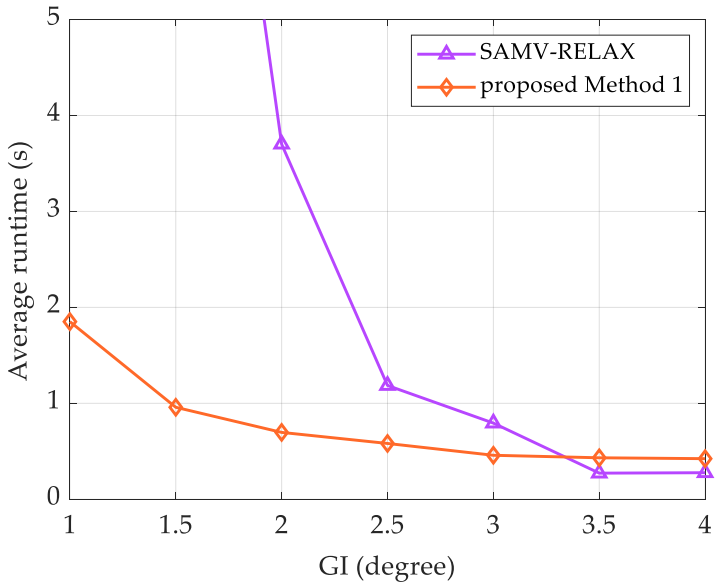
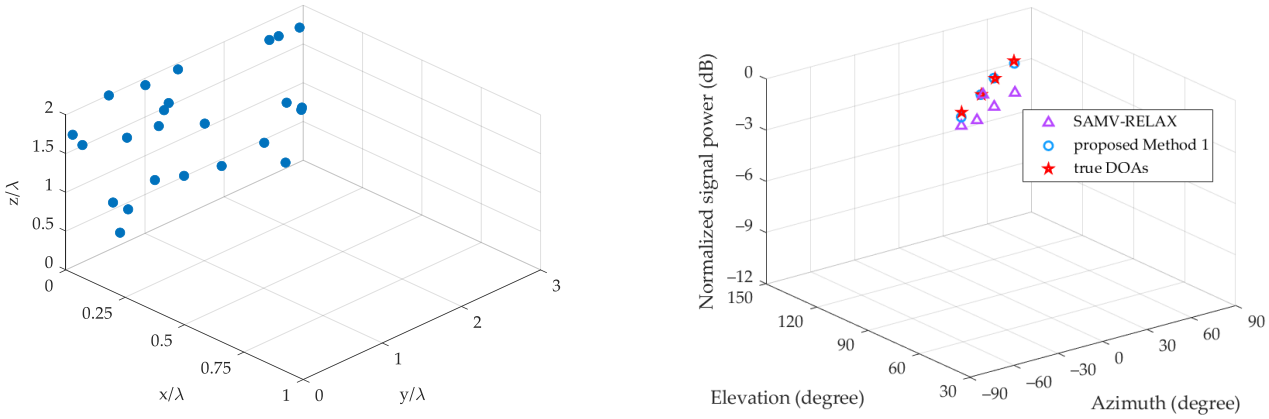


Figure 9. Average runtime versus GI, where a URA is utilized, SNR is 10 dB, and L is 50.

Simulation 5 verifies the adaptability of Method 1 to array configurations. The previous simulations utilize a URA with $M = 24$ sensors, while Simulation 5 adopts a non-rectangular array with the same number of sensors to estimate the DOAs of the aforementioned $K = 4$ signals from Simulation 1. The locations of array sensors $r_m = [0, y_m, z_m]^T$, $m = 1, \cdots, M$ are randomly generated within the range of $0 \leq y_m \leq 3\lambda$ and $0 \leq z_m \leq 2\lambda$. Only the proposed Method 1 and SAMV-REALX are available in this arbitrary planar array case, with SNR set to 0 dB and the snapshots to 50. The result of 2D DOA estimation obtained from a random trial is shown in Figure 10. One can see that Method 1 succeeds in localizing all four signals correctly, achieving better estimation than SAMV-RELAX.



(**a**) (**b**)

Figure 10. Result of 2D DOA estimation using an arbitrary planar array, where SNR is 0 dB, and L is 50: (**a**) array configuration; (**b**) 2D DOA estimates.

5. Conclusions

In this paper, we propose a fast 2D DOA estimator using array manifold matrix learning in the SBL framework. The proposed estimator achieves rapid and accurate 2D DOA estimation through grid selection and optimization processes based on marginal likelihood maximization. In grid optimization, an efficient dichotomizing-based solver is proposed for arbitrary planar arrays, while a 2D zoom-FFT based solver is introduced for URAs with improved computational efficiency. Simulation results demonstrate a superior balance between estimation accuracy and computational efficiency of the proposed DOA estimator compared to state-of-the-art methods. Moreover, the proposed estimator exhibits improvements in resolution probability.

Author Contributions: Conceptualization, J.L.; methodology, J.L. and L.Y.; validation, Y.Y.; Writing—original draft, J.L.; Writing—review and editing, L.Y., Y.Y. and L.W.; supervision, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant no. 12474442, grant no. 62431023 and grant no. 12204100) and the Science and Technology on Sonar Laboratory (grant no. 2023-JCJQ-LB-32/07).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Krim, H.; Viberg, M. Two Decades of Array Signal Processing Research: The Parametric Approach. *IEEE Signal Process. Mag.* **1996**, *13*, 67–94. [CrossRef]
2. Schmidt, R. Multiple Emitter Location and Signal Parameter Estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280. [CrossRef]
3. Roy, R.; Kailath, T. ESPRIT-Estimation of Signal Parameters via Rotational Invariance Techniques. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 984–995. [CrossRef]
4. Stoica, P.; Nehorai, A. MUSIC, Maximum Likelihood, and Cramer-Rao Bound. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 720–741. [CrossRef]
5. Malioutov, D.; Cetin, M.; Willsky, A.S. A Sparse Signal Reconstruction Perspective for Source Localization with Sensor Arrays. *IEEE Trans. Signal Process.* **2005**, *53*, 3010–3022. [CrossRef]
6. Yardibi, T.; Li, J.; Stoica, P.; Xue, M.; Baggeroer, A.B. Source Localization and Sensing: A Nonparametric Iterative Adaptive Approach Based on Weighted Least Squares. *IEEE Trans. Aerosp. Electron. Syst.* **2010**, *46*, 425–443. [CrossRef]
7. Stoica, P.; Babu, P.; Li, J. SPICE: A Sparse Covariance-Based Estimation Method for Array Processing. *IEEE Trans. Signal Process.* **2011**, *59*, 629–638. [CrossRef]

8. Abeida, H.; Zhang, Q.; Li, J.; Merabtine, N. Iterative Sparse Asymptotic Minimum Variance Based Approaches for Array Processing. *IEEE Trans. Signal Process.* **2013**, *61*, 933–944. [CrossRef]
9. Zhu, H.; Leus, G.; Giannakis, G.B. Sparsity-Cognizant Total Least-Squares for Perturbed Compressive Sampling. *IEEE Trans. Signal Process.* **2011**, *59*, 2002–2016. [CrossRef]
10. Yang, Z.; Xie, L.; Zhang, C. Off-Grid Direction of Arrival Estimation Using Sparse Bayesian Inference. *IEEE Trans. Signal Process.* **2013**, *61*, 38–43. [CrossRef]
11. Dai, J.; Bao, X.; Xu, W.; Chang, C. Root Sparse Bayesian Learning for Off-Grid DOA Estimation. *IEEE Signal Process. Lett.* **2017**, *24*, 46–50. [CrossRef]
12. Zhang, Y.; Yang, Y.; Yang, L.; Guo, X. Root Sparse Asymptotic Minimum Variance for Off-Grid Direction-of-Arrival Estimation. *Signal Process.* **2019**, *163*, 225–231. [CrossRef]
13. Park, H.-R.; Li, J. Efficient Sparse Parameter Estimation Based Methods for Two-Dimensional DOA Estimation of Coherent Signals. *IET Signal Process.* **2020**, *14*, 643–651. [CrossRef]
14. Li, J.; Zheng, D.; Stoica, P. Angle and Waveform Estimation via RELAX. *IEEE Trans. Aerosp. Electron. Syst.* **1997**, *33*, 1077–1087. [CrossRef]
15. Wu, X.; Zhu, W.-P.; Yan, J. A Fast Gridless Covariance Matrix Reconstruction Method for One- and Two-Dimensional Direction-of-Arrival Estimation. *IEEE Sens. J.* **2017**, *17*, 4916–4927. [CrossRef]
16. Zoltowski, M.D.; Haardt, M.; Mathews, C.P. Closed-Form 2-D Angle Estimation with Rectangular Arrays in Element Space or Beamspace via Unitary ESPRIT. *IEEE Trans. Signal Process.* **1996**, *44*, 316–328. [CrossRef]
17. Mao, Y.; Guo, Q.; Ding, J.; Liu, F.; Yu, Y. Marginal Likelihood Maximization Based Fast Array Manifold Matrix Learning for Direction of Arrival Estimation. *IEEE Trans. Signal Process.* **2021**, *69*, 5512–5522. [CrossRef]
18. Tipping, M.E. Sparse Bayesian Learning and the Relevance Vector Machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
19. Faul, A.; Avenuse, J. Fast Marginal Likelihood Maximisation for Sparse Bayesian Models. In *Proceedings of Machine Learning Research*; MLResearchPress: Singapore, 2002.
20. Liu, Z.-S.; Li, J. Implementation of the RELAX Algorithm. *IEEE Trans. Aerosp. Electron. Syst.* **1998**, *34*, 657–664.
21. Wipf, D.P.; Rao, B.D. An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem. *IEEE Trans. Signal Process.* **2007**, *55*, 3704–3716. [CrossRef]
22. Liu, Z.-M.; Huang, Z.-T.; Zhou, Y.-Y. An Efficient Maximum Likelihood Method for Direction-of-Arrival Estimation via Sparse Bayesian Learning. *IEEE Trans. Wirel. Commun.* **2012**, *11*, 1–11. [CrossRef]
23. Gerstoft, P.; Mecklenbräuker, C.F.; Xenaki, A.; Nannuru, S. Multisnapshot Sparse Bayesian Learning for DOA. *IEEE Signal Process. Lett.* **2016**, *23*, 1469–1473. [CrossRef]
24. Dai, X.; Zhang, X.; Wang, Y. Extended DOA-Matrix Method for DOA Estimation via Two Parallel Linear Arrays. *IEEE Commun. Lett.* **2019**, *23*, 1981–1984. [CrossRef]
25. Gershman, A.B. Direction Finding Using Beamspace Root Estimator Banks. *IEEE Trans. Signal Process.* **1998**, *46*, 3131–3135. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

Gridless DOA Estimation Method for Arbitrary Array Geometries Based on Complex-Valued Deep Neural Networks

Yuan Cao, Tianjun Zhou and Qunfei Zhang *

School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China; chaya66@mail.nwpu.edu.cn (Y.C.); zhoutianjun@mail.nwpu.edu.cn (T.Z.)
* Correspondence: zhangqf@nwpu.edu.cn

Abstract: Gridless direction of arrival (DOA) estimation methods have garnered significant attention due to their ability to avoid grid mismatch errors, which can adversely affect the performance of high-resolution DOA estimation algorithms. However, most existing gridless methods are primarily restricted to applications involving uniform linear arrays or sparse linear arrays. In this paper, we derive the relationship between the element-domain covariance matrix and the angular-domain covariance matrix for arbitrary array geometries by expanding the steering vector using a Fourier series. Then, a deep neural network is designed to reconstruct the angular-domain covariance matrix from the sample covariance matrix and the gridless DOA estimation can be obtained by Root-MUSIC. Simulation results on arbitrary array geometries demonstrate that the proposed method outperforms existing methods like MUSIC, SPICE, and SBL in terms of resolution probability and DOA estimation accuracy, especially when the angular separation between targets is small. Additionally, the proposed method does not require any hyperparameter tuning, is robust to varying snapshot numbers, and has a lower computational complexity. Finally, real hydrophone data from the SWellEx-96 ocean experiment validates the effectiveness of the proposed method in practical underwater acoustic environments.

Keywords: gridless direction of arrival (DOA) estimation; complex-valued deep neural network; arbitrary array geometry; high resolution

Citation: Cao, Y.; Zhou, T.; Zhang, Q. Gridless DOA Estimation Method for Arbitrary Array Geometries Based on Complex-Valued Deep Neural Networks. *Remote Sens.* **2024**, *16*, 3752. https://doi.org/10.3390/rs16193752

Academic Editors: Xuebo Zhang, Lei Zhang, Jingyue Lu, Haiyong Xu, Marc Pinto and Farman Ali

Received: 11 September 2024
Revised: 6 October 2024
Accepted: 7 October 2024
Published: 9 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Direction of arrival (DOA) estimation is one of the most fundamental problems in array signal processing, which obtains azimuthal information about targets by sampling the signal in the spatial domain [1,2]. Conventional DOA estimation methods are mainly based on beamforming, subspace decomposition, or sparse reconstruction. Conventional beamforming (CBF) is the simplest beamforming technique and has been widely used in practice due to its very low complexity and high robustness. However, the resolution of CBF can only be enhanced by expanding the array aperture and cannot be improved by increasing the number of snapshots or the signal-to-noise ratio (SNR) [3]. To enhance the performance of CBF when the array aperture is limited, superdirectivity beamforming methods, such as Capon, have been proposed. These methods greatly enhance the resolution of small-aperture arrays, but at the cost of increased sensitivity to errors, which affects robustness [4]. Subspace decomposition is another classical class of DOA estimation methods, of which the MUSIC [5] and ESPRIT [6] algorithms are the most representative. The advantage of such methods is their excellent performance when the number of snapshots is sufficient. However, if the snapshot number is insufficient, the signal covariance matrix cannot be accurately estimated, leading to a significant decline in algorithm performance [7].

Subsequently, sparse reconstruction DOA estimation methods have been proposed based on signal sparsity. These methods are generally divided into three categories based on the signal model: on-grid, off-grid, and gridless approaches [8]. On-grid methods

discretize the continuous angular domain into a finite set of points, assuming that the actual DOAs coincide with these discrete points [9–12]. While these methods enable the direct application of compressed sensing techniques, they are susceptible to grid mismatch errors. Off-grid methods, though they estimate DOAs within the continuous angular domain, still rely on the discretized grid model [13–16]. To alleviate grid mismatch, these methods employ strategies such as first-order Taylor approximations of the steering vector near the grid [14] or adaptive grid adjustments [15]. However, their performance is contingent on the accuracy of these approximations, and they often introduce non-convex optimization problems, complicating the solution process. In contrast, gridless methods perform DOA estimation directly in the continuous angular domain without grid discretization. These approaches typically rely on techniques like Toeplitz covariance matrix reconstruction or trigonometric polynomial rooting [8,17]. Initially, gridless methods were limited to uniform linear arrays (ULAs) and sparse linear arrays (SLAs), but they have since been extended to coprime arrays through interpolation techniques [18]. Recent developments have further broadened their applicability to arbitrary array geometries by representing the steering vector as a Fourier series [19–23]. Despite these advancements, gridless methods often require semi-definite programming, which can result in high computational complexity. Additionally, in methods that use convex optimization to solve for the spatial spectrum and obtain DOA estimates, hyperparameter tuning presents a significant challenge. Improper hyperparameter settings can severely degrade the performance of these algorithms [9].

In recent years, deep learning (DL) has been extensively applied to DOA estimation tasks due to its non-linear capabilities and feature-learning potential [24]. Compared to sparse reconstruction methods, DL-based approaches offer several advantages. First, DOA estimation using a trained deep network bypasses the need for high-dimensional matrix inversion and multiple iterations, leading to a significant reduction in computational complexity. Second, the process does not require hyperparameter tuning, which helps avoid performance degradation due to improper parameter settings. Currently, most DL methods frame the DOA estimation problem as a multi-label classification task, where the deep network outputs the probability of targets being present in different directions. However, this approach is still fundamentally based on the discrete grid model [25–28]. While increasing the grid resolution can theoretically reduce grid mismatch errors, it also increases the output dimensionality of the network, making the training process more difficult. Although gridless DL-based DOA estimation methods have also been explored, similar to gridless sparse reconstruction approaches, they are typically limited to ULA, SLA [29–31], or coprime arrays [32].

In this paper, we first derive the Toeplitz structure of the angular-domain covariance matrix for arbitrary array geometries. Then, we design and train a complex-valued deep neural network (CDNN) to reconstruct this Toeplitz matrix from the sample covariance matrix (SCM) and subsequently employ the Root-MUSIC method to achieve gridless DOA estimation for arbitrary array geometries. The main contributions of this paper are as follows: (1) We represent the steering vector as a Fourier series and derive the Toeplitz structure of the angular-domain covariance matrix for arbitrary array geometries. Furthermore, we establish the relationship between this matrix and the element-domain covariance matrix, providing a theoretical foundation for gridless DOA estimation using SCM in arbitrary array geometries. (2) We propose a novel approach that utilizes CDNN to reconstruct the Toeplitz matrix from SCM and leverages Root-MUSIC for gridless DOA estimation. This method avoids grid mismatch errors, eliminates the need for hyperparameter tuning, and delivers strong performance with low computational complexity. (3) We analyze the impact of the Toeplitz matrix's dimensionality on DOA estimation performance.

The rest of this paper is organized as follows: Section 2 presents the DOA estimation problem model and derives the Toeplitz structure of the angular-domain covariance matrix from the element-domain covariance matrix. Section 3 describes the training of the CDNN for the subarray of horizontal linear array sorting (HLAS) in the Swellex-96 experiment, to reconstruct the Toeplitz structure of the angular-domain covariance matrix. In Section 4, the

performance of the proposed method is assessed through numerical simulations. Section 5 demonstrates the algorithm's effectiveness using real data obtained from sea trials. Finally, Section 6 provides the conclusions and discussions of this paper.

2. Problem Formulation

In this section, we present the mathematical model for gridless DOA estimation. Specifically, Section 2.1 introduces the general model for the DOA estimation problem in a two-dimensional (2D) plane. Section 2.2 derives the Toeplitz structure of the angular-domain covariance matrix for arbitrary array geometries in the DOA estimation problem.

2.1. Signal Model

In a 2D plane, we consider the scenario where K far-field narrowband plane waves with DOAs $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_K]$ impinge on an array of M omnidirectional sensors. The mutual coupling effects between sensors are not considered. It is assumed that $K < M$ and the number of sources K is known. The received snapshots of the array can be expressed as

$$\mathbf{y}_l = \mathbf{A}(\boldsymbol{\theta})\mathbf{x}_l + \mathbf{n}_l, \tag{1}$$

where $\mathbf{A}(\boldsymbol{\theta}) = [\mathbf{a}(\theta_1), \mathbf{a}(\theta_2), \cdots, \mathbf{a}(\theta_K)] \in \mathbb{C}^{M\times K}$ is the steering matrix of the array corresponding to the directions $\boldsymbol{\theta}$. $\mathbf{x}_l \in \mathbb{C}^K$ and $\mathbf{n}_l \in \mathbb{C}^M$ are the signal vector and noise vector at the lth snapshot, respectively, and $l = 1, 2, \cdots, L$ is the snapshot index. For an array with arbitrary geometry, the steering vector can be expressed as

$$\mathbf{a}(\theta) = [e^{j2\pi f r_1 \cos(\theta-\varphi_1)/c}, e^{j2\pi f r_2 \cos(\theta-\varphi_2)/c}, \cdots, e^{j2\pi f r_M \cos(\theta-\varphi_M)/c}]^T, \tag{2}$$

where the superscript ^{T} denotes the transpose operator, f is the center frequency of the narrowband signal, and (r_m, φ_m) are the polar coordinates of the mth array element, with c being the speed of waves.

The covariance matrix of the received signal in the element domain can be written as

$$\mathbf{R}_y = \mathbf{A}\mathbf{R}_x\mathbf{A}^H + \mathbf{R}_n, \tag{3}$$

where the superscript ^{H} denotes the conjugate transpose operator, and $\mathbf{R}_x$ and $\mathbf{R}_n$ are the covariance matrices of the source signals and noise, respectively. For uncorrelated sources, $\mathbf{R}_x = diag({\alpha_1}^2, {\alpha_2}^2, \cdots, {\alpha_K}^2)$, where ${\alpha_k}^2$ is the power of the kth source. For simplicity, we abbreviate $\mathbf{A}(\boldsymbol{\theta})$ as $\mathbf{A}$. In practical scenarios with limited observation data, the SCM $\widehat{\mathbf{R}_y} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{y}_l{\mathbf{y}_l}^H$ is often used to estimate $\mathbf{R}_y$. When the noise is Gaussian, $\widehat{\mathbf{R}_y}$ provides an unbiased estimate of $\mathbf{R}_y$.

2.2. Angular-Domain Covariance Matrix

In a 2D plane, the steering vector $\mathbf{a}(\theta)$ exhibits periodicity with respect to θ, with a period of 2π. Therefore, it can be expanded into a Fourier series in terms of θ, given by

$$[\mathbf{a}(\theta)]_m = \sum_{n=-\infty}^{\infty} b_{nm}e^{jn\theta} \approx \sum_{n=-N}^{N} b_{nm}e^{jn\theta}, \tag{4}$$

where $[\mathbf{a}(\theta)]_m$ represents the mth element of $\mathbf{a}(\theta)$, n is the Fourier series index, and b_{nm} is the Fourier coefficient. When $b_{Nm} \approx 0$, the Fourier series can be truncated at $n = N$, and the resulting expression in matrix form is

$$\mathbf{a}(\theta) = \mathbf{B}^H\mathbf{f}(\theta), \tag{5}$$

where $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_M] \in \mathbb{C}^{(2N+1)\times M}$ is the Fourier coefficient matrix, $\mathbf{b}_m = [b_{-Nm}, \cdots, b_{0m}, \cdots, b_{Nm}]^H$ is the coefficient vector, and $\mathbf{f}(\theta) = [e^{-jN\theta}, \cdots, 1, \cdots, e^{jN\theta}]^T$. Therefore, the steering matrix $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \mathbf{B}^H \mathbf{F}, \tag{6}$$

where $\mathbf{F} = [\mathbf{f}(\theta_1), \mathbf{f}(\theta_2), \cdots, \mathbf{f}(\theta_K)]$. In the absence of noise, substituting (6) into (3) yields

$$\widetilde{\mathbf{R}_y} = \mathbf{B}^H \mathbf{F} \mathbf{R}_x \mathbf{F}^H \mathbf{B} = \mathbf{B}^H \mathbf{T}(\mathbf{u}) \mathbf{B}, \tag{7}$$

where $\mathbf{T}(\mathbf{u}) = \mathbf{F}\mathbf{R}_x\mathbf{F}^H$ is a Hermitian Toeplitz matrix, and the vector $\mathbf{u} \in \mathbb{C}^{2N+1}$ represents the first row of the matrix $\mathbf{T}(\mathbf{u})$. According to [33], the Fourier coefficients $b_{nm} = j^n J_n(2\pi f r_m/c) e^{-jn\varphi_m}$, where $J_n(\bullet)$ denotes the nth-order Bessel function of the first kind. We refer to $\mathbf{T}(\mathbf{u})$ as the angular-domain covariance matrix. It is evident that $\mathbf{T}(\mathbf{u})$ has a rank of K and is positive semi-definite. According to the Vandermonde decomposition theorem for semi-definite Toeplitz matrices, $\mathbf{F}$ and $\mathbf{R}_x$ can be uniquely recovered from $\mathbf{T}(\mathbf{u})$, enabling gridless DOA estimation [8,34].

It is important to note that cross-correlation between signals, as well as between signals and noise, causes the SCM to deviate from the true element-domain covariance matrix [35,36]. This cross-correlation introduces inaccuracies in the signal subspace, resulting in degraded performance of subspace-based algorithms [36]. By recovering a rank-deficient Toeplitz matrix $\mathbf{T}(\mathbf{u})$, the influence of signal correlation on the signal subspace can be mitigated, thereby improving the performance of subspace algorithms. However, directly reconstructing $\mathbf{T}(\mathbf{u})$ from $\widehat{\mathbf{R}_y}$ is challenging. Therefore, we propose using a CDNN to learn the mapping between $\widehat{\mathbf{R}_y}$ and $\mathbf{T}(\mathbf{u})$, and subsequently employing the Root-MUSIC algorithm [37], which obtains DOA estimates by utilizing the roots of a polynomial derived from $\mathbf{T}(\mathbf{u})$. This approach aims to achieve gridless DOA estimation and improve performance for arbitrary array geometries.

3. Angular-Domain Covariance Matrix Reconstruction Network

In this section, we propose a CDNN to model the mapping between the SCM $\widehat{\mathbf{R}_y}$ and the angular-domain covariance matrix $\mathbf{T}(\mathbf{u})$ using the subarray composed of elements #1 to #8 of the HLAS in the Swellex-96 experiment [38]. Section 3.1 introduces the CDNN, Section 3.2 provides the array geometry and parameter settings, Section 3.3 describes the dataset used for training the network, and Section 3.4 presents the proposed network framework along with the training process and results.

3.1. Complex-Valued Deep Neural Network

CDNNs [39] can capture the phase information between the real and imaginary components of complex numbers, which are often crucial in signal processing tasks. As a result, CDNNs have been widely applied in these tasks [40–42], achieving better performance compared to real-valued deep neural networks (RDNNs). To achieve better performance in covariance matrix reconstruction with fewer parameters, we employ CDNN. The principles of CDNN are similar to those of RDNN. In the case of a fully connected neural network, it consists of an input layer, an output layer, and multiple cascaded hidden layers. Within each hidden layer, the input undergoes a linear transformation followed by a nonlinear activation function. This process can be expressed as

$$\mathbf{o} = \sigma(\mathbf{z}) = \sigma(\mathbf{W}\mathbf{o}' + \mathbf{b}), \tag{8}$$

where $\mathbf{W}$ and $\mathbf{b}$ are the complex weight matrix and bias vector, respectively, $\mathbf{o}'$ and $\mathbf{o}$ are the complex outputs of the previous layer and the current layer, respectively, and $\sigma(\bullet)$

represents the complex nonlinear activation function. The complex matrix multiplication in real space can be expressed as

$$\begin{bmatrix} \Re(\mathbf{z}) \\ \Im(\mathbf{z}) \end{bmatrix} = \begin{bmatrix} \Re(\mathbf{W}) & -\Im(\mathbf{W}) \\ \Im(\mathbf{W}) & \Re(\mathbf{W}) \end{bmatrix} \bullet \begin{bmatrix} \Re(\mathbf{o}') \\ \Im(\mathbf{o}') \end{bmatrix} + \begin{bmatrix} \Re(\mathbf{b}) \\ \Im(\mathbf{b}) \end{bmatrix}, \tag{9}$$

where $\Re(\bullet)$ and $\Im(\bullet)$ denote the real and imaginary parts of the complex number, respectively. According to (8) and (9), a complex-valued layer can be expressed as

$$\mathbf{h}(\mathbf{o}) = (\mathbf{h}_r(\mathbf{o}_r) - \mathbf{h}_i(\mathbf{o}_i)) + j(\mathbf{h}_r(\mathbf{o}_i) + \mathbf{h}_r(\mathbf{o}_i)), \tag{10}$$

where $\mathbf{o}_r = \Re(\mathbf{o})$, $\mathbf{o}_i = \Im(\mathbf{o})$, $\mathbf{h}_r(\mathbf{x}) = \Re(\mathbf{W})\mathbf{x} + \Re(\mathbf{b})$, $\mathbf{h}_i(\mathbf{x}) = \Im(\mathbf{W})\mathbf{x} + \Im(\mathbf{b})$. Thus, CDNN is an extension of RDNN, and a CDNN of the same scale has twice the number of parameters as an RDNN.

3.2. Array Geometry and Parameter Settings

We use the subarray composed of elements #1 to #8 from HLAS in Swellex-96 to generate a dataset for training the CDNN for angular-domain covariance matrix reconstruction. With element #5 as the origin of the Cartesian coordinate system, the coordinates of the 8-element array are shown in Table 1. The array is approximated as a linear array with unequal element spacing, with element #1 aligned in the direction of azimuth $\theta = 0$. The DOA is defined as the angle between the end-fire direction of the #1 element and the direction of the incoming source, with the counterclockwise direction considered positive, ranging from 0° to 180°. The center frequency of the narrowband far-field plane wave is set to $f = 50$ Hz, and the speed of sound in water is assumed to be $c = 1490$ m/s. The maximum element spacing in the linear array is $d_{\max} = 0.21\lambda$, and the minimum element spacing is $d_{\min} = 0.12\lambda$.

Table 1. Array Coordinates.

Element No.	#1	#2	#3	#4	#5	#6	#7	#8
x (m)	11.93	9.26	6.50	3.41	0	−3.75	−7.87	−12.43
y (m)	−10.4	−8.24	−5.78	−3.06	0	3.40	7.34	11.62

According to [19,21], when the wavelength of the narrowband signal is fixed, the error between the steering vector approximated by (5) and the true steering vector is related to the distance of the array elements from the coordinate origin, as well as the truncation order N. Figure 1a shows the magnitude of the Fourier coefficients for different element distances when $f = 50$ Hz, $\theta = 0$, and $\varphi_m = 0$, with the Fourier coefficients expressed in decibels, i.e., $20\lg|b_n|$. Figure 1b shows the relationship between the steering vector error and truncation order for different azimuth angles in the array specified in Table 1, with the steering vector error also expressed in decibels, i.e., $20\lg\left|\mathbf{a}(\theta) - \mathbf{B}^H\mathbf{f}(\theta)\right|$. We find that the magnitude of the Fourier coefficients decreases with increasing order, so higher truncation orders reduce the approximation error in (4) and (5), making $\widehat{\mathbf{R}_y}$ closer to the true element-domain covariance matrix. At the same time, the dimensionality of $\mathbf{T}(\mathbf{u})$ increases with higher truncation orders, theoretically enhancing the performance of DOA estimation. However, this increase in dimensionality also raises the complexity of the network by expanding the number of output parameters, making the network more challenging to train. Therefore, we set the truncation orders $N = 4$, $N = 7$ and $N = 10$ in the following experiments and analyze the effect of truncation order on the performance of CDNN in reconstructing $\mathbf{T}(\mathbf{u})$ through simulation results.

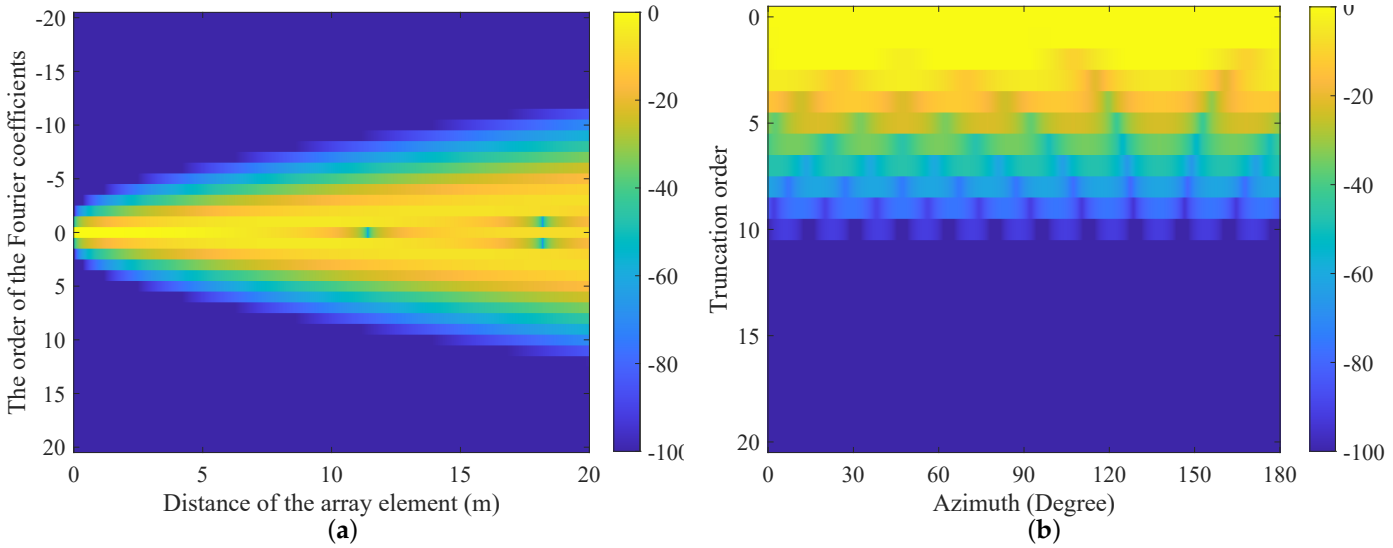


Figure 1. (**a**) Magnitude of Fourier coefficients at various orders. (**b**) Relationship between array steering vector error and N.

3.3. Dataset

We use MATLAB to generate the dataset for training the CDNN according to the array geometry and parameters provided in Section 3.2. The number of incoherent sources in the dataset is denoted as $K = \{2, 3\}$, with data corresponding to $K = 2$ and $K = 3$ each comprising half of the dataset. In the data numbered i, the DOA θ_{ik} for the kth incoherent source are randomly generated within the range of $[0, \pi]$, following an independent uniform distribution. The source amplitudes $\alpha_{i1} = 1$, and α_{ik}, $k \neq 1$ follow independent uniform distributions over $[0, 1]$. To enhance the robustness of the CDNN in real noise environments, we introduce spatially and temporally uncorrelated Gaussian white noise into the data, with the SNR uniformly distributed over $[0, 20]$dB. For multiple sources, the SNR is defined as $\frac{\sum_{k=1}^{K} {\alpha_k}^2}{\sigma^2}$, where σ^2 represents the noise power. Substituting the above parameters into (1), we obtain the array snapshot signal $\mathbf{Y}_i = [\mathbf{y}_{i1}, \mathbf{y}_{i2}, \cdots, \mathbf{y}_{iL}]$, where $\mathbf{y}_{il} \sim CN(\mathbf{0}, \mathbf{A}_i \mathbf{R}_{xi} {\mathbf{A}_i}^H + \frac{\sum_{k=1}^{K} {\alpha_{ik}}^2}{SNR_i}\mathbf{I})$, $\mathbf{R}_{xi} = diag({\alpha_{i1}}^2, \cdots, {\alpha_{iK}}^2)$, $\mathbf{A}_i = [\mathbf{a}(\theta_{i1}), \cdots, \mathbf{a}(\theta_{iK})]$, and $\mathbf{I}$ is the identity matrix. To enhance the randomness of the dataset, we set the number of snapshots for ith data sample as $L_i = \mathrm{mod}\,(i, 10) \times 10$, where $\mathrm{mod}\,(x, y)$ represents the remainder when dividing x by y.

We model the mapping from $\widehat{\mathbf{R}_y}$ to $\mathbf{T}(\mathbf{u})$ as a regression task, with $\widehat{\mathbf{R}_y}$ as the input and $\mathbf{T}(\mathbf{u})$ as the target label. The input data $\widehat{\mathbf{R}_y}$ is the SCM of the array's snapshot signal $\mathbf{Y}$, which is a complex-valued matrix of dimension $M \times M$. For the ith input data $\widehat{\mathbf{R}_{y_i}} = \mathbf{Y}_i {\mathbf{Y}_i}^H / L_i$, the corresponding label is the ideal angular-domain covariance matrix $\mathbf{T}(\mathbf{u})_i = {\mathbf{F}_i}^H \mathbf{R}_{xi} \mathbf{F}_i$, where $\mathbf{F}_i = [\mathbf{f}(\theta_{i1}), \mathbf{f}(\theta_{i2}), \cdots, \mathbf{f}(\theta_{iK})]$ is a complex-valued matrix of dimension $(2N + 1) \times (2N + 1)$. Therefore, the ith data pair in the dataset is represented as $\left(\widehat{\mathbf{R}_{y_i}}, \mathbf{T}(\mathbf{u})_i\right)$, forming a dataset $\varepsilon_{tra} = \left\{\left(\widehat{\mathbf{R}_{y_1}}, \mathbf{T}(\mathbf{u})_1\right), \left(\widehat{\mathbf{R}_{y_2}}, \mathbf{T}(\mathbf{u})_2\right), \cdots, \left(\widehat{\mathbf{R}_{y_{E_{tra}}}}, \mathbf{T}(\mathbf{u})_{E_{tra}}\right)\right\}$ of size $E_{tra} = 120,000$ to be used as the training set. Similarly, we generate a validation set ε_{val} of size $E_{val} = 10,000$ for model validation during training. Additionally, we construct a test set ε_{test} of size $E_{test} = 10,000$ to evaluate the performance of the trained CDNN.

3.4. Network Architecture and Training

As shown in Figure 2, the proposed angular-domain covariance matrix reconstruction CDNN consists of input, output, and dense layers. Since the input SCM is a Hermitian

matrix, its upper triangular elements cover all the matrix information. Therefore, the input layer takes a vector $\hat{\mathbf{r}}_y \in \mathbb{C}^{M(M+1)/2}$ composed of the upper triangular part of the SCM. The output layer produces a vector $\mathbf{u} \in \mathbb{C}^{2N+1}$, composed of the first row elements of $\mathbf{T}(\mathbf{u})$. After post-processing $f_p(\mathbf{u}) = \mathbf{T}(\mathbf{u})$, the reconstructed angular-domain covariance matrix is obtained. The CDNN can be represented as

$$\widehat{\mathbf{T}(\mathbf{u})} = f_{CDNN}\left(\widehat{\mathbf{R}_y}\right) = f_p\left(f_7\left(f_6 \cdots f_1\left(\hat{\mathbf{r}}_y\right)\right)\right), \tag{11}$$

where $\{f_i(\bullet)\}_{i=1,2,\cdots,6}$ is followed by Complex ReLU (CReLU) as the activation function. In $\{f_i(\bullet)\}_{i=1,2,3,4}$, Complex LayerNorm (CLayerNorm) is inserted between the fully connected layers and CReLU for normalization. A Dropout layer with a dropout rate of 0.3 is applied after $f_2(\bullet)$ to mitigate overfitting. The CReLU and CLayerNorm functions can be expressed as

$$\mathrm{CRelu}(\mathbf{x}) = \mathrm{Relu}(\Re(\mathbf{x})) + j\mathrm{Relu}(\Im(\mathbf{x})), \tag{12}$$

$$\mathrm{CNorm}(\mathbf{x}) = \frac{\mathbf{x} - \mu}{\sqrt{\sigma^2 - \varepsilon}}, \tag{13}$$

where $\mathbf{x} = [x_1, x_2, \cdots, x_D]^T$ is the output of a certain layer in the network, $\mu = \frac{1}{D}\sum_{d=1}^{D} x_d$, $\sigma^2 = \frac{1}{D}\sum_{d=1}^{D} |x_d - \mu|^2$, and ε is a small value to prevent division by zero in (13).

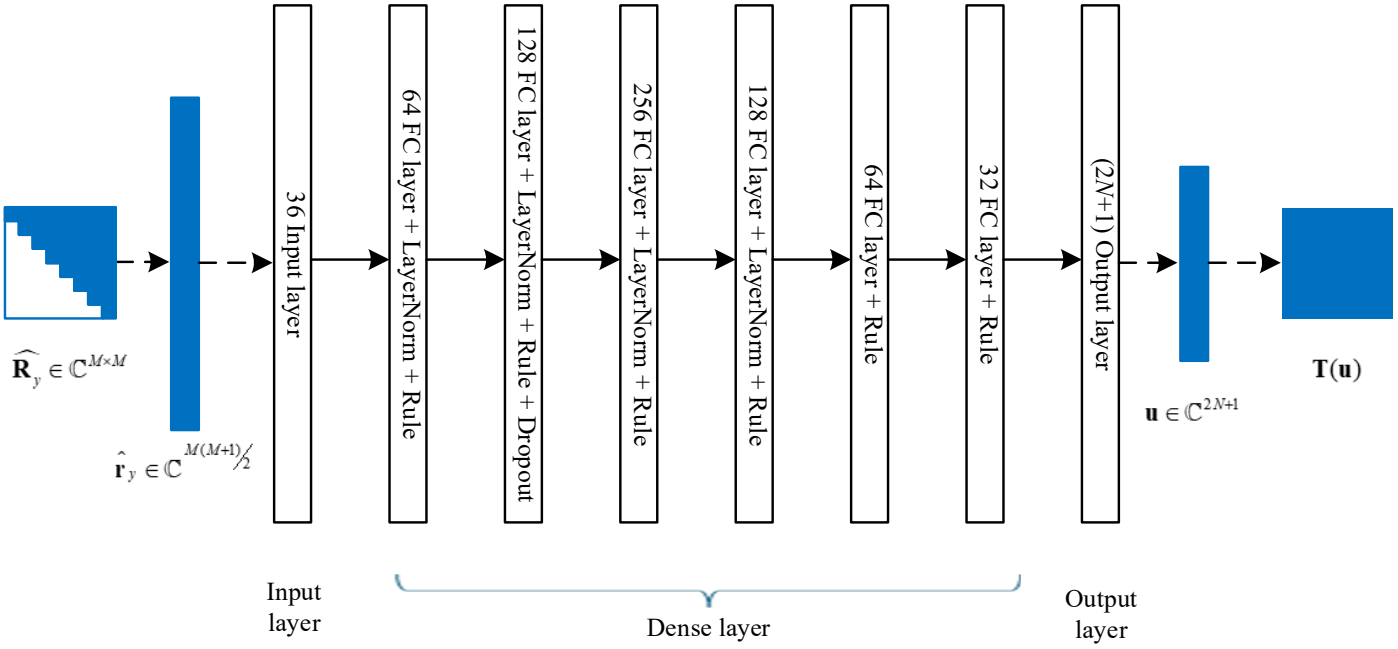


Figure 2. Angular-domain covariance matrix reconstruction network architecture.

Since $\mathbf{T}(\mathbf{u})$ is expressed in terms of power, we chose the mean absolute error between $f_{CDNN}\left(\widehat{\mathbf{R}_{y_i}}\right)$ and $\mathbf{T}(\mathbf{u})_i$ as the loss function during network training to ensure that the network's output closely approximates the true angular-domain covariance matrix. We use the Adam optimizer [43] to optimize the network parameters, with an initial learning rate of 0.01 and parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. Additionally, a 20-epoch warmup is introduced to stabilize network convergence. The training process adopts mini-batch gradient descent with a batch size of 1024. At the beginning of each epoch, the training data are shuffled to ensure that the distribution of each mini-batch is random. Then, the data are divided into multiple mini-batches, with each mini-batch undergoing forward propagation, loss computation, backpropagation, and parameter updates. To improve training efficiency

and prevent overfitting, we introduce ReduceLROnPlateau and Early Stopping techniques. If the loss on the validation set does not decrease within 10 epochs, ReduceLROnPlateau reduces the learning rate by half, allowing the optimizer to search the parameter space more finely. If the loss on the validation set does not improve for 20 consecutive epochs, Early Stopping terminates training to prevent overfitting on the training set.

Figure 3 illustrates the changes in training loss, validation loss, and learning rate over epochs for different values of $N = 4$, $N = 7$ and $N = 10$. Table 2 summarizes the training results of the three CDNN models, including the final losses on the training and validation sets, as well as the losses on the test set after training. Based on the results shown in Table 2 and Figure 3, it can be observed that all three CDNN models converge well and do not exhibit significant overfitting. Additionally, since we only varied the output dimension across these three CDNN models, we find that while all models achieve good training performance, the CDNN with a higher output dimension tends to exhibit larger average errors on the test set.

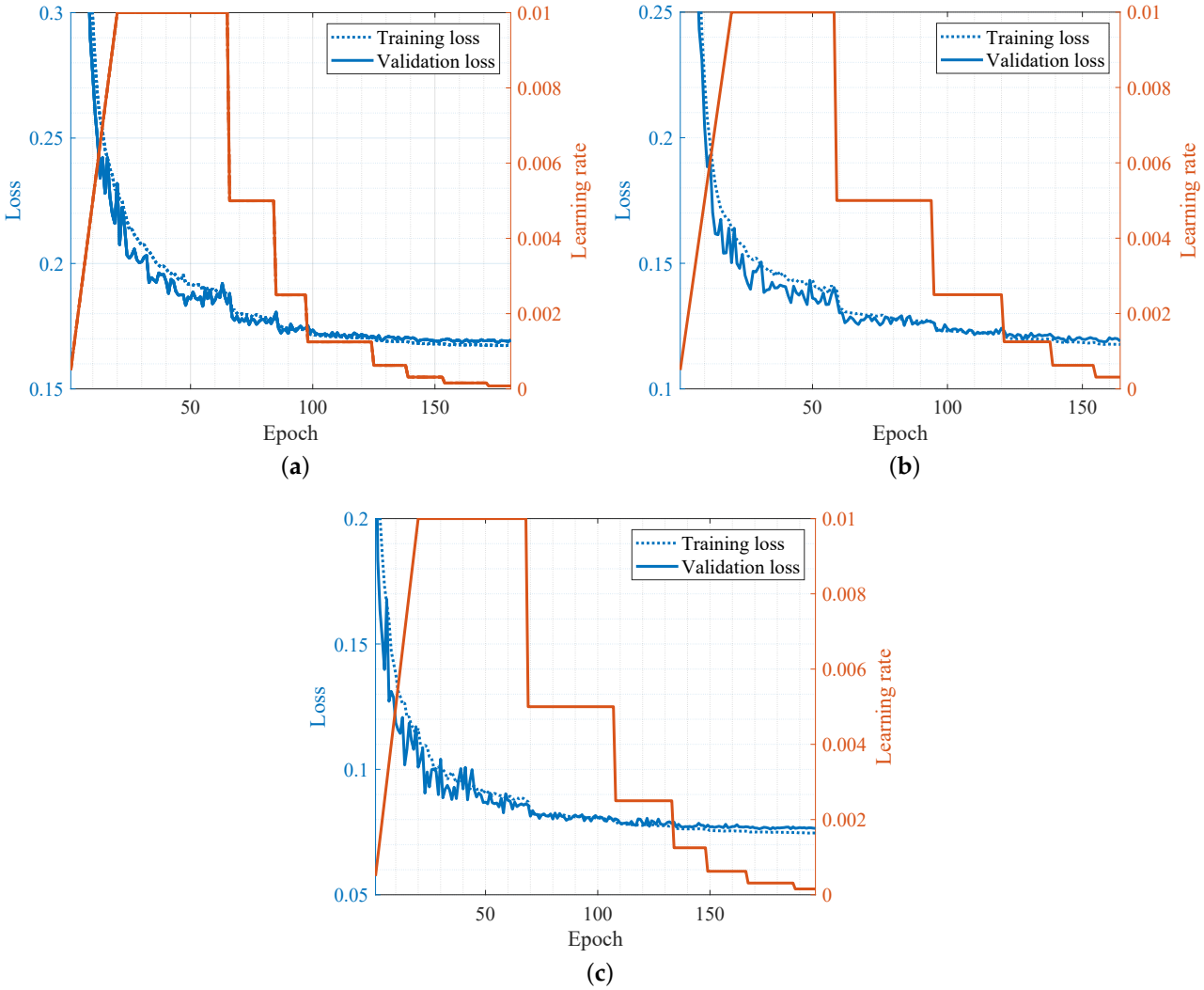


Figure 3. Training loss, validation loss, and learning rate variation with epochs for the following: (**a**) $N = 4$. (**b**) $N = 7$. (**c**) $N = 10$.

Table 2. CDNN training results.

	Training Loss	Validation Loss	Test Loss
$N = 4$	0.0750	0.0762	0.0766
$N = 7$	0.1185	0.1187	0.1197
$N = 10$	0.1676	0.1685	0.1692

4. Simulation

In this section, we will analyze the performance of our proposed method through numerical simulations and compare it with existing DOA estimation methods such as MUSIC [5], SPICE [12], and SBL [44]. First, we compare the gridless DOA estimation performance of CDNN models with truncation orders $N = 4$, $N = 7$ and $N = 10$ under different SNR conditions. Then, we present the DOA estimation results of these methods for varying numbers of sources within the 0° to 180° range. Finally, we evaluate the impact of the SNR, target angular separation, and the number of snapshots on the performance of the aforementioned algorithms in uncorrelated source scenarios through Monte Carlo simulations.

We consider K uncorrelated narrowband acoustic sources with equal power at a center frequency of $f = 50$ Hz, whose DOAs are $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_K]$, assuming that the number of sources is known. In the Monte Carlo simulations, the number of sources is set to $K = 2$. The performance of DOA estimation is evaluated using resolution probability (RP) and root mean square error (RMSE) [9,45,46]. The Cramer–Rao Bound (CRB) provides the lower bound of RMSE for unbiased estimators, serving as a reference for the DOA estimation algorithms studied in this paper [47]. For grid-based methods, DOA estimates are obtained from the peaks of the spatial spectrum. Specifically, the spatial spectrum obtained from the DOA estimation algorithm is first normalized, and then peaks with power greater than -10 dB are identified, where each peak corresponds to a target. If the algorithm identifies two targets and the DOA estimates $\widehat{\boldsymbol{\theta}} = [\widehat{\theta_1}, \widehat{\theta_2}]$ satisfy $\left|\widehat{\theta_1} - \theta_1\right| + \left|\widehat{\theta_2} - \theta_2\right| < |\theta_1 - \theta_2|$, the algorithm is considered to have successfully resolved the targets. To ensure the accuracy of the simulations, we only simulate the RMSE of the algorithm for cases where $RP \geq 0.5$. The RMSE of DOA estimation is defined as $\sqrt{\frac{1}{2T}\left[\sum_{t=1}^{T}\left(\widehat{\theta_{1t}} - \theta_1\right)^2 + \sum_{t=1}^{T}\left(\widehat{\theta_{2t}} - \theta_2\right)^2\right]}$, where T is the number of successful resolutions, and $\widehat{\boldsymbol{\theta}}_t = [\widehat{\theta_{1t}}, \widehat{\theta_{2t}}]$ is the estimate of $\boldsymbol{\theta}$ in the tth successful simulation. Additionally, the number of Monte Carlo simulations is set to 1000. For SPICE and SBL, the number of iterations is set to 10 and 1000, respectively. For the MUSIC algorithm, we divide the azimuth angle grid at 0.1° intervals, while for SPICE and SBL, the grid is divided at 1° intervals.

4.1. Impact of N on DOA Estimation Performance

Figure 4 illustrates the relationship between RMSE and RP with respect to the SNR for the gridless DOA estimation performance of CDNNs with truncation orders $N = 4$, $N = 7$, and $N = 10$ under different scenarios: $\boldsymbol{\theta} = [85^\circ, 95^\circ]$, $\boldsymbol{\theta} = [80^\circ, 100^\circ]$ and $\boldsymbol{\theta} = [70^\circ, 110^\circ]$. In the simulation, the noise is modeled as a zero-mean Gaussian white random process, assumed to be spatially and temporally uncorrelated. The SNR is defined as $\frac{{\alpha_1}^2 + {\alpha_2}^2}{\sigma^2}$, where ${\alpha_1}^2$, ${\alpha_2}^2$ and σ^2 represent the power of the two sources and the noise at the array elements, respectively. The number of snapshots is set to 50. As shown in Figure 4d,f, even when the angular separation between two targets is large and the SNR is high, the CDNN with $N = 10$ still occasionally fails to distinguish the two targets. Additionally, as depicted in Figure 4b, when the angular separation is small, the RP for $N = 10$ deteriorates more significantly as the SNR decreases. Conversely, as seen in Figure 4c,e, the CDNN with $N = 4$ exhibits higher robustness to Gaussian white noise when the target angular separations are 20° and 40°, providing better performance at low SNR levels while maintaining similar performance to $N = 7$ and $N = 10$ at higher SNRs. However, as shown in Figure 4a,

when the target angular separation decreases to 10° and the SNR exceeds 5 dB, the DOA estimation error for the CDNN with $N = 4$ becomes the largest. In summary, we observe that while $N = 10$ provides better resolution for closely spaced targets, its DOA estimation performance becomes unstable when the target angular separation is large and the SNR is high. The instability observed at $N = 10$ is due to the network needing to fit outputs of a higher dimensionality, which can lead to higher fitting errors. On the other hand, $N = 4$ delivers superior DOA estimation performance under conditions of larger angular separation and a lower SNR, but its resolution is limited when the angular separation is small. To balance the DOA estimation performance across varying angular separations, we choose the CDNN with $N = 7$, where the corresponding angular-domain covariance matrix $\mathbf{T}(\mathbf{u})$ is a 15×15 Toeplitz matrix.

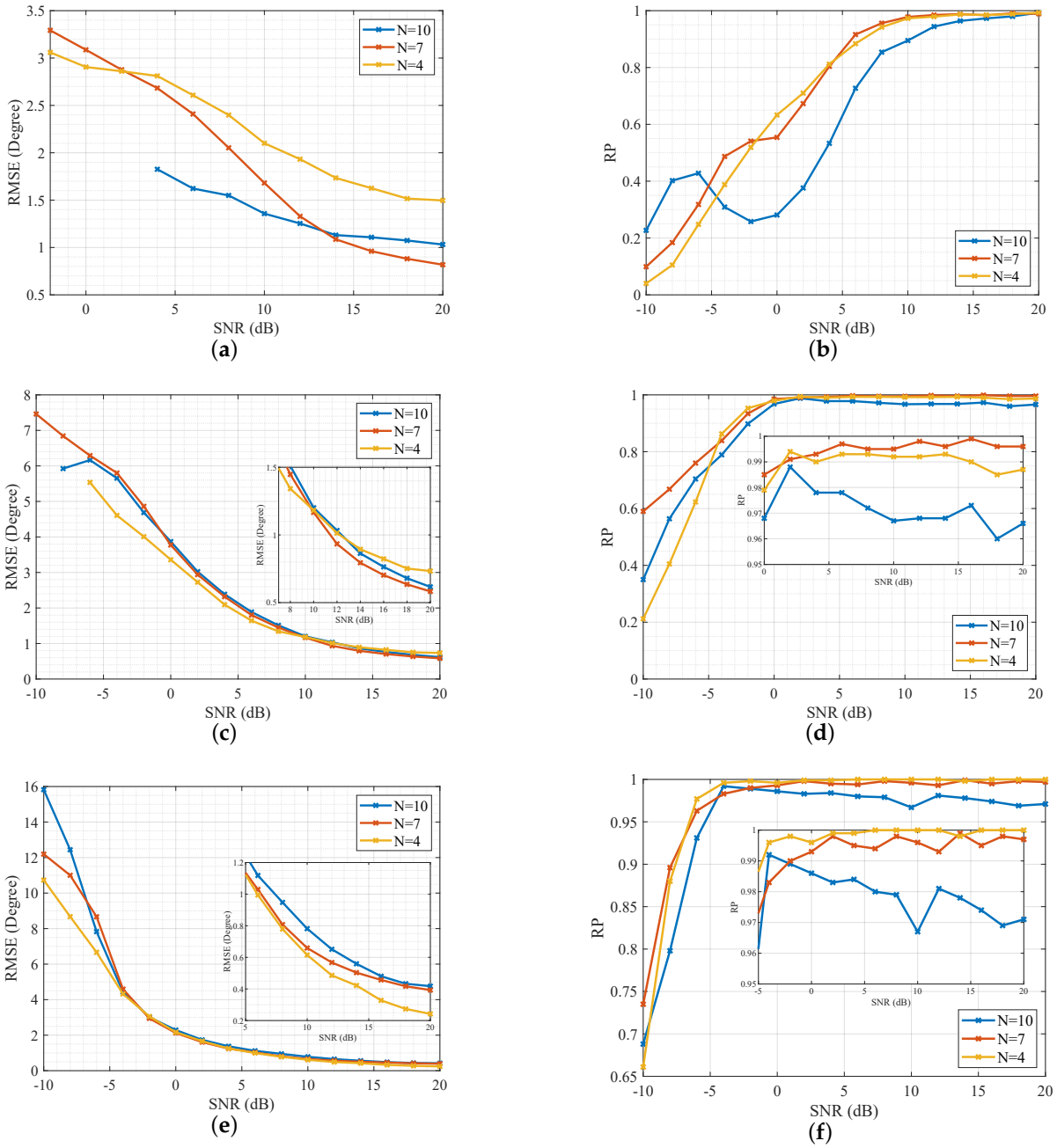


Figure 4. DOA estimation performance of CDNNs with truncation orders $N = 4$, $N = 7$, and $N = 10$ at different target angular separations. (**a**) RMSE and (**b**) RP for $\boldsymbol{\theta} = [85°, 95°]$. (**c**) RMSE and (**d**) RP for $\boldsymbol{\theta} = [80°, 100°]$. (**e**) RMSE and (**f**) RP for $\boldsymbol{\theta} = [70°, 110°]$.

4.2. DOA Estimation Results

In this section, we present the DOA estimation results using the methods under study for scenarios with $K \in \{1, 2, 3\}$ sources within an angular range of from 0° to 180°. When the number of sources is 2 or 3, we set the angular separation between adjacent sources as $\Delta\theta = 40°$, meaning that all source DOAs are $\boldsymbol{\theta} = [\theta_1, \theta_1 + \Delta\theta, \cdots, \theta_1 + (K-1)\Delta\theta]$. We consider the case where $\boldsymbol{\theta}$ satisfying $\theta_1 \in [0° : 0.1° : 180° - (K-1)\Delta\theta]$, with the SNR set to 10 dB, and with the noise being spatiotemporally uncorrelated Gaussian white noise. The number of snapshots is set to 50. The DOA estimation results for various methods are shown in Figure 5, and the RMSE for different angular ranges is provided in Table 3. When the number of sources $K = 1$, the results in Table 3 indicate that the proposed method and MUSIC achieve more accurate DOA estimates than SPICE and SBL. Comparing Figure 5a,d,g,j, it can be seen that even when dividing the grid densely with a separation of 1°, grid mismatch errors still severely affect the performance of SPICE and SBL. When the number of sources $K = 2$ and $K = 3$, and all sources are within the angular range $[45°, 135°]$, all four methods can obtain relatively accurate DOA estimation results. However, when the range of source DOAs is confined to $[0°, 180°]$, the proposed method shows a significant RMSE advantage over the other three methods. According to the simulation results shown in Figure 5, when all sources are far from the end-fire direction, all four methods can effectively distinguish targets and achieve accurate DOA estimation. However, when a source is close to the end-fire direction, the DOA estimation performance of the other three methods deteriorates significantly, making it difficult to distinguish adjacent targets or leading to large DOA estimation errors. In contrast, the performance degradation of the proposed method is relatively mild.

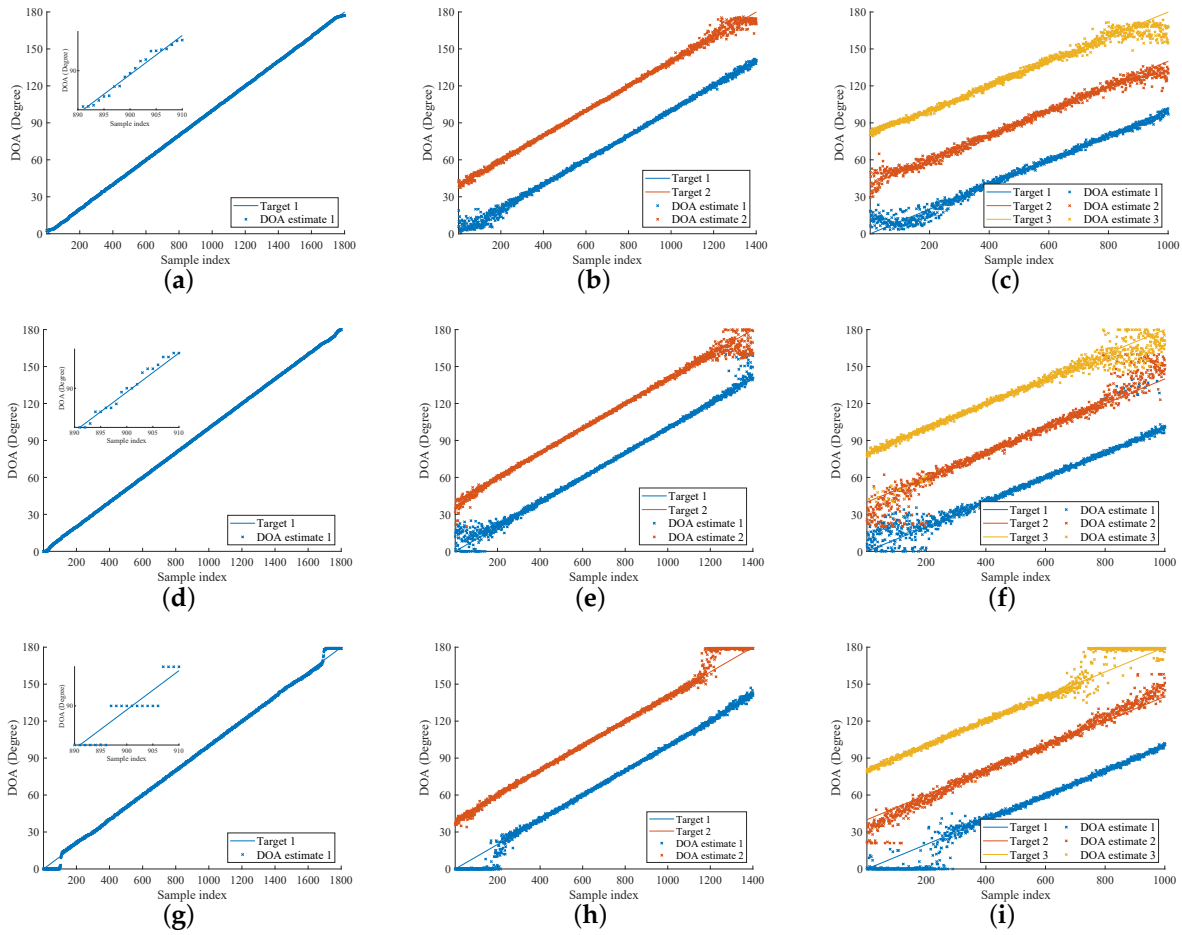


Figure 5. *Cont.*

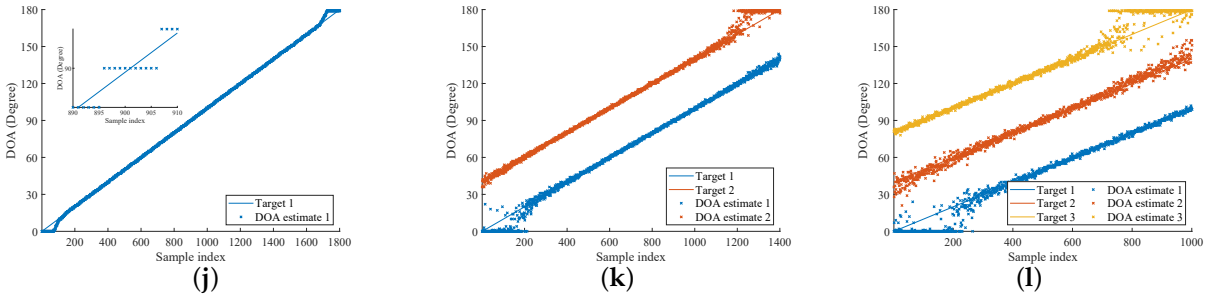


Figure 5. DOA estimation results. The proposed method, source numbers (**a**) $K = 1$, (**b**) $K = 2$, and (**c**) $K = 3$. MUSIC, source numbers (**d**) $K = 1$, (**e**) $K = 2$, and (**f**) $K = 3$. SPICE, source numbers (**g**) $K = 1$, (**h**) $K = 2$, and (**i**) $K = 3$. SBL, source numbers (**j**) $K = 1$, (**k**) $K = 2$, and (**l**) $K = 3$.

Table 3. RMSE of DOA estimation.

K	The Range of $\theta_k, \forall k = 1, 2, \ldots, K$	Proposed Method	MUSIC	SPICE	SBL
1	$[45°, 135°]$	0.1278	0.1098	0.2676	0.2678
	$[0°, 180°]$	0.2239	0.1885	1.0060	0.6724
2	$[45°, 135°]$	0.9051	0.8279	0.9340	0.9146
	$[0°, 180°]$	2.0294	2.9906	3.9232	3.2688
3	$[45°, 135°]$	2.4844	2.2639	2.1821	2.0246
	$[0°, 180°]$	5.0632	9.1194	8.6687	7.1800

4.3. Relationship between Algorithm Performance and SNR

We analyze the relationship between algorithm performance and the SNR under two scenarios: $\boldsymbol{\theta} = [80°, 100°]$ and $\boldsymbol{\theta} = [70°, 110°]$. The other definitions and settings in the simulation are the same as those in Section 4.1, and the simulation results are presented in Figure 6. When $\boldsymbol{\theta} = [80°, 100°]$, as shown in Figure 6a, the proposed method achieves a lower RMSE than the other three methods when the SNR is below 14 dB. Additionally, as illustrated in Figure 6b, the proposed method maintains an RP above 0.6 for distinguishing the two sources across the SNR range of -10 to 20 dB. In contrast, the RP of the other three methods rapidly decreases with a decreasing SNR, falling below 0.2 at an SNR of 0 dB, indicating that they almost fail to distinguish the two sources. For the scenario with $\boldsymbol{\theta} = [70°, 110°]$, Figure 6c,d show that in the SNR range of -2 to 12 dB, the proposed method achieves comparable or lower RMSE compared to the other methods, while all methods maintain nearly perfect RP values close to 1, successfully distinguishing the two targets. When the SNR drops below -4 dB, although the RMSE of the proposed method exceeds that of SBL and MUSIC, it still achieves the highest RP for source resolution, with RP remaining above 0.7 even when the SNR decreases to -10 dB.

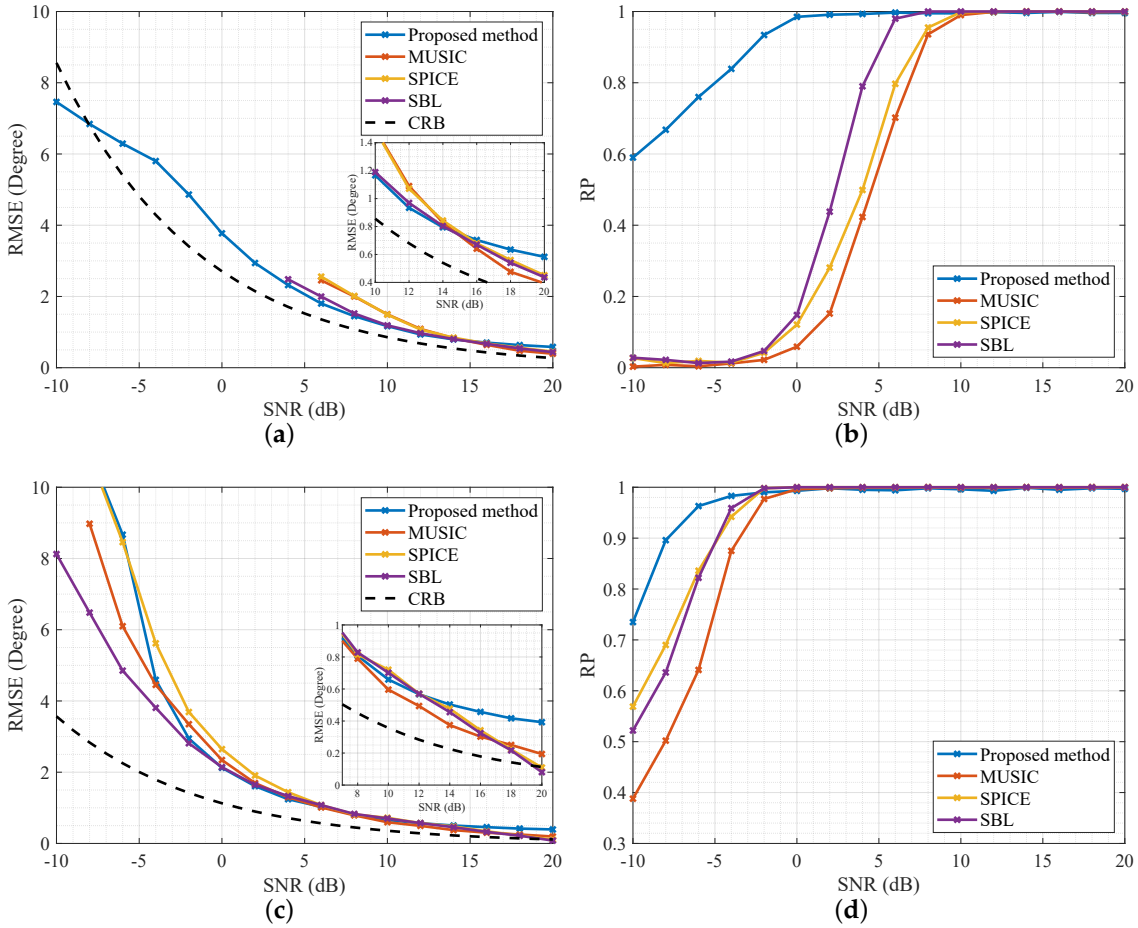


Figure 6. Relationship between SNR and both RMSE and RP under spatio-temporal Gaussian white noise conditions.(**a**) RMSE and (**b**) RP for $\boldsymbol{\theta} = [80°, 100°]$. (**c**) RMSE and (**d**) RP for $\boldsymbol{\theta} = [70°, 110°]$.

4.4. Relationship between Algorithm Performance and Angle Separation

We set $\theta_1 = 90° - \Delta\theta/2$, $\theta_2 = 90° + \Delta\theta/2$, where $\Delta\theta$ represents the angular separation between the two sources. The SNR in the simulation is set to 10 dB, and other conditions remain the same as in Section 4.1. The relationship between the angle separation and the DOA estimation performance, as measured by RP and RMSE, is shown in Figure 7. When the angular separation between the two sources is between 20° and 50°, the proposed method shows similar RMSE values compared to the other three methods, with the RMSE difference among the four methods being less than 0.2°. Additionally, all four methods are able to distinguish between the two sources with an RP close to 1. However, when the angular separation between the two sources is less than 20°, the proposed method exhibits the lowest DOA estimation RMSE and can still differentiate between the two sources with an RP close to 1. In contrast, the RP of the other three methods decreases rapidly with decreasing angular separation. When the angular separation is reduced to 10°, the RP of all three existing methods drops below 0.2, making them almost unable to distinguish between the sources. The simulation results demonstrate that the proposed method outperforms the existing methods in DOA estimation when the angular separation is small.

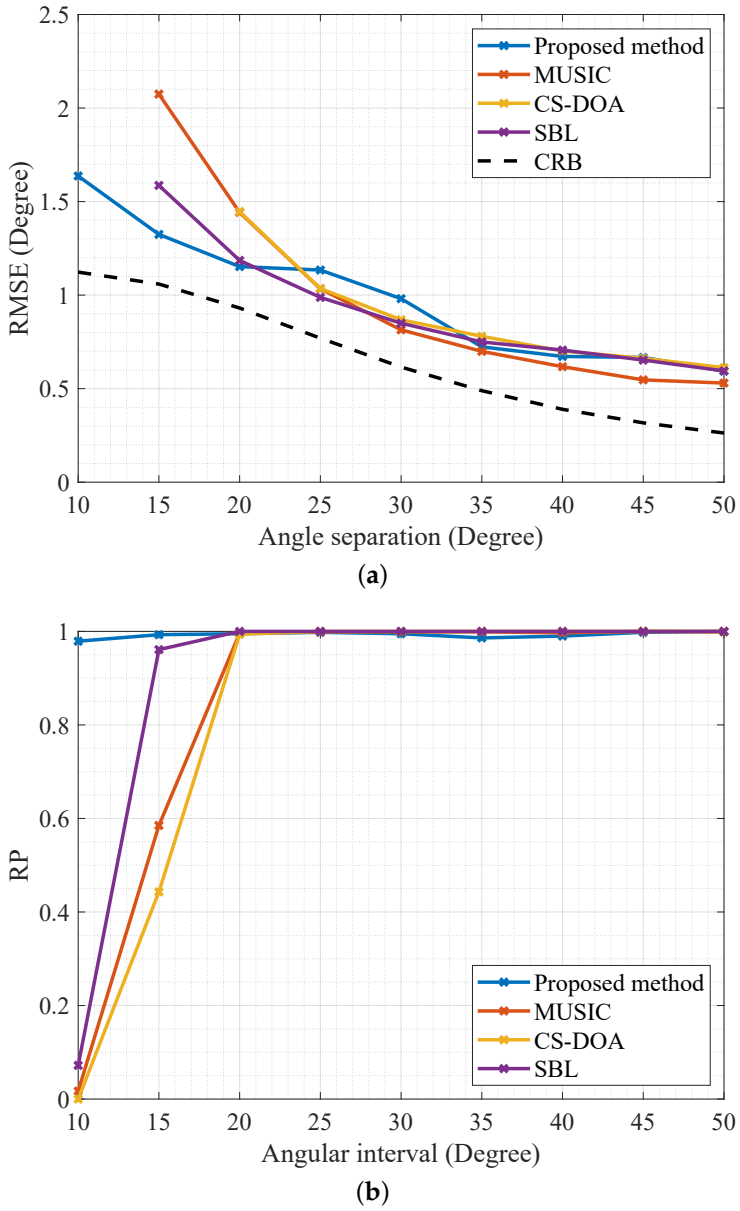


Figure 7. Relationship between RP and RMSE with respect to angle separation. (**a**) RMSE. (**b**) RP.

4.5. Impact of Snapshot Number on Algorithm Performance

In this section, we analyze the impact of the number of snapshots on the algorithm's performance. Figure 8a,b illustrate the RMSE and RP of the proposed method and the three existing methods for DOA estimation under different snapshot numbers. In the simulation, we set $\boldsymbol{\theta} = [78^\circ, 102^\circ]$ and the SNR to 10 dB, with other conditions being the same as in Section 4.1. As shown in Figure 8, under the simulation conditions, when the number of snapshots exceeds 50, the proposed method exhibits the highest RMSE, while all four methods are able to distinguish the two sources with an RP close to 1. When the number of snapshots is less than 30, the RMSE of MUSIC and SPICE increases, and their RP decreases rapidly as the number of snapshots decreases. When the number of snapshots falls below 10, the RP of MUSIC and SPICE drops below 0.5, indicating a high probability of failure in accurate DOA estimation. At this point, the proposed method can still distinguish the two sources with an RP close to 1, and it also achieves lower RMSE compared to the existing three methods. Based on the simulation results, we conclude that the proposed method outperforms the other three existing methods when the number of snapshots is low, demonstrating higher adaptability to fewer snapshots.

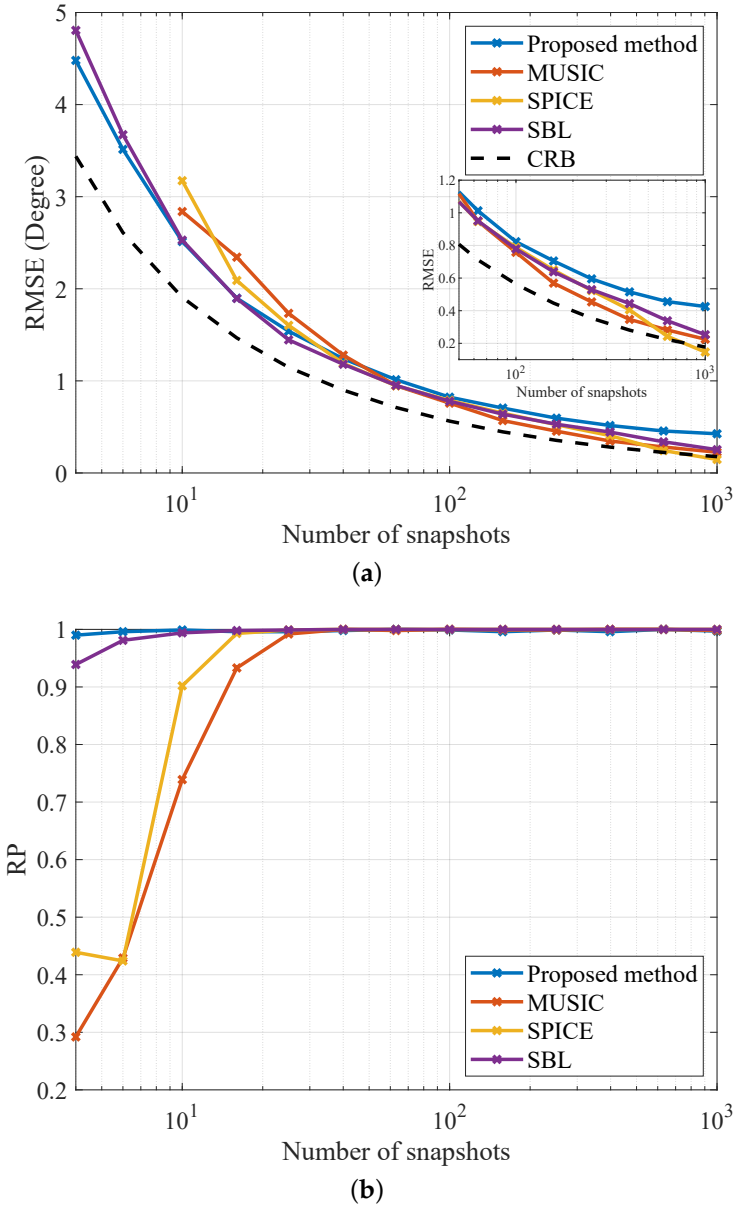


Figure 8. Algorithm performance under different snapshot conditions. (**a**) RMSE. (**b**) RP.

5. Swellex-96 Event S59 Experimental Results

In this section, we validate the effectiveness of the proposed method using experimental data from the Swellex-96 Event S59. The SWellEx-96 experiment [38] was conducted in shallow waters approximately 12 km off the tip of Point Loma near San Diego, California, from 10 to 18 May 1996. As shown in Figure 9, the event scenario depicts the source's trajectory in blue and the interferer's trajectory in red. The markers on the tracks represent GPS-recorded coordinates at 5-minute intervals, and the entire experiment lasted for 65 min. During the experiment, the source continuously emitted signals at various frequencies, with the higher power frequencies being 49, 64, 79, 94, 112, 130, 148, 166, 201, 235, 283, 338, and 388 Hz. The HLA subarray used in our study had a sampling frequency of $f_s = 3276.8$ Hz, and the subarray formed by elements #1 to #8 effectively met the far-field condition during the experiment.

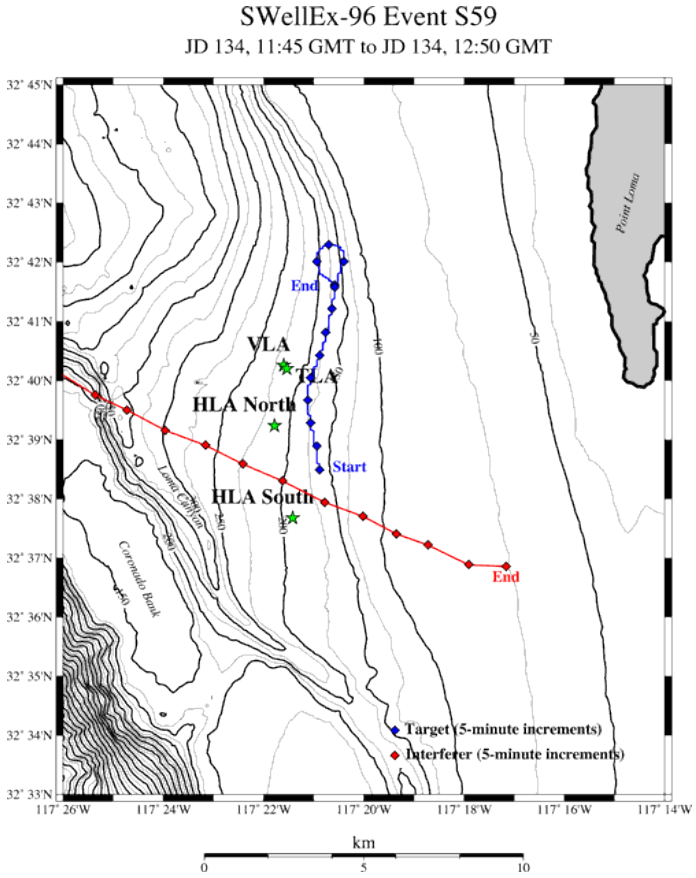


Figure 9. Schematic of the Swellex-96 Event S59 experiment scenario [38].

The data received by the 8-element subarray are denoted as $\mathbf{Y} = [\mathbf{Y}(:,1), \mathbf{Y}(:,2), \cdots, \mathbf{Y}(:,8)]$, where the column vector $\mathbf{Y}(:,m)$ represents the data received by the mth array element. We divided the data into $Q = 361$ segments, each starting at time $t_{start} \in \{t_1, t_2, \cdots, t_Q\} = [0s : 10s : 3600s]$ and lasting for a duration $T_{seg} = 30s$. The data received by the mth array element during the qth segment can be expressed as $\bar{\mathbf{y}}_{mq} = \mathbf{Y}\big(\lfloor f_s t_q \rfloor + 1 : \lfloor f_s (t_q + T_{seg}) \rfloor, m\big)$, where $\lfloor \bullet \rfloor$ denotes the floor function. Each segment was further divided into $L = 100$ snapshots, with each snapshot containing $N_{snap} = 2048$ sampling points. The duration of each snapshot was $T_{snap} = 0.625$ s, and the overlap between consecutive snapshots was $\eta = 53\%$. Therefore, the lth snapshot within $\bar{\mathbf{y}}_{mq}$ is represented as $\tilde{\mathbf{y}}_{mql} = \bar{\mathbf{y}}_{mq}\big(\lfloor (l-1)(1-\eta) f_s T_{snap} \rfloor + 1 : \lfloor (l-1)(1-\eta) f_s T_{snap} \rfloor + N_{snap}\big)$. We will apply different DOA estimation methods based on the snapshot model described above to perform narrowband DOA estimation at a frequency of $f_c = 50$ Hz for each data segment, thereby obtaining the DOA trajectories.

Figure 10 presents the bearing time records (BTRs) obtained using different DOA estimation methods based on the Swellex-96 Event S59 data, along with the GPS-recorded tracks of the source and interferer. As shown in Figure 10b, for CBF, prior to 1800 s, due to the small angular separation between the source and the interferer, the two targets cannot be distinguished. After 2400 s, although the angular separation between the source and interferer increases and two peaks appear in the CBF spectrum, high sidelobes severely affect the performance of DOA estimation. As shown in Figure 10d–f, the three existing methods demonstrate better resolution compared to CBF, allowing the source and interferer tracks to be distinguished. However, all three methods exhibit high levels of pseudospectra around the DOA track, which negatively impact target trajectory identification. Moreover, for SPICE and SBL, the interferer's DOA is close to the end-fire direction of the array,

making it difficult to identify the interferer prior to 1800 s in most cases. As shown in Figure 10c, the proposed method directly achieves gridless DOA estimation, avoiding the influence of pseudospectra on BTR trajectories. Throughout the entire BTR process, the proposed method consistently identifies both the source and interferer with a high probability, making it more effective for trajectory determination. Figure 11 illustrates the results of processing the Swellex-96 Event S59 data using different methods. The RMSE was calculated by interpolating the source trajectory with its GPS angles, and the data were processed using an AMD Ryzen 76800H CPU. The results show that the proposed method significantly outperforms the other three methods in terms of RP while maintaining RMSE values similar to the other methods. Additionally, the proposed method exhibits complexity comparable to MUSIC and is significantly more efficient than SPICE and SBL, which require iterative operations.

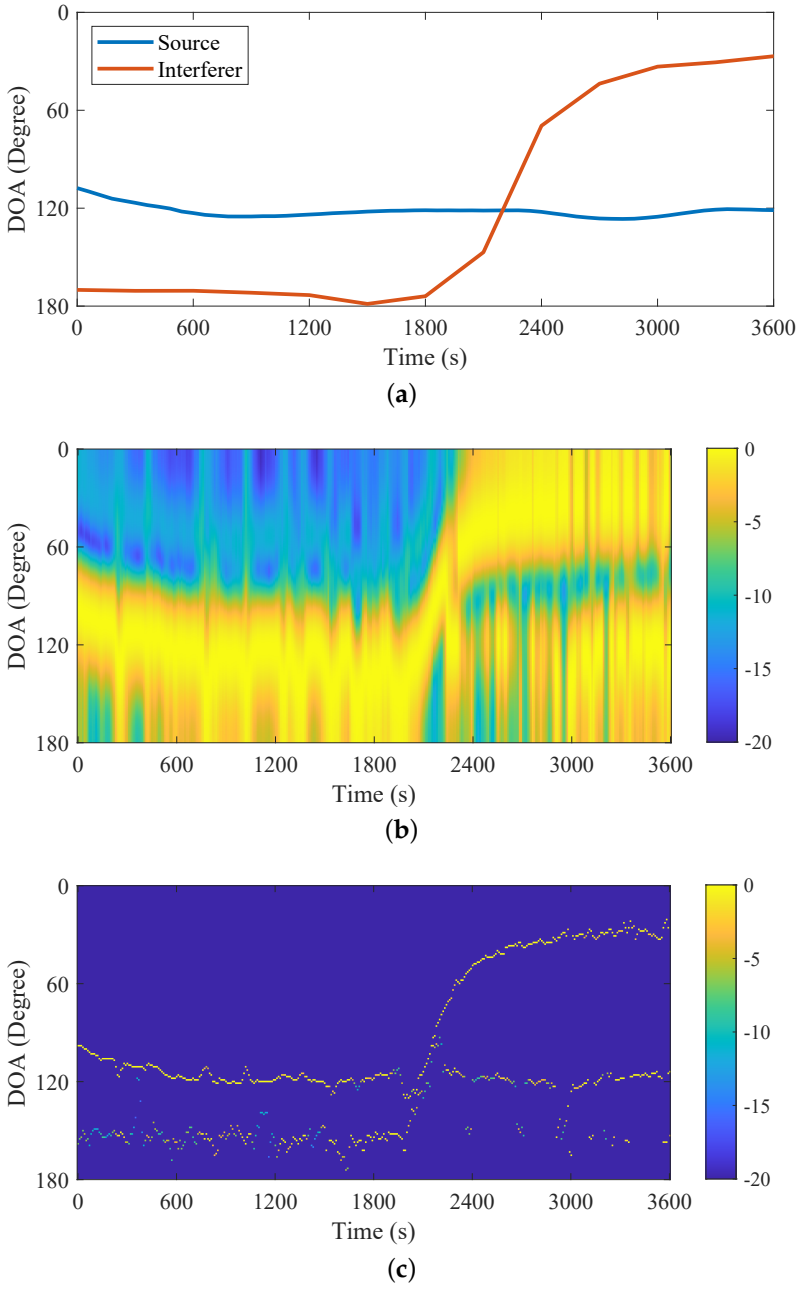


Figure 10. *Cont.*

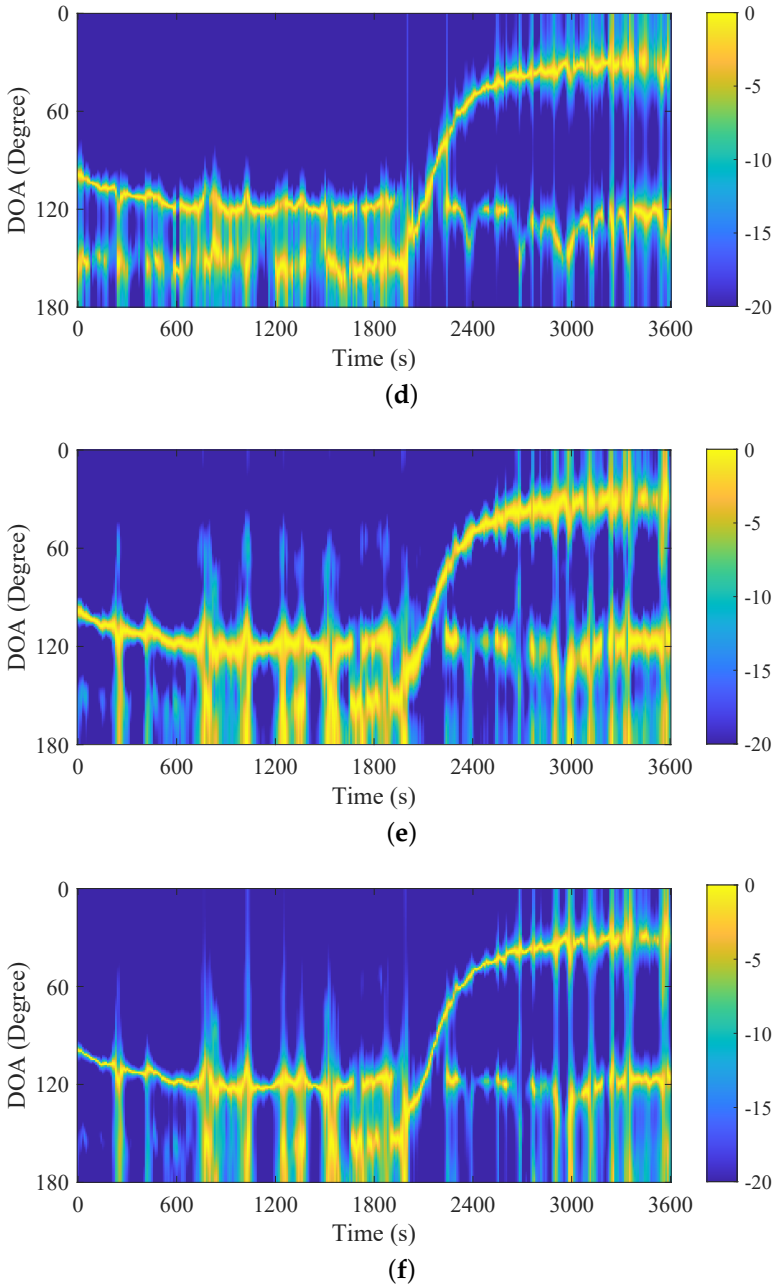


Figure 10. BTR results using different methods. (**a**) GPS. (**b**) CBF. (**c**) Proposed method. (**d**) MUSIC. (**e**) SPICE. (**f**) SBL.

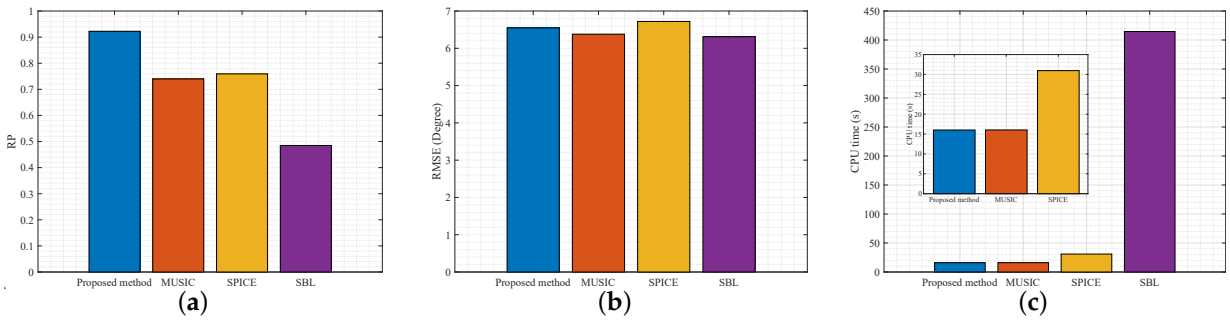


Figure 11. Processing results of Swellex-96 Event S59 data using different methods. (**a**) RP. (**b**) RMSE. (**c**) CPU time.

6. Conclusions and Discussion

In this paper, we propose a novel gridless DOA estimation method for arbitrary array geometries based on a CDNN. First, by expanding the steering vector into a Fourier series, the relationship between the element-domain covariance matrix and the angular-domain-covariance matrix for arbitrary arrays is derived. Then, a CDNN is designed and trained to reconstruct the Toeplitz-structured angular-domain covariance matrix from the SCM. Finally, Root-MUSIC is applied to the reconstructed angular-domain covariance matrix to obtain gridless DOA estimates. Our proposed method demonstrates superior resolution and DOA estimation accuracy compared to existing methods, particularly when target angles are closely spaced. The proposed method does not require the setting of any hyperparameters and exhibits high adaptability to different snapshot counts. Additionally, its lower computational complexity makes it well suited for applications that demand real-time performance or deployment on low-power platforms with limited computational resources. In this paper, we only analyze the performance of the proposed method for single-frequency narrowband signals. In our future work, we will continue to optimize the network structure and training process to make the proposed method applicable to multiple frequencies without needing prior information about the number of sources, thereby enhancing its practical utility in real-world scenarios, such as providing accurate angular estimates for localization and object tracking.

Author Contributions: Conceptualization, Y.C. and T.Z.; methodology, T.Z.; writing—review and editing, Y.C., T.Z. and Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (62371393, 62271397) and the Fundamental Research Funds for the Central Universities (D5000240231).

Data Availability Statement: Data are contained within this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Godara, L.C. Application of antenna arrays to mobile communications. II. Beam-forming and direction-of-arrival considerations. *Proc. IEEE* **1997**, *85*, 1195–1245. [CrossRef]
2. Massa, A.; Rocca, P.; Oliveri, G. Compressive Sensing in Electromagnetics—A Review. *IEEE Antennas Propag. Mag.* **2015**, *57*, 224–238. [CrossRef]
3. Trees, H.L.V. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*; Wiley: Hoboken, NJ, USA, 2004; pp. 17–79.
4. Capon, J. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* **1969**, *57*, 1408–1418. [CrossRef]
5. Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280. [CrossRef]
6. Roy, R.; Paulraj, A.; Kailath, T. ESPRIT—A subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Trans. Acoust., Speech, Signal Process.* **1986**, *34*, 1340–1342. [CrossRef]
7. Thomas, J.K.; Scharf, L.L.; Tufts, D.W. The probability of a subspace swap in the SVD. *IEEE Trans. Signal Process.* **1995**, *43*, 730–736. [CrossRef]
8. Yang, Z.; Li, J.; Stoica, P.; Xie, L. Chapter 11 - Sparse methods for direction-of-arrival estimation. In *Array, Radar and Communications Engineering*; Chellappa, R., Theodoridis, S., Eds.; Academic Press: Cambridge, MA, USA, 2018; pp. 509–581.
9. Zhang, G.; Liu, K.; Sun, S.; Fu, J.; Wang, J. DOA estimation method for underwater acoustic signals based on two-dimensional power distribution (TPD) for few element array. *Appl. Acoust.* **2021**, *184*, 108352. [CrossRef]
10. Zhang, Y.; Zhang, L.; Han, J.; Ban, Z.; Yang, Y. A new DOA estimation algorithm based on compressed sensing. *Cluster Comput.* **2019**, *22*, 895–903.
11. Li, X.; Ma, X.; Yan, S.; Hou, C. Single snapshot DOA estimation by compressive sampling. *Appl. Acoust.* **2013**, *74*, 926–930. [CrossRef]
12. Stoica, P.; Babu, P.; Li, J. SPICE: A Sparse Covariance-Based Estimation Method for Array Processing. *IEEE Trans. Signal Process.* **2011**, *59*, 629–638. [CrossRef]
13. Tan, Z.; Yang, P.; Nehorai, A. Joint Sparse Recovery Method for Compressed Sensing with Structured Dictionary Mismatches. *IEEE Trans. Signal Process.* **2014**, *62*, 4997–5008. [CrossRef]
14. Wu, X.; Zhu, W.; Yan, J. Direction of Arrival Estimation for Off-Grid Signals Based on Sparse Bayesian Learning. *IEEE Sens. J.* **2014**, *16*, 2004–2016. [CrossRef]

15. Wang, Q.; Yu, H.; Li, J.; Chen, F. Adaptive Grid Refinement Method for DOA Estimation via Sparse Bayesian Learning. *IEEE J. Ocean Eng.* **2023**, *48*, 806–819. [CrossRef]
16. Jagannath, R.; Hari, K.V.S. Block Sparse Estimator for Grid Matching in Single Snapshot DoA Estimation. *IEEE Signal Process. Lett.* **2013**, *11*, 1038–1041. [CrossRef]
17. Xenaki, A.; Gerstoft, P. Grid-free compressive beamforming. *J. Acoust. Soc. Am.* **2015**, *137*, 1923–1935. [CrossRef]
18. Zhou, C.; Gu, Y.; Fan, X.; Shi, Z.; Mao, G.; Zhang, Y. Direction-of-Arrival Estimation for Coprime Array via Virtual Array Interpolation. *IEEE Trans. Signal Process.* **2018**, *66*, 5956–5971. [CrossRef]
19. Chu, Z.; Liu, Y.; Yang, Y.; Yang, Y. A preliminary study on two-dimensional grid-free compressive beamforming for arbitrary planar array geometries. *J. Acoust. Soc. Am.* **2021**, *149*, 3751–3757. [CrossRef]
20. Raj, A.G.; McClellan, J.H. Super-resolution DOA Estimation for Arbitrary Array Geometries Using a Single Noisy Snapshot. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019.
21. Raj, A.G.; McClellan, J.H. Single Snapshot Super-Resolution DOA Estimation for Arbitrary Array Geometries. *IEEE Signal Process. Lett.* **2019**, *26*, 119–123.
22. Yang, Y.; Chu, Z.; Yin, S. Two-dimensional grid-free compressive beamforming with spherical microphone arrays. *Mech. Syst. Signal Proc.* **2022**, *169*, 108642. [CrossRef]
23. Mahata, K.; Hyder, M.M. Grid-less T.V minimization for DOA estimation, Signal Processing. *Signal Process.* **2017**, *132*, 155–164. [CrossRef]
24. Grumiaux, P.; Kitić, S.; Girin, L.; Guérin, A. A survey of sound source localization with deep learning methods. *J. Acoust. Soc. Am.* **2022**, *152*, 107–151. [CrossRef]
25. Papageorgiou, G.K.; Sellathurai, M.; Eldar, Y.C. Deep Networks for Direction-of-Arrival Estimation in Low SNR. *IEEE Trans. Signal Process.* **2021**, *69*, 3714–3729. [CrossRef]
26. Liu, Y.; Chen, H.; Wang, B. DOA estimation based on CNN for underwater acoustic array. *Appl. Acoust.* **2021**, *172*, 107594. [CrossRef]
27. Nie, W.; Zhang, X.; Xu, J.; Guo, L.; Yan, Y. Adaptive Direction-of-Arrival Estimation Using Deep Neural Network in Marine Acoustic Environment. *IEEE Sens. J.* **2023**, *23*, 15093–15105. [CrossRef]
28. Liang, C.; Liu, M.; Li, Y.; Wang, Y.; Hu, X. LDnADMM-Net: A Denoising Unfolded Deep Neural Network for Direction-of-Arrival Estimations in A Low Signal-to-Noise Ratio. *Remote Sens.* **2024**, *16*, 554. [CrossRef]
29. Wu, X.; Yang, X.; Jia, X.; Tian, F. A Gridless DOA Estimation Method Based on Convolutional Neural Network with Toeplitz Prior. *IEEE Signal Process. Lett.* **2022**, *29*, 1247–1251. [CrossRef]
30. Gao, S.; Ma, H.; Liu, H.; Yang, J.; Yang, Y. A Gridless DOA Estimation Method for Sparse Sensor Array. *Remote Sens.* **2023**, *15*, 5281. [CrossRef]
31. Wu, X.; Wang, J.; Yang, X.; Tian, F. A Gridless DOA Estimation Method Based on Residual Attention Network and Transfer Learning. *IEEE Trans. Veh. Technol.* **2024**, *73*, 9103–9108. [CrossRef]
32. Cui, Y.; Yang, F.; Zhou, M.; Hao, L.; Wang, J.; Sun, H.; Kong, A.; Qi, J. Gridless Underdetermined DOA Estimation for Mobile Agents with Limited Snapshots Based on Deep Convolutional Generative Adversarial Network. *Remote Sens.* **2024**, *16*, 626. [CrossRef]
33. Huang, G.; Benesty, J.; Chen, J. On the Design of Frequency-Invariant Beampatterns with Uniform Circular Microphone Arrays. *IEEE-ACM Trans. Audio Speech Lang.* **2017**, *25*, 1140–1153. [CrossRef]
34. Tang, G.; Bhaskar, B.N.; Shah, P.; Recht, B. Compressed Sensing Off the Grid. *IEEE Trans. Inf. Theory* **2013**, *59*, 7465–7490. [CrossRef]
35. Wu, X.; Zhu, W.; Yan, J. A Toeplitz Covariance Matrix Reconstruction Approach for Direction-of-Arrival Estimation. *IEEE Trans. Veh. Technol.* **2017**, *66*, 8223–8237. [CrossRef]
36. Pal, P.; Vaidyanathan, P.P. A Grid-Less Approach to Underdetermined Direction of Arrival Estimation Via Low Rank Matrix Denoising. *IEEE Signal Process. Lett.* **2014**, *21*, 737–741. [CrossRef]
37. Rao, B.D.; Hari, K.V.S. Performance analysis of Root-Music. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 1939–1949. [CrossRef]
38. The SWellEX-96 Experiment. 1996. Available online: http://swellex96.ucsd.edu/ (accessed on 5 September 2024).
39. Trabelsi, C.; Bilaniuk, O.; Serdyuk, D.; Subramanian, S.; Santos, J.F.; Mehri, S.; Rostamzadeh, N.; Bengio, Y.; Pal, C.J. Deep Complex Networks. *arXiv* **2017**, arXiv:1705.09792.
40. Mohammadzadeh, S.; Nascimento, V.H.; Lamare, R.C.; Hajarolasvadi, N. Robust Beamforming Based on Complex-Valued Convolutional Neural Networks for Sensor Arrays. *IEEE Signal Process. Lett.* **2022**, *29*, 2108–2112. [CrossRef]
41. Zhang, Y.; Zeng, R.; Zhang, S.; Wang, J.; Wu, Y. Complex-Valued Neural Network with Multistep Training for Single-Snapshot DOA Estimation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
42. Fan, Z.; Tu, Y.; Lin, Y.; Shi, Q. Class-Incremental Learning for Recognition of Complex-Valued Signals. *IEEE Trans. Cogn. Commun. Netw.* **2024**, *10*, 417–428. [CrossRef]
43. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
44. Xenaki, A.; Boldt, J.B.; Christensen, M.G. Sound source localization and speech enhancement with sparse Bayesian learning beamforming. *J. Acoust. Soc. Am.* **2018**, *143*, 2912–3921. [CrossRef]

45. Zhou, T.; He, Z.; Shi, Q.; Lin, C.; Zhang, S. Multisnapshot High-Resolution Gridless DOA Estimation for Uniform Circular Arrays. *IEEE Signal Process. Lett.* **2024**, *31*, 1705–1709. [CrossRef]
46. Liu, Z.; Zhang, Y.; Wang, W.; Li, X.; Li, H.; Shi, W.; Ali, W. Infinite Weighted p-Norm Sparse Iterative DOA Estimation via Acoustic Vector Sensor Array under Impulsive Noise. *J. Mar. Sci. Eng.* **2023**, *11*, 1798. [CrossRef]
47. Stoica, P.; Nehorai, A. MUSIC, maximum likelihood, and Cramer-Rao bound. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 720–741. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

Dual-Feature Fusion Learning: An Acoustic Signal Recognition Method for Marine Mammals

Zhichao Lü [1], Yaqian Shi [1], Liangang Lü [2], Dongyue Han [1], Zhengkai Wang [1] and Fei Yu [1,*]

[1] College of Ocean Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China; lvzhichao@hrbeu.edu.cn (Z.L.); shiyaqian@sdust.edu.cn (Y.S.); handongyue@sdust.edu.cn (D.H.); wzk18724791626@sdust.edu.cn (Z.W.)

[2] First Institute of Oceanography, Ministry of Natural Resources, Qingdao 266061, China; lvlg@fio.org.cn

* Correspondence: feiyu@sdust.edu.cn

Abstract: Marine mammal acoustic signal recognition is a key technology for species conservation and ecological environment monitoring. Aiming at the complex and changing marine environment, and because the traditional recognition method based on a single feature input has the problems of poor environmental adaptability and low recognition accuracy, this paper proposes a dual-feature fusion learning method. First, dual-domain feature extraction is performed on marine mammal acoustic signals to overcome the limitations of single feature input methods by interacting feature information between the time-frequency domain and the Delay-Doppler domain. Second, this paper constructs a dual-feature fusion learning target recognition model, which improves the generalization ability and robustness of mammal acoustic signal recognition in complex marine environments. Finally, the feasibility and effectiveness of the dual-feature fusion learning target recognition model are verified in this study by using the acoustic datasets of three marine mammals, namely, the Fraser's Dolphin, the Spinner Dolphin, and the Long-Finned Pilot Whale. The dual-feature fusion learning target recognition model improved the accuracy of the training set by 3% to 6% and 20% to 23%, and the accuracy of the test set by 1% to 3% and 25% to 38%, respectively, compared to the model that used the time-frequency domain features and the Delay-Doppler domain features alone for recognition.

Keywords: marine mammals; acoustic signals; feature fusion; target recognition; neural networks

Citation: Lü, Z.; Shi, Y.; Lü, L.; Han, D.; Wang, Z.; Yu, F. Dual-Feature Fusion Learning: An Acoustic Signal Recognition Method for Marine Mammals. *Remote Sens.* **2024**, *16*, 3823. https://doi.org/10.3390/rs16203823

Academic Editor: Marc Pinto

Received: 8 September 2024
Revised: 11 October 2024
Accepted: 11 October 2024
Published: 14 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The ability of marine mammals to use sound for communication is one of the key features of their adaptation to the underwater living environment, and their acoustic signals exhibit remarkable diversity. American scholar Lilly categorized these acoustic signals into three main types based on their functions and parameters, Click, Whistle, and Burst Pulse, with each sound type serving distinct purposes [1]. Studying the acoustic signals of marine mammals holds great significance for understanding their biological behavior, rational utilization of marine resources, and conservation efforts to protect these species. Passive Acoustic Monitoring (PAM) of marine mammals is a widely employed biomonitoring method. Since acoustic waves propagate with lower energy attenuation in the ocean than light waves, they can travel longer distances, making them particularly suitable for monitoring and identifying marine mammals [2]. In the pursuit of developing PAM systems for detecting and classifying marine mammals, numerous algorithms have been devised specifically to analyze the acoustic characteristics of these animals. Notably, most of these detection and classification algorithms rely on the distinct acoustic features exhibited by marine mammals [3].

Feature extraction is a crucial step in acoustic signal processing. By employing feature extraction techniques, one can obtain the acoustic features of the target, which subsequently serve as a reliable basis for target identification. Ibrahim et al. demonstrated the effectiveness of using Mel-Frequency Cepstral Coefficients (MFCC) and Discrete Wavelet

Transformation Coefficients in classifying marine mammal calls using Support Vector Machines (SVM) [4]. Brown et al. pioneered the use of the Gaussian Mixture Model combined with the Hidden Markov Model (GMM-HMM) for recognizing MFCC features [5]. In their study, they achieved a classification consistency of over 90% for a set of 75 killer whale calls. Dugan, on the other hand, extracted time-frequency features from cetacean calls and classified them using three different models, achieving the highest assignment rate of 86.45% [6]. Maheen et al. acoustically classified six marine mammal species by fusing one-dimensional localized binary patterns with MFCC features. Their approach yielded a training test accuracy of 90.4% [7]. Meanwhile, Zhong Mingtuo et al. fused MFCC, Linear Cepstral Coefficients, and time-domain features from 61 species of marine mammals as feature parameters [8]. By utilizing SVM for classification, they managed to improve the recognition rate by 5.5% compared to traditional MFCC-based methods. Li Songbin et al. extracted six features—MFCC, FBanks, PNCC, PSRCC, GFCC, and MSRCC—from three marine mammal species for comparative analysis [9]. They employed a Convolutional Neural Network-Gated Recurrent Unit (CNN-GRU) structure for recognition and achieved a classification accuracy of 74%. Some of these studies use one feature to identify marine mammal acoustic signals, and some use multiple, but none of the features they use can be separated from the time-frequency domain. Since climatic factors, such as rainfall and typhoons, have a great impact on the marine acoustic environment, the recognition of marine mammal acoustic signals only by time-domain features or frequency-domain features has certain limitations. Zhang Xuebo et al. coherently synthesized signals in the range-Doppler domain associated with each receiver after performing range cell migration correction (RCMC) for each receiver, and then corrected the azimuth offset [10,11]. To improve the simulation efficiency of multi-receiver synthetic aperture sonar (SAS) echo signal, Zhang Xuebo et al. multiplied the spectrum of the transmitted signal with the phase shift related to the delay so that the spectrum of the echo signal can be accurately obtained [12,13]. Compared with the traditional echo simulation algorithm, this method significantly improves the simulation efficiency of the echo signal without losing performance. Although the MFCC features have achieved good recognition results in the current field, the Delay-Doppler (DD) domain features can respond to the speed information of the target and are not easily affected by environmental factors such as climate, so combining them with the MFCC features can increase the reliability of the recognition, and the use of the dual-feature fusion learning method may be beneficial to improve the accuracy of the recognition.

Convolutional Neural Networks (CNNs) are widely utilized in the field of speech recognition. Traditionally, acoustic signal recognition involves time-frequency analysis of the acoustic signal to generate a spectrogram, which is then used to identify the acoustic signal based on its unique patterns. Detailed information on machine learning and deep learning-based methods and underwater sound sources, features, classifiers, datasets, related techniques, challenges, and future trends for marine ship sound classification and fish sound classification are discussed by Aslam et al. [14]. Bianco et al. presented the development of machine learning in four acoustic research areas: source localization in speech processing, source localization in marine acoustics, bioacoustics, and environmental sounds in everyday scenes [15]. In the process of feature extraction for speech signals, the original acoustic features are often replaced with acoustic feature images, and CNN-based image recognition techniques are employed to recognize the acoustic signals. This approach has proven to achieve accuracy rates that are difficult to match using traditional methods, especially when dealing with large-sample datasets. Zhang Xuebo et al. focus on the application of a nonlinear chirp scaling algorithm in SAS and validate the proposed method through simulation and real data. The processing results show that the imaging efficiency is greatly improved compared with the phase center approximation (PCA) method [16]. Wang et al. proposed a comprehensive underwater image enhancement framework, the metalantis framework, which enhances state-of-the-art physical models of underwater imaging by utilizing virtually generated data for reinforcement learning [17,18]. Meanwhile,

they gave two examples in [19,20]. Shiu et al. explored the use of deep CNNs and Recurrent Neural Networks (RNNs) with spectrograms to detect vocalizations of North Atlantic Right Whales [21]. Their findings indicated that deep learning architectures can produce false positive rates several orders of magnitude lower than other algorithms. Griffiths et al. proposed a multivariate clustering method to identify distinct Click vocal clusters of Dall's Porpoise in the U.S.A., and the validity of the three clusters was verified using the Random Forest method [22]. Cai et al. designed a multichannel-based classification model with a parallel structure, fusing predictions and introducing data enhancement techniques to further improve classification accuracy [23]. Duan Dexin et al. trained a random forest classifier using time-frequency graph features to detect and distinguish echolocation signals, achieving higher recall and accuracy under low Signal-to-Noise Ratio (SNR) conditions [24]. Cominelli et al. combined pre-trained acoustic classification models (VGGish, NOAA, and Google Humpback Whale Detector), dimensionality reduction (UMAP), and balanced random forest algorithms to demonstrate how machine-learned acoustic features can capture different aspects of the marine acoustic environment [25]. The current state of research suggests that applying deep learning to marine mammal acoustic recognition has become a trend, but most of the methods proposed so far are based on a feature-based recognition model. Although neural network models for marine mammal acoustic recognition offer higher accuracy and reduce time and labor costs, traditional neural network models often suffer from complexity, computational costs, and other problems.

To address the limitations of traditional recognition methods that rely solely on a single feature input, this paper proposes a dual-feature fusion learning target recognition model. To validate the effectiveness of the proposed model for marine mammal acoustic recognition, this study combines three common CNN recognition models with two signal features, respectively, for single-feature recognition, which is compared with two-feature recognition. The innovations of the dual-feature fusion learning method presented in this paper are as follows:

The marine mammal acoustic signal is preprocessed using adaptive filtering to enhance the SNR and mitigate the interference of environmental noise.

1. Delay-Doppler domain features are introduced into the acoustic feature recognition of marine mammals, effectively addressing the impact of seasonal changes in the marine environment on marine mammal acoustic signals.
2. A dual-feature fusion learning target recognition model is developed, capable of recognizing both MFCC features and Delay-Doppler domain features simultaneously. This model exhibits high recognition accuracy and strong generalization ability for mammal acoustic signal recognition in complex marine environments.

2. Theory

2.1. Least Mean Square Adaptive Filter

To mitigate the effects of environmental noise on marine mammal identification, this study employs the Least Mean Square (LMS) adaptive filter to process the acoustic signals of marine mammals, thereby enhancing the SNR [26]. LMS is an adaptive filter based on the minimum mean square error criterion. The core idea is to minimize the sum of squares of the errors between the output signal of the filter and the desired signal by continuously adjusting the coefficients of the filter. The LMS algorithm is based on the Wiener filtering theory, and adopts the algorithm of estimating the gradient vector by instantaneous values to update the adaptive filter weights coefficients by minimizing the energy of the error signal. The structure of the adaptive filter is shown in Figure 1, the input signal $x(n)$ is passed through a parameter-adjustable digital filter to produce an output signal $y(n)$, which is compared with the desired signal $d(n)$ to form an error signal $e(n)$, and the filter-parameters are adjusted by an adaptive algorithm to minimize the mean square value of $e(n)$ [27].

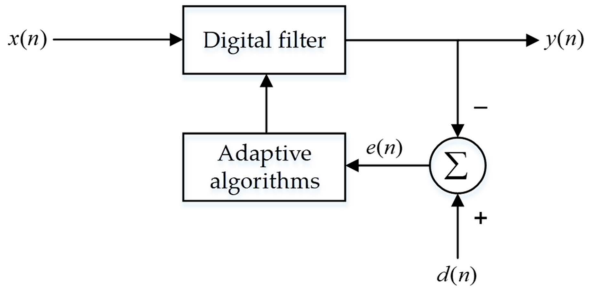


Figure 1. The structure of the LMS adaptive filter.

Let the input signal be a vector $x(n)$ and the output of the LMS adaptive filter be $y(n)$. Then we have [28]:

$$y(n) = \sum_{i=0}^{N} s_0 x(n-i) = S^T(n)X(n) \tag{1}$$

$$e(n) = d(n) - y(n) = d(n) - S^T(n)X(n) \tag{2}$$

where $S(n)$ is the weight coefficient of the filter and N is the order of the filter.

Adaptive filtering can utilize the results of the filter parameters that have been obtained at the previous moment to automatically adjust the filter parameters at the current moment to adapt to the unknown or time-varying statistical characteristics of the signal and noise, thus achieving optimal filtering to improve the SNR.

2.2. Mel-Frequency Cepstral Analysis

The Mel-Frequency Cepstral Coefficient is a cepstral parameter derived from the Mel scale frequency domain. By employing equally spaced band divisions on the Mel scale, MFCC more closely mimics the human auditory system compared to linearly spaced bands used in traditional logarithmic cepstral analysis. This results in superior recognition performance, particularly when the SNR is low [29]. Figure 2 illustrates the steps involved in extracting MFCC features from a signal.

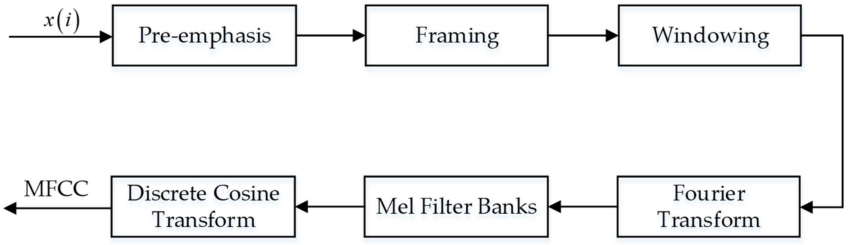


Figure 2. The procedure for extracting MFCC features.

(1) Preprocessing: Pre-emphasis, framing, windowing. The purpose of pre-emphasis is to boost the high-frequency part so that the gap between the peaks of the spectrum of the whole signal is reduced by passing the audio signal through a high-pass filter. Due to the non-smooth and short-time smooth characteristics of the speech signal, the speech signal is divided into frames, N samples are gathered into one frame, and there should be a section of overlapping area between two neighboring frames to avoid too large variations between two neighboring frames. To increase the continuity of the left and right ends of the frame, it is necessary to multiply the audio signal of each frame by a window function, and the Hamming window matrix C is multiplied by the post-split-frame matrix S to get the post-windowed matrix SC.

(2) Fast Fourier Transform (FFT) and calculation of energy spectrum. The spectrum of each frame is obtained by performing FFT on each frame of the signal after framing and windowing [30]:

$$X(i) = \sum_{n=o}^{N-1} x(n)e^{-j\frac{2\pi i}{n}}, 0 \le i \le N \tag{3}$$

where $x(n)$ is the input signal and N represents the number of points of the Fourier transform.

(3) Mel filter. The filter bank consists of 32 triangular filters, with dense filters and high thresholds at low frequencies and sparse filters and low thresholds at high frequencies. The conversion relationship between frequency and the Mel scale is [31]:

$$Mel(f) = 2595lg\left(1 + \frac{f}{700}\right) \tag{4}$$

where f is the frequency in Hz.

(4) Discrete Cosine Transform (DCT). The filter bank coefficients are highly correlated, which may be problematic in machine learning algorithms; as such, the DCT is applied to de-correlate the filter bank coefficients to finally obtain the MFCC features [31]:

$$C(n) = \sum_{k=1}^{p} logX(k)cos\left(\frac{\pi(k-0.5)n}{p}\right), n = 1, 2, \ldots, L \tag{5}$$

where L is the MFCC coefficient order and p is the number of triangular filters.

2.3. Delay-Doppler Domain Analysis

The Delay-Doppler domain combines the properties of a signal in the time and frequency domains, mapping the signal into a two-dimensional time-frequency lattice by placing lattice points in both the time-delay dimension and the Doppler dimension. This representation allows the signal to maintain relatively stable transmission characteristics in a complex Doppler frequency shift environment. The relative motion between the target and the signal-receiving equipment induces a Doppler effect on the signal. The Delay-Doppler domain analysis of the signal can reveal the frequency shift caused by this effect. Since different marine mammals exhibit distinct average speeds, they generate unique Doppler shifts, allowing the Delay-Doppler domain features of the signal to reflect the movement characteristics of these animals.

There exists a defined relationship between the time-frequency domain and the Delay-Doppler domain of the signal. Specifically, the time-domain signal undergoes FFT to convert it into the frequency domain. Subsequently, two-dimensional sampling with specific periods and frequency intervals is performed to obtain the time-frequency domain features. Finally, the time-frequency domain signal is subjected to the Symplectic Finite Fourier Transform (SFFT) to derive Delay-Doppler domain features. The expression for SFFT is provided in [32]:

$$Y[p, q] = \frac{1}{\sqrt{PQ}} \sum_{n=0}^{Q-1} \sum_{m}^{P-1} X[l, k]e^{-j2\pi(\frac{qk}{Q} - \frac{pl}{P})} \tag{6}$$

where X [l, k] is the signal after Fourier transform and sampling of the time domain signal, M and N denote the number of time delay dimensions and Doppler dimensions, respectively, and $0 \le p \le P$, $0 \le q \le Q$.

Inverse Symplectic Finite Fourier Transform (ISFFT) is the inverse transformation of SFFT, which is able to simultaneously transform the time-delay domain to the frequency domain and the Doppler domain to the time domain, so as to complete the transformation of the received signal from the Delay-Doppler domain to the time-frequency domain. The

transform relationships in the time-frequency domain and Delay-Doppler domain are shown in Figure 3.

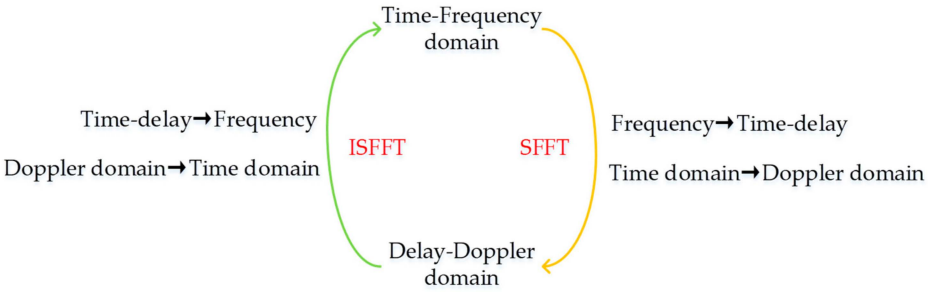


Figure 3. Transformation relations in the time-frequency and Delay-Doppler domains. The green arrow indicates Inverse Symplectic Finite Fourier Transform (ISFFT), the yellow arrow indicates Symplectic Finite Fourier Transform (SFFT). The red font indicates the two transformations.

2.4. Convolutional Neural Network Model

CNN is a deep learning model commonly used in fields such as image and audio recognition The core idea is to extract the features of the data such as images through convolutional operations to achieve the purpose of classification and recognition of the data. Figure 4 illustrates the typical architecture of a CNN, which typically encompasses convolutional layers, pooling layers, and fully connected layers.

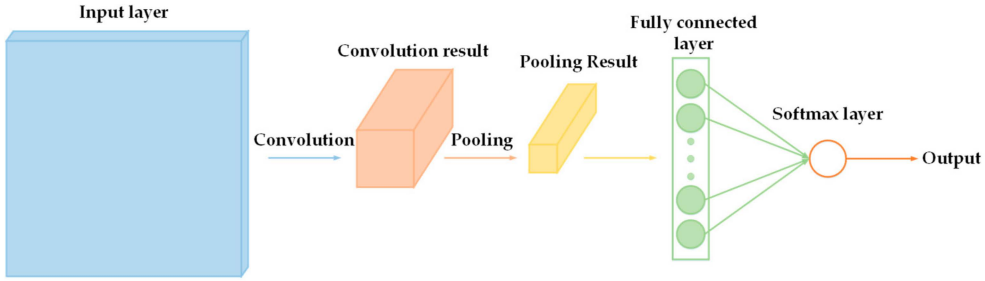


Figure 4. The architecture of CNN.

The convolutional layer consists of multiple convolutional kernels and the complete feature map is obtained by using different convolutional kernels. The eigenvalue of position (i, j) in the kth feature map of the lth layer is [33]:

$$v_{p,q,k}^{l} = W_{k}^{l^T} i_{p,q}^{l} + B_{k}^{l} \tag{7}$$

where W_k^{lT} is the weight vector of the kth filter of the lth layer, B_k^l is the bias term of the kth filter of the lth layer, and $i_{p,q}^l$ is the input segment centered at position (i, j).

As hardware technology and computational capabilities continue to advance at a rapid pace, so too do the design and application of convolutional neural network models. These models are tailored to specific tasks and application scenarios to optimize performance and efficiency, with notable examples including VGG16, GoogleNet, ResNet, and others.

The VGG16 model consists of 13 convolutional layers and 3 fully connected layers; the convolutional part uses a smaller 3×3 convolutional kernel and a convolutional operation with a step size of 1 [34]. This design approach allows the network to be deeper. Between every two convolutional layers, VGG16 also uses a 2×2 maximum pooling layer to reduce the size of the feature map and retain the most salient features. After the final convolutional layer, VGG16 uses 3 fully connected layers, each with 4096 hidden units, and the last fully connected layer outputs the predictions of the model.

The GoogleNet model has about 22 layers, including convolutional layers, pooling layers, fully connected layers, and the Inception module [35]. Its most important feature

is the use of the Inception module, which allows the simultaneous use of multiple convolutional kernels of different sizes and pooling layers to extract features, thus increasing the expressive power and accuracy of the network. GoogleNet also uses global average pooling instead of maximal pooling, which averages the entire feature map in an average operation to obtain a feature vector as the final output, reducing the model's number of parameters and preventing overfitting.

The ResNet model is proposed to solve the network degradation problem when there are too many hidden layers in a deep neural network [36]. While traditional neural networks try to learn a function of the input and target mapping, ResNet learns the residuals between the input and target. The residuals are obtained by comparing the input signal with the desired output signal and then learning the residuals; this approach helps to solve the problems such as gradient vanishing. The ResNet model can learn a deep network with 152 layers and can obtain higher accuracy than the VGG model and GoogleNet model.

3. Method

First, this research performs adaptive filtering on acoustic signals of marine mammals to enhance their SNR. Subsequently, it extracts the MFCC and DD domain features of marine mammals as the two input features of this recognition method. This study constructs a dual-feature fusion learning target recognition model, which can be trained by inputting two marine mammal acoustic signal features at the same time and can improve the target recognition accuracy. The overall idea of the paper is shown in Figure 5.

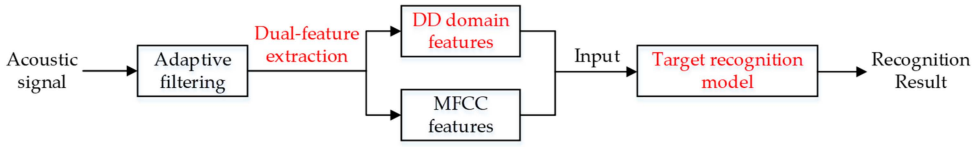


Figure 5. Overview of the experimental procedure. The red font indicates the innovations of this paper: dual-feature extraction, DD domain features and target recognition model.

3.1. Framing and Normalization

Framing: The purpose of framing is to extract a series of shorter, discrete time segments (frames) from a continuous signal so that each frame can be further analyzed and processed. A frame is a fixed-length sequence of samples extracted from an audio signal. The length of a frame is usually a power of 2, such as 256, 512, or 1024, because such a length makes subsequent Fast Fourier Transform (FFT) calculations more efficient. The frameshift is the number of samples between two consecutive frames. The frameshift determines the degree of overlap between frames. Smaller frameshifts provide higher temporal resolution, but increase computational effort; larger frameshifts reduce computational effort, but decrease temporal resolution.

For example, after inputting a segment of the signal, set the number of samples per frame to 1024 and the frameshift to 512, i.e., move 512 samples at a time to get a new frame, which ensures 50% overlap, and draw the data of the first three frames as shown in Figure 6.

Normalization: Min-Max Normalization is a method of scaling data features to a specific range (usually between 0 and 1). This method is implemented through the maximum and minimum values of each feature, using a linear transformation to map the data to the new range. The formula for this is [37]:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{8}$$

where x is the original data point, $min(x)$ is the minimum value in the data set, $max(x)$ is the maximum value in the data set, and x' is the normalized data point.

The advantage of Min-Max Normalization is that it is sensitive to outliers, since changes in the maximum and minimum values directly affect the normalized result, and the data can be easily scaled to any specified range.

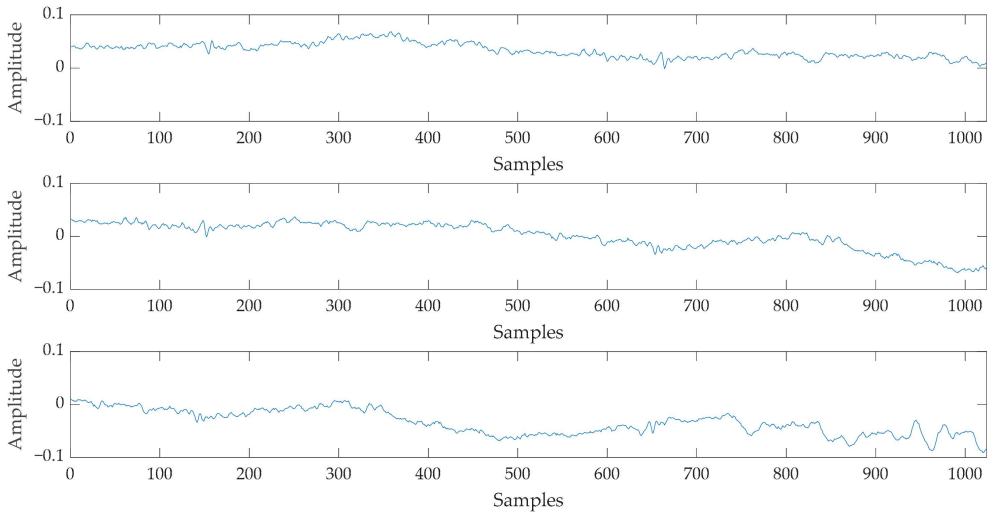


Figure 6. Time-domain plot of three frames.

3.2. Dual-Feature Extraction Analysis

Due to its excellent nonlinear perception ability, MFCC has been widely applied in acoustic signal recognition research. However, this feature remains a traditional time-frequency domain feature, susceptible to environmental changes. In this study, we extract the DD domain feature of marine mammals, which complements the MFCC feature and can reflect the motion characteristics of marine mammals. This approach addresses the limitations of recognizing marine mammals using a single feature.

Figure 7 depicts the MFCC features of the Fraser's Dolphin acoustic signal. The horizontal axis represents the frame rate, where each frame encompasses a specific number of samples with a certain overlap between neighboring frames. The more frames there are, the higher the temporal resolution, which can more accurately capture the dynamic properties of the signal and changes in short-term characteristics. This is because more frames mean that the signal is more finely segmented, which better reflects changes in the signal over time. The vertical axis indicates the MFCC parameter dimension, signifying the number of MFCCs extracted per frame. This number determines the dimensionality of the feature vector for each frame. For example, if the order of the DCT is 30, then 30 MFCC coefficients will be generated for each frame, but usually only the first 13 coefficients are retained because these lower-order coefficients contain the main spectral information, while the higher-order coefficients tend to be associated with noise.

Figure 8 illustrates the Delay-Doppler domain features of the Fraser's Dolphin acoustic signal. The horizontal axis represents the Doppler frequency shift, indicating the change in signal frequency resulting from the relative motion between the target and the receiver. The vertical axis depicts the time delay, which is the duration of signal propagation. The Delay-Doppler domain features effectively reflect the speed characteristics of organisms, as varying speeds among different organisms inherently leads to distinct Delay-Doppler domain features. As evident from the figure, the signal exhibits a pronounced frequency shift at approximately 8 Hz.

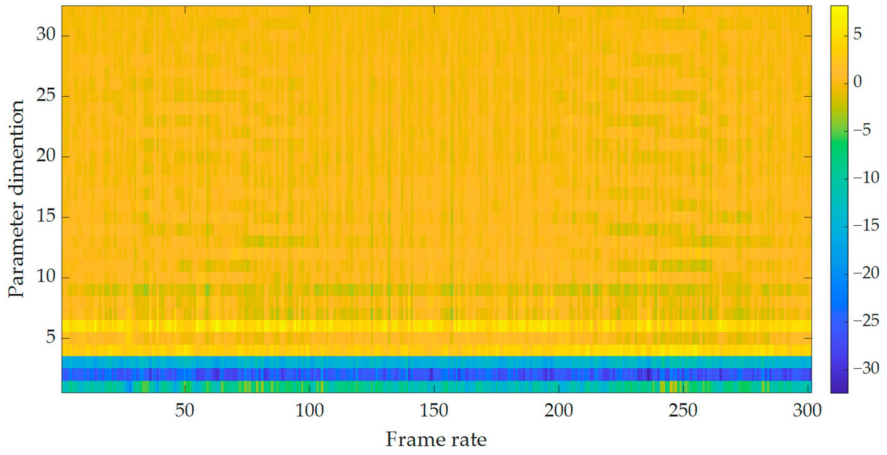


Figure 7. MFCC features.

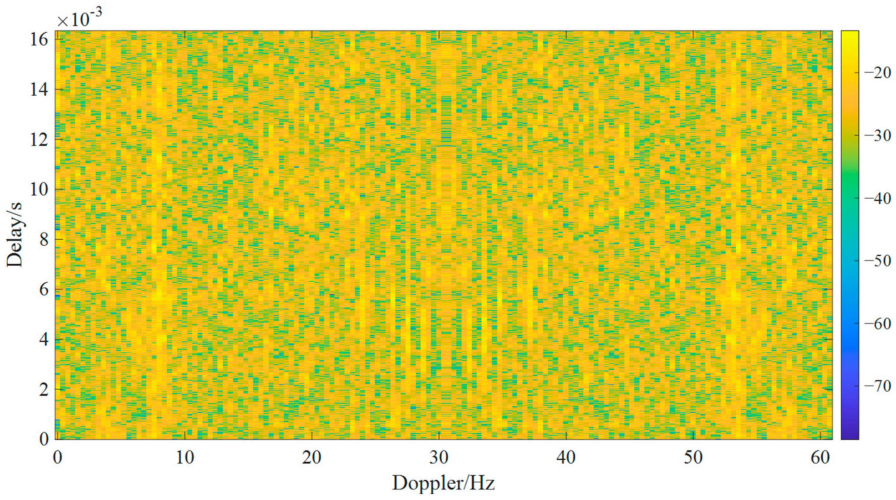


Figure 8. Delay-Doppler features.

3.3. Dual-Feature Fusion Learning Target Recognition Model

Traditional CNN models are typically based on a single feature and can only be trained for that specific feature, which often limits their generalization ability and robustness. Consequently, in this paper, we introduce a dual-feature fusion learning target recognition model that is capable of simultaneously inputting features from both the time-frequency and Delay-Doppler domains. The acoustic signals generated by different targets may have frequency overlap, so it is challenging to rely only on the time-frequency domain features of the signals to identify the targets. Since different targets move at different speeds, they are characterized differently in the DD domain. The target attributes in the other feature domain of the signal can be described by DD domain features, and the classification approach is equivalent to combining the MFCC features and the DD domain features as a joint feature and classifying the different marine mammals by using this joint feature as an axis, which corresponds to different points on this axis. Due to the increased dimensionality of the described signal, the target information can be reflected more comprehensively. We compare the recognition performance of this model with three widely used single-feature models: VGG16, GoogleNet, and ResNet.

As depicted in Figure 9, our dual-feature fusion learning target recognition model comprises nine convolutional layers and two fully connected layers. Notably, a max pooling layer is inserted between every two convolutional layers, and the recognition task is ultimately carried out by a SoftMax layer. The maximum pooling layer reduces the size of the feature map by selecting the maximum value of each region, which reduces the amount of computation and the number of parameters in the subsequent layers, helping to improve

the computational efficiency of the model and reduce the risk of overfitting. In addition to this, the maximum pooling layer provides a degree of translation invariance, so that even if there is a small translation of the image, the maximum pooling layer can extract the same features. With the pooling operation, the model is able to retain the most important features and ignore unimportant details, thus improving the robustness of the features. The SoftMax layer transforms the output of the neural network into a probability distribution, such that the output value of each category is between 0 and 1, and the sum of the output values of all the categories is 1. This allows the model to perform better on multi-classification tasks and improves the accuracy of the classification. The fully connected layer is able to integrate the features extracted from the convolutional and pooling layers and map these features to the sample labeling space, enhancing the feature integration capability of the model.

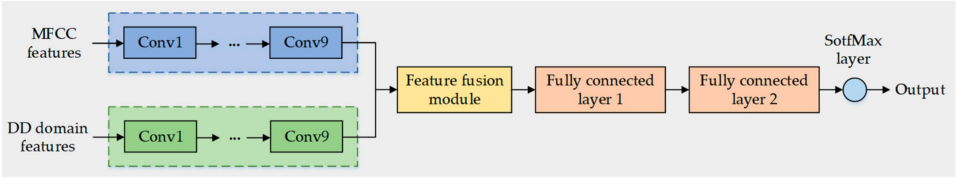


Figure 9. The architecture of the dual-feature fusion learning target recognition model.

Figure 10 presents the specific parameters of the nine convolutional layers. For instance, the parameter "1 × 1 × 32" signifies the following: the convolutional kernel size is 1 × 1, allowing for cross-channel information integration without altering the spatial dimensions; furthermore, the number of convolutions is 32, indicating that the convolutional operation yields 32 feature maps that reflect the outcomes of convolving the input data with this kernel. Additionally, the activation function employed is ReLU (Rectified Linear Unit), renowned for its simplicity in computation, rapid convergence, and effectiveness in mitigating the gradient vanishing problem. The normalization method used is Min-Max Normalization, which is implemented by the maximum and minimum values of each feature, using linear transformations to map the data to new ranges.

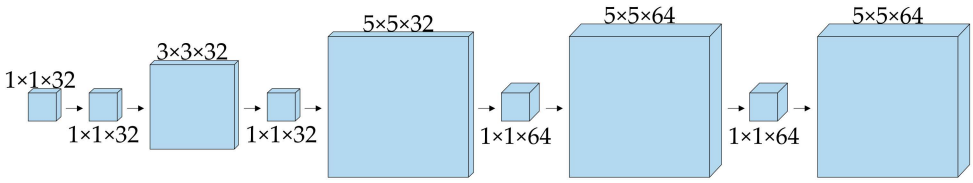


Figure 10. Parameters of the convolutional layers in the model.

4. Experiment and Analysis

In this section, the effectiveness of the aforementioned dual-feature fusion learning target recognition model is verified. First, the pre-processed signal undergoes MFCC and DD domain feature extraction. Subsequently, these two types of features are input into VGG16, GoogleNet, ResNet, and the dual-feature fusion learning target recognition model for training. The recognition performances of these different models are then analyzed and compared. The experimental environment configuration for this study is as follows: the GPU is NVIDIA GeForce RTX 4080 (NVIDIA Corporation, Santa Clara, CA, USA), the CPU is Intel i9-13900K (Intel Corporation, Santa Clara, CA, USA), and the neural network is trained to utilize the GPU. The operating system is Windows 10, with 128 GB of RAM (Samsung, Seoul, Republic of Korea). The Python version used is 3.9, and the neural network is constructed using the PyTorch framework. The development environment is PyCharm 2018.

4.1. Experimental Data and Evaluation Metrics

The marine mammal acoustic data used in this study were obtained from the Watkins Marine Mammal Sound Database open-source database, which was collected by William

Watkins, one of the founding fathers of marine mammal bioacoustics, and contains recordings from seven decades, from the 1940s to the 2000s. The database contains approximately 2000 unique recordings of more than 60 species of marine mammals, ranging in length from one second to several minutes, all in .wav format. Figure 11 shows three marine mammals used in the experiment [38].

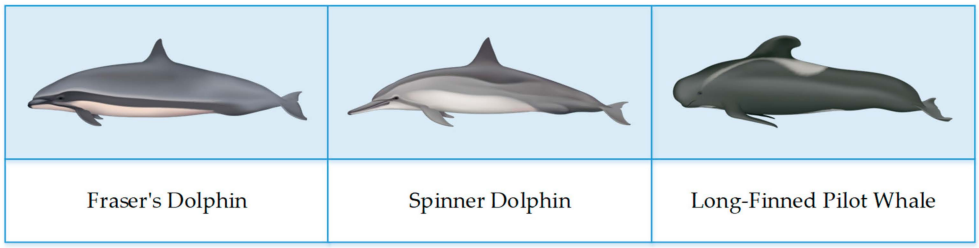


Figure 11. Types of marine mammals used in this study.

Due to the varying lengths of the audio data in the dataset and the short duration of marine mammal vocalizations, in this paper, the longer audio data are clipped into multiple short segments, which increases the amount of input data and improves the accuracy of the training. To ensure that each short segment after clipping contains at least one complete animal vocalization, this paper utilizes Adobe Audition 2024 software to analyze the time-frequency diagrams of the audio data before clipping. The dataset is divided into training sets and test sets in the ratio of 4:1, and the following experimental results are analyzed from these two aspects.

In this study, four metrics were selected to evaluate the experimental results: Accuracy, Precision, Recall, and F1 Score. These metrics are derived from the confusion matrix, which evaluates the model's accuracy by comparing the predicted category labels with the actual category labels. The structure of the confusion matrix is presented in Table 1 [39].

Table 1. Structure of the confusion matrix.

	Actual Category: Positive	**Actual Category: Negative**
Predicted category: Positive	*TP*	*FP*
Predicted category: Negative	*FN*	*TN*

True Positives (*TP*): the number of samples that the model correctly predicts as positive.
False Positives (*FP*): the number of samples that the model incorrectly predicts as positive.
True Negatives (*TN*): the number of samples that the model correctly predicts as negative.
False Negatives (*FN*): the number of samples that the model incorrectly predicts as negative.

Accuracy is the proportion of correctly predicted samples to the total number of samples. Precision is the proportion of samples that are positive classes among all samples that are predicted to be positive classes. Recall is the proportion of samples that are correctly predicted to be positive classes among all samples that are positive classes. The F1-score is the harmonic mean of precision and recall, which is used to measure the balanced performance of the model. They are calculated as follows [40]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

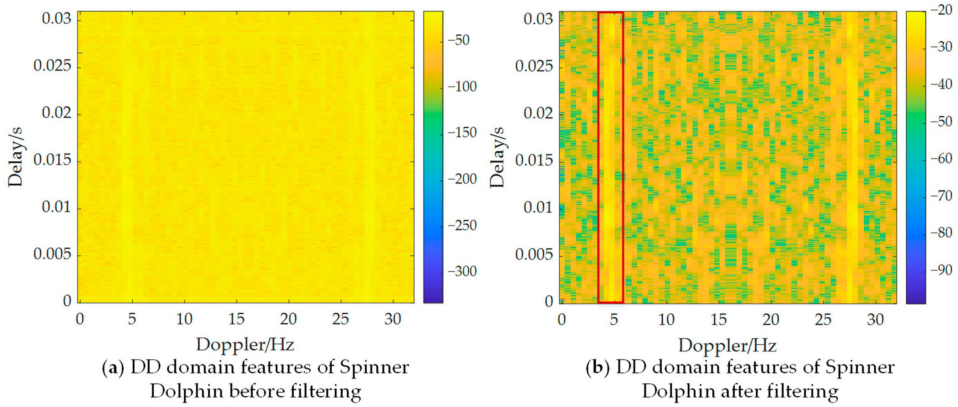
$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

4.2. Experimental Validation

In this paper, Section 2 mentions that a set of triangular filters is used for extracting MFCC features, which serve to improve the SNR of the original signal. The DD features are extracted without filtering the signal, whereas the signal is analyzed by LMS adaptive filtering before DD feature extraction. The analysis results are presented in Figure 12, which shows the DD domain feature extraction of the Spinner Dolphin before and after filtering, respectively. Notably, the results are significantly different, with the DD domain feature extraction after filtering making the visualization features more obvious and enhancing image contrast.



(**a**) DD domain features of Spinner Dolphin before filtering

(**b**) DD domain features of Spinner Dolphin after filtering

Figure 12. Comparison of DD domain features before and after filtering.

In this study, all marine mammal acoustic signals were minimum mean square filtered, and then the DD-domain features of the original signals and the DD-domain features of the filtered signals were input into the three single feature models of VGG16, GoogleNet, and ResNet as the target recognition features, respectively. The accuracy, precision, recall, and F1 scores of the two groups were obtained for comparison, and the results are shown in Tables 2 and 3:

Table 2. Recognition results of DD domain features before filtering.

Model-Feature	Acc (%)	Pre (%)	Re (%)	F1 (%)
VGG16-DD	62.47/64.71	62.76/75.87	62.47/64.71	62.31/64.42
GoogleNet-DD	72.60/54.90	73.13/55.90	72.59/54.90	72.55/54.01
ResNet-DD	74.82/57.84	75.15/57.84	74.82/57.84	74.69/57.78

Table 3. Recognition results for DD domain features after filtering.

Model-Feature	Acc (%)	Pre (%)	Re (%)	F1 (%)
VGG16-DD	75.56/59.80	75.71/67.08	75.56/59.80	75.55/59.22
GoogleNet-DD	79.26/72.55	79.29/73.23	79.26/72.55	79.24/72.74
ResNet-DD	77.28/59.80	77.44/60.48	77.28/59.80	77.33/59.82

The table contains the results of the training set and the test set; the two kinds of data are separated by "/", the data in front of "/" is the result of the training set, the data on the right side of "/" is the result of the test set. By comparing the recognition results of DD domain features from single-feature models before and after filtering, it becomes evident that the recognition accuracy of each model improves by 3% to 13% after filtering. Consequently, applying LMS adaptive filtering before feature extraction and recognition can effectively enhance recognition accuracy.

MFCC and DD domain features are extracted from preprocessed data and individually input into VGG16, GoogleNet, and ResNet models for single-feature training. Subsequently, these two types of features are simultaneously input into the dual-feature fusion learning

target recognition model for training. Table 4 presents the recognition results for the two features using the four models.

Table 4. Recognition results (Fraser's Dolphin, Spinner Dolphin, Long-Finned Pilot Whale).

Model-Feature	Acc (%)	Pre (%)	Re (%)	F1 (%)
VGG16-MFCC	94.07/95.10	94.09/95.10	94.07/95.10	94.08/95.13
VGG16-DD	75.56/59.80	75.71/67.08	75.56/59.80	75.55/59.22
GoogleNet-MFCC	93.09/95.10	93.22/95.18	93.09/95.10	93.11/95.10
GoogleNet-DD	79.26/72.55	79.29/73.23	79.26/72.55	79.24/72.74
ResNet-MFCC	96.05/97.06	96.09/97.11	96.05/97.06	96.06/97.07
ResNet-DD	77.28/59.80	77.44/60.48	77.28/59.80	77.33/59.82
Dual-feature	99.26/98.04	99.28/98.04	99.26/98.04	99.26/98.04

In the final line, "Dual-feature" signifies that the dual-feature fusion learning target recognition model is employed to recognize both MFCC features and DD domain features. Analyzing the training results reveals that the accuracy rate stands at 59% to 80% for DD domain features alone, 93% to 98% for MFCC features alone, and a remarkable 98% to 99% when both MFCC features and DD domain features are utilized through the dual-feature fusion learning target recognition model. Therefore, the dual-feature fusion learning target recognition model is better than the other models in recognizing the acoustic signals of three marine mammals, namely, the Fraser's Dolphin, the Spinner Dolphin, and the Long-Finned Pilot Whale. Comparison with other models alone does not adequately demonstrate the superiority of the model proposed in this paper. As such, this study conducts generalizability experiments and ablation experiments to validate the model's ability to generalize and the reasonableness of the methodology.

4.3. Generalization Ability Analysis

To verify the generalizability of this dual-feature fusion learning target recognition model, two distinct marine mammal acoustic signals—the Ross seal and the Bearded seal—were chosen for training in this study. The training results are presented in Table 5.

Table 5. Recognition results (Ross seal, Bearded seal).

Model-Feature	Acc (%)	Pre (%)	Re (%)	F1 (%)
VGG16-MFCC	97.73/73.91	97.84/73.91	97.73/73.91	97.73/73.91
VGG16-DD	88.64/69.57	88.74/77.43	88.64/69.57	88.64/70.15
GoogleNet-MFCC	92.05/73.91	92.09/74.47	92.05/73.91	92.06/73.61
GoogleNet-DD	90.91/86.96	91.03/89.97	90.91/86.96	90.92/86.96
ResNet-MFCC	97.73/86.96	97.73/87.39	97.73/86.96	97.73/87.01
ResNet-DD	77.29/59.80	77.44/60.48	77.28/59.80	77.33/59.82
Dual-feature	98.86/91.30	98.89/92.55	98.86/91.30	98.87/91.20

When DD domain features are recognized in isolation, the accuracy ranges from 59% to 90%. When MFCC features are recognized alone, the accuracy lies between 73% and 97%. However, when the dual-feature fusion learning target recognition model recognizes both MFCC and DD domain features, the accuracy soars to 91% to 98%. The superior recognition accuracy achieved by the dual-feature fusion learning target recognition model, as compared to other models, underscores its excellent generalization capabilities.

4.4. Ablation Experiment

Ablation Study is an experimental design method commonly used in scientific research, especially in the fields of machine learning and deep learning. The core idea is to gain a deeper understanding of how the model works and how the components interact with each other by systematically removing or modifying certain parts of the model

(e.g., layers, nodes, features, parameters, etc.) and observing how such changes affect the model's performance.

An ablation experiment was conducted in this study to validate the efficacy of MFCC features, DD domain features, and LMS adaptive filtering within the proposed method for marine mammal acoustic signal recognition. By systematically removing each component from the model, we aimed to assess their contributions to the overall performance.

Ablation Experiment 1: MFCC features are removed and DD domain features are trained using CNN.

Ablation Experiment 2: Remove DD domain features and train on MFCC features using CNN.

Ablation Experiment 3: Remove LMS adaptive filtering and directly perform feature extraction on the original signal.

The training results of the three ablation experiments and the complete target recognition model are compared in Table 6, which shows that the removal of MFCC features decreases the target recognition accuracy by about 43%; the removal of DD domain features decreases the target recognition accuracy by 0–3%; and the removal of LMS adaptive filtering decreases the target recognition accuracy by 1–3%.

Table 6. Results of ablation experiments.

Ablation Experiment	Acc (%)	Pre (%)	Re (%)	F1 (%)
1	56.30/54.90	56.38/57.62	56.30/54.90	56.15/54.35
2	96.54/98.04	96.54/98.10	96.54/98.04	96.53/98.03
3	97.29/95.10	97.40/95.34	97.28/95.10	97.30/95.10
Dual-feature	99.26/98.04	99.28/98.04	99.26/98.04	99.260/98.04

These findings demonstrate that the inclusion of MFCC features, DD domain features, and LMS adaptive filtering is crucial for achieving optimal performance in the target recognition model. Removing any of these components leads to a decrease in model accuracy, highlighting their contributions to the overall performance.

4.5. Qualitative Validation

The loss function is a non-negative real-valued function used to quantify the difference between model predictions and true labels [41]. By calculating the value of the loss function, the accuracy of the model predictions can be quantified, and thus the performance of the model can be evaluated. By minimizing the value of the loss function during training, the parameters of the model can be optimized so that the model's predictions are closer to the true labels. To provide a more intuitive analysis of the processes and effects of each model and feature recognition, we have plotted the loss function curves during the training process and the recognition accuracies of each model in Figures 13 and 14.

An explanation of the legend follows:

VGG-MFCC. Recognition of MFCC domain features using the VGG16 model.

VGG-DD. Recognition of DD domain features using the VGG16 model.

GoogleNet-MFCC. Recognition of MFCC features using the GoogleNet model.

GoogleNet-DD. Recognition of DD features using the GoogleNet model.

ResNet-MFCC. Recognition of MFCC features using the ResNet model.

ResNet-DD. Recognition of DD domain features using the ResNet model.

MFCC-DD. Recognition of MFCC and DD domain features using the dual-feature fusion learning target recognition model.

Figure 13 illustrates the variation of the loss function during the training of the three single-feature models and the dual-feature fusion learning target recognition model. Figure 14, on the other hand, shows the accuracies of different models for two types of feature recognition, including both the training and test sets, enabling a more direct comparison of the recognition performance of each model.

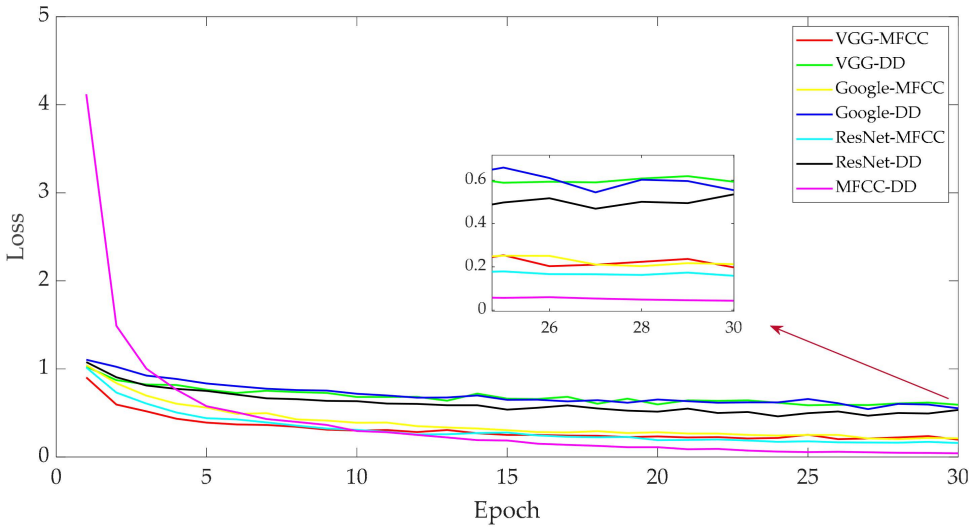


Figure 13. Loss function of each model during the training process.

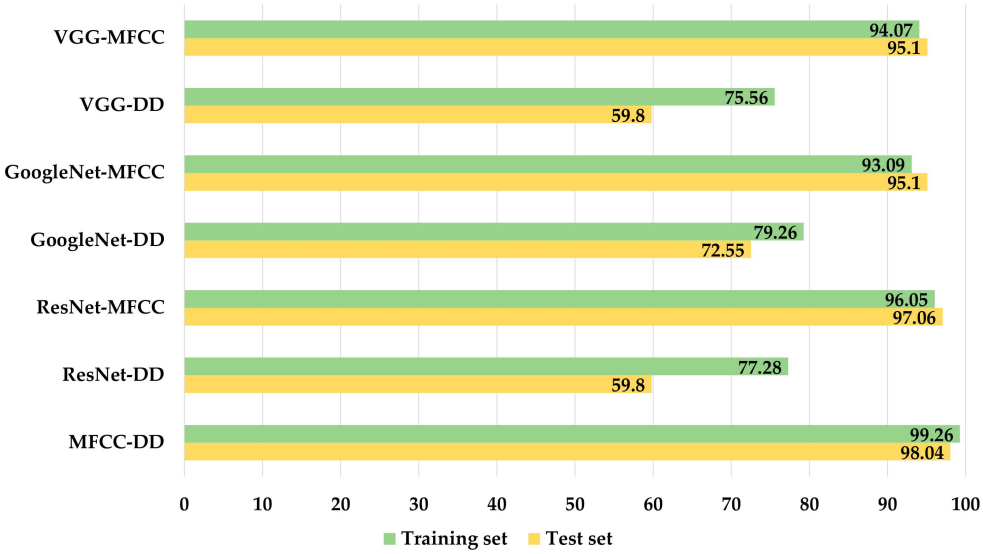


Figure 14. Recognition accuracy of each model.

(1) When recognizing DD domain features alone, the loss function decreases slowly and exhibits significant fluctuations. In contrast, recognizing MFCC features alone results in a relatively faster decrease in the loss function with fewer fluctuations. Notably, the dual-feature fusion learning target recognition model, which incorporates both MFCC and DD domain features, achieves an even faster decrease in the loss function with a smoother curve.

(2) Compared to the model using MFCC features alone, the dual-feature fusion learning target recognition model improves the accuracy of the training set by 3% to 6% and the accuracy of the test set by 1% to 3%. When compared to the model using DD domain features alone, the improvement in accuracy is even more pronounced, with an increase of 20% to 23% for the training set and 25% to 38% for the test set. Additionally, compared to models utilizing single features such as VGG16, GoogleNet, and ResNet, the structure of the dual-feature fusion learning target recognition model is simpler, contributing to an overall enhancement in model recognition efficiency.

5. Conclusions

In this paper, we propose a dual-feature fusion learning method that mainly consists of two parts: feature extraction and target recognition.

(1) Feature Extraction: The MFCC and DD domain features of marine mammals are extracted as input features. The MFCC, being closer to the human auditory system than other spectral features, captures the auditory characteristics of the marine mammals' vocalizations. Meanwhile, the DD domain features reflect the motion characteristics of these animals, providing complementary information. By combining these two features, the model ensures robust recognition performance even under low SNR conditions.

(2) Dual-Feature Fusion Learning Target Recognition Model: We introduce a novel dual-feature fusion learning target recognition model that can simultaneously input both features into a convolutional neural network for target recognition. In addition, generalizability experiments and ablation experiments are carried out in this study, which prove that the model has good generalization ability.

Compared with the traditional single-feature recognition model, the method proposed in this paper simplifies the model structure, improves the recognition accuracy and training efficiency, and has good generalization ability, which can provide some references for research in marine mammal acoustic recognition and other related fields. The method is based on the recognition of marine mammals by passive sonar, so it is suitable for the recognition of marine mammals that can actively emit sound, but it is less effective for the recognition of some fish. The acoustic signals emitted by underwater targets are transmitted through acoustic channels in the ocean, and their signal-to-noise ratio is bound to be greatly reduced, so noise in the ocean is an important issue affecting target recognition. With the increasing changes in the marine environment, including the impacts of climate change and human activities, the habitats and behavioral patterns of marine mammals are also changing [42]. Effective acoustic signal recognition techniques can help monitor these changes and study and protect animals and their habitats non-invasively and at ecologically relevant temporal and spatial scales. The future of this research can be applied to (1) the combination of active and passive sonar for the identification of marine organisms; (2) the combination of marine environmental noise filtering technology for detection; and (3) the upgrading of hardware and software technology, which will be mounted on a variety of marine observation platforms to observe organisms in the ocean in real-time, allowing us to achieve the goal of monitoring and protecting the marine biological environment.

In this paper, three models, VGG16, GoogleNet, and ResNet, are used as references in the comparison experiments, and future research will consider using more up-to-date algorithms for comparisons to improve the recognition efficiency of other models and to find more appropriate models. This study mainly relies on marine mammal acoustic signal data from the Watkins Marine Mammal Sound Database open-source database, and future studies will consider using more diverse datasets, including real-world data, to explore the performance of the target recognition model with dual-feature fusion learning in different environments.

Author Contributions: Conceptualization, Y.S. and Z.L.; methodology, Y.S.; validation, Y.S., Z.L. and Z.W.; formal analysis, Y.S.; investigation, Y.S. and Z.W.; writing-original draft preparation, Y.S.; writing-review and editing, Y.S., Z.L. and D.H.; supervision, Z.L., L.L., Z.W., F.Y. and D.H.; and project administration, Y.S. and Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (2022YFC2808003) and the Natural Science Foundation of Qingdao Municipality (23-2-1-100-zyyd-jch).

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to laboratory confidentiality regulations.

Acknowledgments: The authors would like to express their gratitude for the data used in this study: the Watkins Marine Mammal Sound Database. This valuable database was instrumental in conducting the research and validating the proposed model.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Lilly, J.C. Sonic-ultrasonic emissions of the bottlenose dolphin. *Whales Dolphins Porpoises* **1966**, *165*, 503–507.
2. Fleishman, E.; Cholewiak, D.; Gillespie, D.; Helble, T.; Klinck, H.; Nosal, E.; Roch, M.A. Ecological inferences about marine mammals from passive acoustic data. *Biol. Rev.* **2023**, *98*, 1633–1647. [CrossRef] [PubMed]
3. Cauchy, P.; Heywood, K.J.; Merchant, N.D.; Risch, D.; Queste, B.Y.; Testor, P. Gliders for passive acoustic monitoring of the oceanic environment. *Front. Remote Sens.* **2023**, *4*, 1106533. [CrossRef]
4. Ibrahim, A.K.; Zhuang, H.; Erdol, N.; Ali, A.M. A new approach for North Atlantic right whale upcall detection. In Proceedings of the 2016 International Symposium on Computer, Consumer, and Control (IS3C), Xi'an, China, 4–6 July 2016; IEEE: New York, NY, USA, 2016; pp. 260–263.
5. Brown, J.C.; Smaragdis, P. Hidden Markov and Gaussian mixture models for automatic call classification. *J. Acoust. Soc. Am.* **2009**, *125*, EL221–EL224. [CrossRef] [PubMed]
6. Dugan, P.J.; Rice, A.N.; Urazghildiiev, I.R.; Clark, C.W. North Atlantic right whale acoustic signal processing: Part I. Comparison of machine learning recognition algorithms. In Proceedings of the 2010 IEEE Long Island Systems, Applications, and Technology Conference, Farmingdale, NY, USA, 7 May 2010; IEEE: New York, NY, USA, 2010; pp. 1–6.
7. Nadir, M.; Adnan, S.M.; Aziz, S.; Khan, M.U. Marine mammals classification using acoustic binary patterns. *Arch. Acoust.* **2020**, *45*, 721–731.
8. Zhong, M.; Cai, W. Marine Mammal Sound Recognition Based on Feature Fusion. *Electron. Sci. Tech.* **2019**, *32*, 32–37.
9. Li, S.; Liu, P.; Yan, J.; Wang, K.; Gan, W.; Wang, J. A marine mammal classification method based on acoustic features. In Proceedings of the 2021–2022 Academic Conference of the Hydroacoustics Branch of the Acoustical Society of China, Qingdao, China, 15 August 2022; South China Sea Research Station, Institute of Acoustics, Chinese Academy of Sciences: Beijing, China, 2022; pp. 432–435.
10. Zhang, X.; Yang, P.; Wang, Y.; Shen, W.; Yang, J.; Wang, J.; Ye, K.; Zhou, M.; Sun, H. A novel multireceiver sas rd processor. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4203611. [CrossRef]
11. Zhang, X.; Yang, P.; Sun, M. Experiment results of a novel sub-bottom profiler using synthetic aperture technique. *Curr. Sci.* **2022**, *122*, 461. [CrossRef]
12. Zhang, X.; Yang, P.; Sun, H. Frequency-domain multireceiver synthetic aperture sonar imagery with Chebyshev polynomials. *Electron. Lett.* **2022**, *58*, 995–998. [CrossRef]
13. Zhang, X. An efficient method for the simulation of multireceiver SAS raw signal. *Multimed. Tools Appl.* **2023**, *83*, 37351–37368. [CrossRef]
14. Aslam, M.A.; Zhang, L.; Liu, X.; Irfan, M.; Xu, Y.; Li, N.; Zhang, P.; Jiangbin, Z.; Yaan, L. Underwater sound classification using learning based methods: A review. *Expert Syst. Appl.* **2024**, *255*, 124498. [CrossRef]
15. Bianco, M.J.; Gerstoft, P.; Traer, J.; Ozanich, E.; Roch, M.A.; Gannot, S.; Deledalle, C.-A. Machine learning in acoustics: Theory and applications. *J. Acoust. Soc. Am.* **2019**, *146*, 3590–3628. [CrossRef] [PubMed]
16. Zhang, X.; Yang, P.; Feng, X.; Sun, H. Efficient imaging method for multireceiver SAS. *IET Radar Sonar Navig.* **2022**, *16*, 1470–1483. [CrossRef]
17. Wang, H.; Zhang, W.; Ren, P. Self-organized underwater image enhancement. *ISPRS J. Photogramm. Remote Sens.* **2024**, *215*, 1–14. [CrossRef]
18. Wang, H.; Zhang, W.; Bai, L.; Ren, P. Metalantis: A Comprehensive Underwater Image Enhancement Framework. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5618319. [CrossRef]
19. Wang, H.; Sun, S.; Bai, X.; Wang, J.; Ren, P. A reinforcement learning paradigm of configuring visual enhancement for object detection in underwater scenes. *IEEE J. Ocean. Eng.* **2023**, *48*, 443–461. [CrossRef]
20. Wang, H.; Sun, S.; Chang, L.; Li, H.; Zhang, W.; Frery, A.C.; Ren, P. INSPIRATION: A reinforcement learning-based human visual perception-driven image enhancement paradigm for underwater scenes. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108411. [CrossRef]
21. Shiu, Y.; Palmer, K.J.; Roch, M.A.; Fleishman, E.; Liu, X.; Nosal, E.-M.; Helble, T.; Cholewiak, D.; Gillespie, D.; Klinck, H. Deep neural networks for automated detection of marine mammal species. *Sci. Rep.* **2020**, *10*, 607. [CrossRef]
22. Griffiths, E.T.; Archer, F.; Rankin, S.; Keating, J.L.; Keen, E.; Barlow, J.; Moore, J.E. Detection and classification of narrow-band high frequency echolocation clicks from drifting recorders. *J. Acoust. Soc. Am.* **2020**, *147*, 3511–3522. [CrossRef]
23. Cai, W.; Zhu, J.; Zhang, M.; Yang, Y. A parallel classification model for marine mammal sounds based on multi-dimensional feature extraction and data augmentation. *Sensors* **2022**, *22*, 7443. [CrossRef] [PubMed]
24. Duan, D.X.; Jiang, Y.; Liu, Z.W.; Yang, C.M.; Lv, L.G. Echolocation signal detection method based on image processing. *Adv. Mar. Sci.* **2022**, *40*, 145–153.
25. Cominelli, S.; Bellin, N.; Brown, C.D.; Rossi, V.; Lawson, J. Acoustic features as a tool to visualize and explore marine soundscapes: Applications illustrated using marine mammal passive acoustic monitoring datasets. *Ecol. Evol.* **2024**, *14*, e10951. [CrossRef] [PubMed]
26. Widrow, B.; McCool, J.; Ball, M. The complex LMS algorithm. *Proc. IEEE* **1975**, *63*, 719–720. [CrossRef]
27. Zhang, X.; Yang, P.; Wang, Y.; Shen, W.; Yang, J.; Ye, K.; Zhou, M.; Sun, H. LBF-based CS algorithm for multireceiver SAS. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1502505. [CrossRef]
28. Liu, J.T.; Xi, B.; Jiang, H. Performance analysis of adaptive filters based on LMS algorithm. *Nav. Electron. Eng.* **2021**, *41*, 36–39.

29. Brummund, M.K.; Sgard, F.; Petit, Y.; Laville, F. Three-dimensional finite element modeling of the human external ear: Simulation study of the bone conduction occlusion effect. *J. Acoust. Soc. Am.* **2014**, *135*, 1433–1444. [CrossRef]
30. Rajaby, E.; Sayedi, S.M. A structured review of sparse fast Fourier transform algorithms. *Digit. Signal Process.* **2022**, *123*, 103403. [CrossRef]
31. Zheng, F.; Zhang, G.; Song, Z. Comparison of different implementations of MFCC. *J. Comput. Sci. Technol.* **2001**, *16*, 582–589. [CrossRef]
32. Xing, W.; Tang, X.; Zhou, Y.; Zhang, C.; Pan, Z. A review of delay-Doppler domain channel estimation methods for OTFS. *J. Commun.* **2022**, *43*, 188–201.
33. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
34. Lavanya, G.; Teja, U.S.M.P.; Dussa, V.K.; Reddy, A.L.; Nitish, Y. Classification of Underwater Fish Species Using Custom-Built Deep Learning Architectures. In Proceedings of the International Conference on Data Science and Applications, Jaipur, India, 14–15 July 2023; Springer Nature Singapore: Singapore, 2023.
35. Altarez, R.D. Faster R–CNN, RetinaNet and Single Shot Detector in different ResNet backbones for marine vessel detection using cross polarization C-band SAR imagery. *Remote Sens. Appl. Soc. Environ.* **2024**, *36*, 101297. [CrossRef]
36. Manikandan, D.L.; Santhanam, S.M. Parallel desires: Unifying local and semantic feature representations in marine species images for classification. *Mar. Geophys. Res.* **2024**, *45*, 16. [CrossRef]
37. Cabello-Solorzano, K.; Ortigosa de Araujo, I.; Peña, M.; Correia, L.; JTallón-Ballesteros, A. The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis. In Proceedings of the International Conference on Soft Computing Models in Industrial and Environmental Applications, Salamanca, Spain, 5–7 September 2023; Springer Nature Switzerland: Cham, Switzerland, 2023.
38. Sayigh, L.; Daher, M.A.; Allen, J.; Gordon, H.; Joyce, K.; Stuhlmann, C.; Tyack, P. The Watkins marine mammal sound database: An online, freely accessible resource. In *Proceedings of Meetings on Acoustics*; AIP Publishing: Melville, NY, USA, 2016; Volume 27.
39. Du, L.; Wang, Z.; Lv, Z.; Han, D.; Wang, L.; Yu, F.; Lan, Q. A Method for Underwater Acoustic Target Recognition Based on the Delay-Doppler Joint Feature. *Remote Sens.* **2024**, *16*, 2005. [CrossRef]
40. Deng, X.; Liu, Q.; Deng, Y.; Mahadevan, S. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Inf. Sci.* **2016**, *340–341*, 250–261. [CrossRef]
41. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef]
42. Ghani, B.; Denton, T.; Kahl, S.; Klinck, H. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Sci. Rep.* **2023**, *13*, 22876. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

Source Range Estimation Using Linear Frequency-Difference Matched Field Processing in a Shallow Water Waveguide

Penghua Song, Haozhong Wang *, Bolin Su, Liang Wang and Wei Gao

School of Marine Technology, Ocean University of China, Qingdao 266100, China; sph2020@stu.ouc.edu.cn (P.S.); subolin@stu.ouc.edu.cn (B.S.); wanglianger@ouc.edu.cn (L.W.); gaowei@ouc.edu.cn (W.G.)
* Correspondence: coolice@ouc.edu.cn

Abstract: Matched field processing (MFP) is an established technique for source localization in known multipath acoustic environments. Unfortunately, in many situations, imperfect knowledge of the actual propagation environment and sidelobes due to modal interference prevent accurate propagation modeling and source localization via MFP. To suppress the sidelobes and improve the method's robustness, a linear frequency-difference matched field processing (LFDMFP) method for estimating the source range is proposed. A two-neighbor-frequency high-order cross-spectrum between the measurement and the replica of each hydrophone of the vertical line array is first computed. The cost function can then be derived from the dual summation or double integral of the high-order cross-spectrum with respect to the depth of the hydrophones and the candidate sources of the replicas, where the range that corresponds to the minimum is the optimal estimation. Because of the larger modal interference distances, LFDMFP can efficiently provide only one optimal range within the same range search interval rather than some conventional matched field processing. The efficiency of the presented method was verified using simulations and experiments. The LFDMFP unambiguously estimated the source range in two experimental datasets with average relative errors of 2.2 and 1.9%.

Keywords: normal mode; matched field processing; linear frequency-difference matched field processing; source range estimation; ocean acoustic waveguide

Citation: Song, P.; Wang, H.; Su, B.; Wang, L.; Gao, W. Source Range Estimation Using Linear Frequency-Difference Matched Field Processing in a Shallow Water Waveguide. *Remote Sens.* **2024**, *16*, 3529. https://doi.org/10.3390/rs16183529

Academic Editors: Xuebo Zhang, Haiyong Xu, Jingyue Lu, Lei Zhang, Marc Pinto and Farman Ali

Received: 12 August 2024
Revised: 12 September 2024
Accepted: 13 September 2024
Published: 23 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Matched field processing (MFP) was developed in the mid-1970s [1] and remained a popular source localization technique through to the 1990s [2]. MFP can be successful under low signal-to-noise ratios (SNRs) when the actual propagation environment is well known. Conventional matched field processing (CMFP) is the simplest scheme. It is a cross-correlation technique developed for matching the values of the replica field computed with the measured values at the output of a vertical linear array. Different from beamforming technology, MFP incorporates the physical multipath effects through the simulated replica field to localize the source.

However, the primary limitation of MFP arises from two aspects: the actual-to-model mismatch and sidelobes due to modal interference. The former usually leads to a lack of robustness in the MFP source localization estimation and increases in severity with an increasing frequency and array-to-source range. Even for a low frequency, the environmental mismatch may induce phase fluctuations and limit the utility of MFP. Prior research [3] indicated that 1 kHz is an approximate nominal upper frequency limit for the successful application of CMFP in shallow water.

Adaptive techniques can provide a higher resolution, such as the minimum variance distortionless response (MVDR) [4,5], multiple constraint method (MCM) [6], and multiple signal classification (MUSIC) [7], which all offer higher source localization resolution than CMFP; however, these techniques based on the direct use of the complex field are sensitive

to environmental mismatch, especially for a high frequency, where incomplete knowledge of the environment can lead to a serious decline in localization performance.

To improve the tolerance of MFP for environmental mismatch, some robust techniques or algorithms based on CMFP were proposed. These include broadband matched field processing [8], matched mode processing (MMP) [9,10], matched beam processing (MBP) [11,12], compressed matched field processing [13–15], frequency-difference matched field processing (FDMFP) [16–19], machine learning techniques [20], multiple constraints methods [21], and sector-focused stability [22]. MMP and MBP specifically transform measurement data from the element domain to the modal and beam domains, respectively, before conducting matching procedures. Compared with CMFP, MMP exhibits reduced susceptibility to environmental mismatch, while MBP mitigates the seabed's impact. FDMFP, which was proposed by Worhtmann in 2015, expanded upon frequency-difference beamforming techniques [23] to adapt MFP for high-frequency source localization in shallow water. The FDMFP technique is realized by matching the autoproduct, which is a quadratic product of the measured field at two nearby frequencies (interval is $\Delta\omega$), with the frequency-difference replicas based on the low-frequency propagation model. It was shown by both simulation and experimental results that FDMFP can be more robust against environmental mismatch than CMFP.

The abovementioned improves the robustness, but to some extent, the resolution of the results is sacrificed. More importantly, the problem of periodic sidelobes caused by modal interference still exists. To suppress the periodic sidelobes and improve the robustness, a linear formulation of FDMFP is proposed in this paper, which differs from the physical mechanism of FDMFP proposed by Worthmann [16]. The formulation is based on normal mode theory, in which the replicas of two nearby frequencies are calculated in the same way as CMFP without recourse to the frequency difference replicas. The distinct advantage of LFDMFP is that we can provide a clear explanation that the mode pair interference distances of LFDMFP are far larger than those of CMFP (the ratio of the modal interference distances of LFDMFP to those of CMFP is approximately $\beta\omega/\Delta\omega$). Therefore, LFDMFP can efficiently provide an optimal range estimation within the distance search interval rather than some optimal ranges of the CMFP. Although the range resolution is degraded, the robustness of LFDMFP to the environmental fluctuation is enhanced. Furthermore, the depth estimation of the source could be evaluated by utilizing the completeness of the normal modes at the optimal range. The formulation was confirmed by numerical simulations and experimental results.

The main contributions of this study are summarized as follows:

(1) Linear frequency-difference matched field processing for estimating the source range is proposed. The LFDMFP utilizes larger modal interference distances to effectively solve the periodic sidelobes, thus greatly reducing the occurrence of "false targets".

(2) The difference processing enhances the tolerance of LFDMFP to environmental mismatches of fluctuations and improves the robustness.

(3) The performance of the LFDMFP was analyzed using simulations, and the effectiveness of the LFDMFP was verified using data from two offshore experiments.

The remainder of this article is organized into four sections. The mathematical derivation and theoretical underpinnings of the proposed method are presented in Section 2, which also includes the specifics for LFDMFP. Section 3 presents the method performance analysis. Section 4 documents the results of this method based on experimental data recorded during the Sound Propagation Experiment in Laoshan Bay in 2005 and the Sound Propagation Experiment in the South China Sea in 2019. Section 5 provides the conclusions of this study.

2. Theory and Algorithms

2.1. The Conventional MFP Method

For simplicity, a range-independent stratified ocean environment was considered in this study. An N-element vertical line array (VLA) of hydrophones is disposed at a distance of r_0 from a source at the depth of z_0. The complex sound pressure of the frequency ω received by the ith hydrophone is given by

$$P(\omega, z_i, r_0, z_0) = S(\omega)G(\omega, z_i, r_0, z_0) + N(\omega, z_i) \tag{1}$$

where $S(\omega)$ is the spectrum of the source, $N(\omega, z_i)$ is the ambient noise considered as the zero-mean Gaussian noise, and $G(\omega, z_i, r_0, z_0)$ denotes the Green's function between the source and the ith hydrophone. The normal mode representation of the replicas (Green's function) in the far field of the source in cylindrical coordinates can be expressed by

$$G(\omega, z_i, r_s, z_s) = \frac{i\rho(z_s)}{\sqrt{8\pi r_s}} e^{i\frac{\pi}{4}} \sum_n \frac{\Phi_n(z_s)\Phi_n(z_i)}{\sqrt{k_n(\omega)}} e^{i[k_n(\omega)+i\alpha_n(\omega)]r_s} \tag{2}$$

where $k_n(\omega) + i\alpha_n(\omega)$ and $\Phi_n(z)$ are the complex eigenvalue and eigenfunction of the nth normal mode, respectively; z_s is the depth of the candidate source and r_s is the distance to the VLA.

CMFP for localizing a source is based on matching the acoustic pressure field measured at the array of sensors with the modeled replica fields $G(\omega, z_i, r_s, z_s)$ computed for the acoustic waveguide via a numerical propagation model over a grid of possible source positions (r_s, z_s), with the position estimate taken to be the grid point of the maximum match. The normalized ambiguity function of CMFP is defined as

$$E(r_s, z_s; \omega) = \frac{\sum_{i=1}^{N} G^*(\omega, z_i, r_s z_s) P(\omega, z_i, r_0, z_0)}{\sqrt{\sum_{i=1}^{N} |G(\omega, z_i, r_s, z_s)|^2 \sum_{i=1}^{N} |P(\omega, z_i, r_0, z_0)|^2}} \tag{3}$$

where the range and depth corresponding to the maximum of $E(r_s, z_s; \omega)$ are the optimal estimations of the source location.

In the far field of the point source, the low-frequency acoustic field comprises modal components that propagate long distances at shallow grazing angles and vertical components that rapidly decay because of the strong interaction with the bottom. It is recognized that the MFP is a generalized beamforming process, where there exists periodic ambiguity (sidelobes) due to modal interferences [24]:

$$\hat{r}_s = r_0 - l\lambda_{mn} (l = 0, \pm 1, \pm 2, \dots\dots) \tag{4}$$

where $\lambda_{mn} = 2\pi/|k_n - k_m|$ is the modal interference distance between the mth and nth modes.

2.2. The LFDMFP Method

Broadband-matched field processing, while capable of partially suppressing sidelobes, requires substantial computational resources due to the need for multiple calculations of replica fields across the band. Instead of broadband-matched field processing, we reformulated LFDMFP to resolve the problem of sidelobes. Initially, the cross-spectrum of the ambiguity functions of two neighbor frequencies is computed for each candidate location of the source. The cross-spectrum process is called the high-order cross-spectrum. The

ambiguity function $\Gamma(r_s;\omega,\Delta\omega)$ of LFDMFP is defined by the integral along the candidate source depth. The process of integration is called dual summation or a double integral:

$$\begin{aligned}\Gamma(r_s;\omega,\Delta\omega) &= \left|\int_0^H E(r_s,z_s;\omega+\Delta\omega)E^*(r_s,z_s;\omega)dz_s\right| \\ &\approx A^2\left|\sum_n \frac{\Phi_n^2(z_0)e^{-[a_n(\omega+\Delta\omega)+a_n(\omega)](r_0+r_S)}}{k_n(\omega)k_n(\omega+\Delta\omega)}e^{i[k_n(\omega+\Delta\omega)-k_n(\omega)](r_0-r_S)}\right|\end{aligned} \tag{5}$$

where $A = 1/\sqrt{\sum_l \phi_l^2(zs)/k_l \sum_m \phi_m^2(zs)/k_m}$ is the normalization coefficient. In the derivation above, the orthogonality of the eigenfunctions of the normal modes is used.

$$\int_0^H \Phi_n(\omega,z_s)\Phi_m(\omega+\Delta\omega,z_s)dz_s \approx \delta_{nm} \tag{6}$$

Provided that the eigenfunctions are weakly dependent on the frequency or $\Delta\omega$ is small ($\omega/\Delta\omega \gg 1$), Equation (6) is a good approximation. The eigenfunctions can then be approximated by neglecting the high-order terms of the Taylor expansion and represented as

$$k_n(\omega+\Delta\omega) \approx k_n(\omega) + \frac{\Delta\omega}{v_{gn}(\omega)}\cdots \tag{7}$$

where $v_{gn} = \partial\omega/\partial k_n$ denotes the group velocity of the nth mode. These properties are well satisfied if ω is significantly larger than the cutoff frequency of the nth mode. Substituting Equation (7) into Equation (5) gives the approximation result of the following ambiguity function of LFDMFP:

$$\Gamma(r_s;\omega,\Delta\omega) \approx A^2\left|\sum_n \frac{\Phi_n^2(z_0)e^{-[\alpha_n(\omega+\Delta\omega)+\alpha_n(\omega)](r_0+r_S)}}{k_n(\omega)k_n(\omega+\Delta\omega)}e^{i\frac{\Delta\omega(r_0-r_S)}{v_{gn}}}\right| \tag{8}$$

The key difference between LDFMFP and CMFP lies in the encoding of the source range information within exponential terms. The source range information in LDFMFP is encoded in the exponential terms of $e^{i\Delta\omega(r_0-r_s)/v_{gn}}$ rather than $e^{ik_n(r_0-r_s)}$; then, the modal interference cycles are given by

$$\Lambda_{mn} = \frac{2\pi}{\Delta\omega\left(\frac{1}{v_{gm}}-\frac{1}{v_{gn}}\right)} \tag{9}$$

Furthermore, for a shallow water waveguide, there is the following approximate estimation:

$$\frac{\Lambda_{mn}}{\lambda_{mn}} = \left|\frac{k_m-k_n}{\Delta\omega\left(\frac{1}{v_{gm}}-\frac{1}{v_{gn}}\right)}\right| = \left|\frac{\omega\left(\frac{1}{v_{pm}}-\frac{1}{v_{pn}}\right)}{\Delta\omega\left(\frac{1}{v_{gm}}-\frac{1}{v_{gn}}\right)}\right| \approx \beta\frac{\omega}{\Delta\omega} \tag{10}$$

The notion of the waveguide invariant is used in Equation (10). In a typical shallow water waveguide, $\beta \approx 1$. Provided that $\Delta\omega \ll \omega$, it is clear that the modal interference distances in the LFDMFP processor are much larger than those of the CMFP. This is the inherent physical principle that causes LFDMFP to be robust against environment fluctuation because Λ_{mn} may be larger than the spatial correlation radius of environment fluctuations in comparison with λ_{mn}. As a penalty, the range resolution of LFDMFP is degraded due to $\Lambda_{mn} \gg \lambda_{mn}$. The influence of the frequency difference on the range estimation is analyzed in Section 3.

2.3. Cost Function

2.3.1. The Cost Function of the Range Estimation

The feasibility of the reformulated LFDMFP was first tested using a numerical simulation in the shallow water waveguide with a negative thermocline shown in Figure 1a. The depth of the water was 60 m, and the sound speed, density, and attenuation of the semi-infinite fluid bottom were set to 1800 m/s, 1.8 g/cm^3, and 1 dB/λ , respectively. The source was arranged at a depth of 15 m and a 60-receiver 1 m interval VLA was deployed at 10 km away from the source. The sound fields at two frequencies of 300 and 310 Hz were computed using the software KRAKENC with the former eight normal modes. The computational grid of the replicas was set to 1 m in depth and 100 m in range.

In this numerical example, the modal interference distances of the first two normal modes were $\lambda_{12} \approx 0.75$ km and $\Lambda_{12} \approx 22.2$ km. Theoretically, the modal interference distance ratio of LFDMFP and CMFP was about 30. In the simulation environment, the modal interference distance ratio of the first two normal modes was about 29.6, which was consistent. Figure 1b shows that in $\Gamma(r_s;\omega,\Delta\omega)$, there were some "spurs" because the orthogonality of normal modes is destroyed by only concerning the sound field in the water. To cancel the "spurs", the cost function to estimate the source range was redefined as

$$Cost(r_s) = |2 * \Gamma(r_s,\omega,\Delta\omega) - \Gamma_\omega(r_s) - \Gamma_{\omega+\Delta\omega}(r_s)| \tag{11}$$

where $\Gamma_\Omega(r_s) = \left|\int_0^H E(r_s,z_s;\Omega)E^*(r_s,z_s;\Omega)dz_s\right| \ldots (\Omega = \omega, \omega+\Delta\omega)$ are called the auto-spectra of the frequencies of ω and $\omega+\Delta\omega$, respectively. $\Gamma(r_s;\omega,\Delta\omega)$ is called the cross-spectrum between ω and $\omega+\Delta\omega$.

Figure 1c shows how the cost function in Equation (11) changed with the range. Because the "spurs" inherent in the auto-spectra $\Gamma_\omega(r_s)$ and $\Gamma_{\omega+\Delta\omega}(r_s)$ and the cross-spectrum $\Gamma(r_s;\omega,\Delta\omega)$ similarly manifested themselves, $Cost(r_s)$ could cancel most of the "spurs" when r_s was close to the real range of the source (Figure 1c). However, the main lobe of the cost function became so fat that the range resolution was degraded. Inspired by Figure 1b, when the estimated range was closer to the real range of the source, the coherent components of the auto-spectra and cross-spectrum decreased, and the difference between them became smaller. When r_s equals r_0, $Cost(r_s)$ reaches the minimum and the estimation of $\hat{r}_s$ is

$$\hat{r}_s = min\{Cost(r_s)\} \tag{12}$$

When the cost function of LFDMFP is similar to that of CMFP, the best source range estimation corresponds to the maximum value of the ambiguity function for observation. The cost function in Equation (12) is normalized by Equation (13), and the final cost function is obtained. To distinguish Equation (12) from Equation (14), the cost function only refers to Equation (14) in the subsequent simulation and experimental verification part:

$$\cos t(r_s) = \min(Cost(r_s))/Cost(r_s) \tag{13}$$

$$\hat{r}_s = \max_{r_s}\{\cos t(r_s)\} \tag{14}$$

Figure 1d illustrates that the performance of the reformulated cost function in Equation (14) changed with the range, where the black line represents the LFDMFP and the red line represents the slice of ambiguity function at the true depth of the CMFP. The resolution of the CMFP was better than the LFDMFP. However, the strong sidelobes certainly disturbed the best range estimation, especially in the condition of low SNRs. Based on the simulation results depicted in Figure 1, the newly formulated LFDMFP effectively mitigated the issue of multiple solutions by ensuring that only one prominent peak was observed within the range search interval, which enhanced the accuracy of the range estimation.

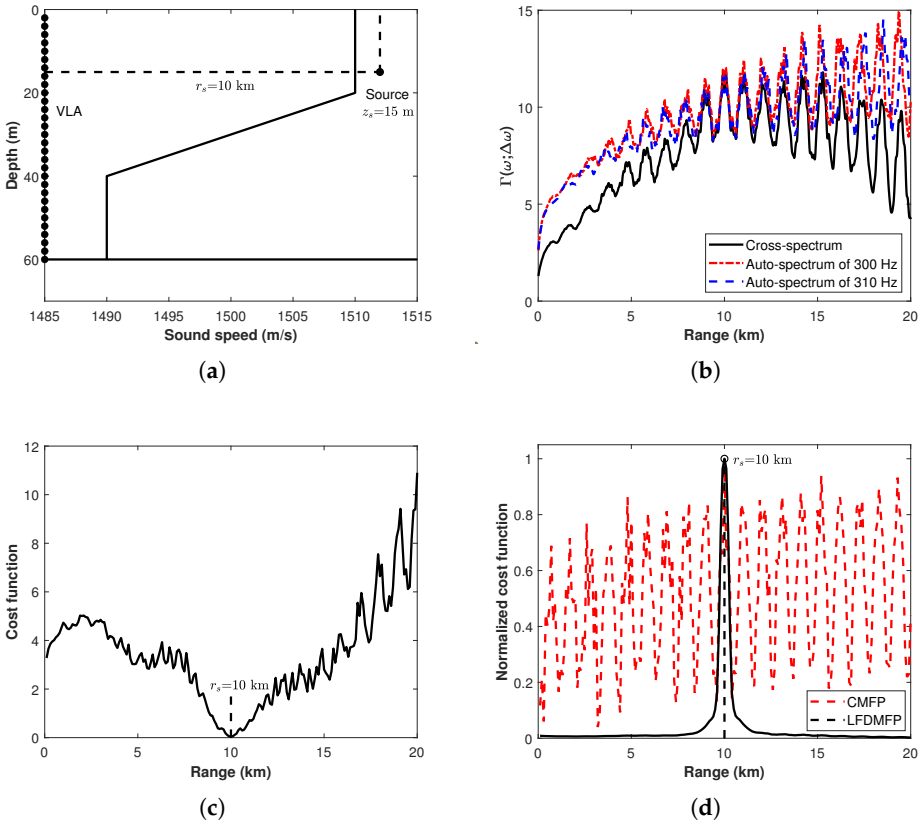


Figure 1. Simulation environment and localization results for the LFDMFP and CMFP. (**a**) The sound speed profile of the simulated environment. (**b**) The dot-dashed curve is the auto-spectrum of 300 Hz, the dashed curve is the auto-spectrum of 310 Hz, and the solid curve is the cross-spectrum between 300 and 310 Hz. (**c**) The cost function using Equation (11). (**d**) The change in the reformulated cost function using Equation (14) with range.

2.3.2. The Cost Function of the Depth Estimation

It is worth mentioning that the depth of the source can be evaluated using the completeness of the normal modes $\sum_n \Phi_n(z_s)\Phi_n(z_0) \approx \delta(z_s - z_0)$ at the optimal range in the case where the environment is known. The cost function of the depth estimation is defined as

$$\Xi(z_s;\omega) = |E(\hat{r}_s, z_s;\omega)| \propto \left| \sum_n \frac{\Phi_n(z_s)\Phi_n(z_0)}{k_n(\omega)} \right| \propto \frac{1}{k(\omega)}\delta(z_s - z_0) \tag{15}$$

According to Equation (15), the depth estimation in the simulation environment of Figure 1a is shown in Figure 2. Although LFDMFP cannot directly obtain the source depth estimation, after obtaining the range estimation, information on the source depth estimation can be obtained using Equation (15). The focus of this study was to explore the advantages of LFDMFP in source range estimation. The depth estimation is just additional information derived from Equation (15).

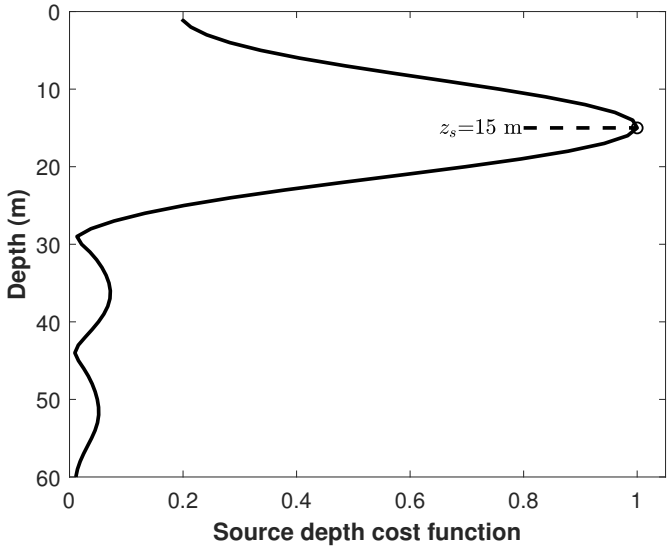


Figure 2. Source depth estimation using Equation (15).

2.4. Algorithm

The pseudocode implementation of the LFDMFP is presented in Algorithm 1. In the following sections, the results of the LFDMFP simulation performance analysis are given in Section 3 and subsequently validated with experimental data in Section 4.

Algorithm 1 LFDMFP

Input: Time domain pressure $p(t)$ recorded by VLA and sound speed profiles
Output: Source range $\hat{r}_s$ and source depth $\hat{z}_s$

1: Compute the power spectrum using the time domain pressure recorded by the VLA and select the two adjacent frequencies ω_1 and ω_2 and their spectra $P(\omega, z_i, r_0, z_0) \ldots \omega = \omega_1, \omega_2$.
2: Divide the search grid and compute the replica field $G(\omega, z_i, r_s, z_s) \ldots \omega = \omega_1, \omega_2$ using the propagation model.
3: Compute the high-order cross-spectra using Equations (3) and (5):
$E(\omega, z_i, r_s, z_s) = \int G * (\omega, z_i, r_s, z_s) P(\omega, z_i, r_0, z_0) dz_i \ldots \omega = \omega_1, \omega_2$.
$\Gamma(r_s; \omega, \Delta\omega) = \int E(\omega, z_i, r_s, z_s) dz_s \ldots \omega = \omega_1, \omega_2; \Delta\omega = \omega_2 - \omega_1$.
4: Compute the cost function of the range estimation using Equations (11) and (13):
$Cost(r_s) = |2 * \Gamma(r_s, \omega, \Delta\omega) - \Gamma_\omega(r_s) - \Gamma_{\omega+\Delta\omega}(r_s)|$.
$\cos t(r_s) = \min(Cost(r_s))/Cost(r_s)$.
5: Compute the source range: $\hat{r}_s = \max_{r_s} \{\cos t(r_s)\}$.
6: Compute the source depth: $\hat{z}_s = |E(\omega, z_i, \hat{r}_s, z_s)|$.

3. Simulation and Performance Analysis

LFDMFP is constrained by various factors, such as the signal-to-noise ratio (SNR), aperture of the linear array, and frequency difference ($\Delta\omega$). Among these factors, some exhibit similar performance to LFDMFP compared with CMFP, such as the SNR and aperture of the linear array, while the frequency difference is unique to LFDMFP. The previous study [25] analyzed the influence of the aperture of a VLA (including the array length and the number of elements) on CMFP. To adequately sample the pressure field, a fully spanning aperture is necessary. Furthermore, the aperture of the VLA affects the orthogonality of the normal mode shown in Equation (6). Once the array satisfies the minimum length requirement, the performance is largely determined by the number of array elements. To provide a unique representation of each mode, the number of elements should be at least equal to the number of modes. Based on this, a VLA with a fully spanning aperture and a number of elements that far exceeded the number of modes was set up in

this simulation experiment. We discuss the influence of the SNR and frequency difference on the method proposed in this paper.

3.1. The Influence of the SNR

In this simulation, the ocean sound channel environment and geometry were consistent with the waveguide environment shown in Figure 1a. The source was arranged at a depth of 15 m and the 60-receiver 1 m interval VLA was deployed 10 km away from the source. The sound fields at two frequencies of 300 and 310 Hz were computed using the software KRAKENC with the former eight normal modes. Figure 3 illustrates the influence of the SNR on the performance of the method, with 1000 Monte Carlo experiments conducted for each SNR. The SNR was set to be uniformly distributed between −20 and 20 dB; the noise type was complex Gaussian white noise and the power signal-to-value ratio was applied, as represented by Equation (17). Figure 3 demonstrates the error of range estimation of the LFDMFP and CMFP under different SNR conditions, with the horizontal axis representing the SNR and the vertical axis representing the RMSE of the range estimation.

$$noise = \sigma A e^{2\pi i B} \tag{16}$$

where σ is the standard deviation of the noise, A satisfies the standard normal distribution, and B is a uniformly distributed random number between 0 and 1. The SNR is defined as

$$SNR = 10\lg[\frac{1}{N\sigma^2}\sum|p|^2] \tag{17}$$

where N is the number of arrays, and p denotes the complex sound pressure, as calculated by Equation (2). The error was computed using Equation (18):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{r}_s - r_0)^2}{n}} \tag{18}$$

According to Figure 3, both the LFDMFP and CMFP produced unsatisfactory range estimations at low SNRs. This was because the MFP necessitated precise spectral matching, which demands that the signal's distinct spectral features remain minimally influenced by noise. Consequently, at low SNRs, the introduction of random spectral variations can compromise the signal's spectral integrity, thereby complicating the matching process. Under low-SNR conditions, the method proposed in this paper exhibited stronger robustness compared with the CMFP; furthermore, the performance was enhanced, which was primarily due to LFDMFP's cross-spectral processing of adjacent frequencies' ambiguity functions, which ensured that noise from different frequencies remained non-interfering, and thus, contributed to its superior robustness against the Gaussian white noise.

3.2. The Influence of the Frequency Difference

The incorporation of the frequency difference imposes constraints on the performance of LFDMFP. Both excessively large and small frequency differences can adversely affect the accuracy of the range estimation. Due to the coupling relationship between the frequency differences, frequency, and waveguide environment, it is difficult to quantitatively determine the applicable range of the frequency difference. Hence, this section discusses the qualitative analysis of the impact of the frequency difference selection on LFDMFP. The frequency difference alters the orthogonality of the normal modes and the phase difference. In this simulation, the SNR was 10 dB, while the sound source and VLA positions, as well as the ocean sound channel environment and geometry, were consistent with the waveguide environment shown in Figure 1a. The lower frequency of the two neighbor frequencies was set to 300 Hz.

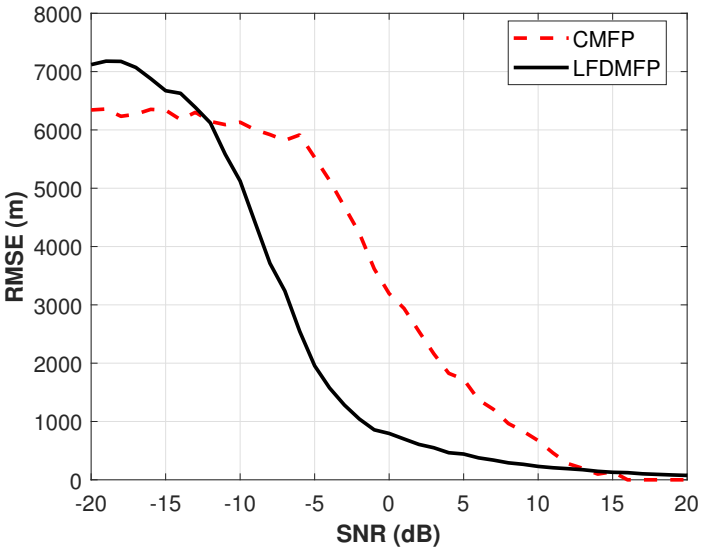


Figure 3. Influence of SNR on the two methods.

The frequency difference influences the phase difference of the normal modes, which plays a dominant role in the localization performance of LFDMFP. The Taylor expansion in Equation (7) requires that the frequency difference should not be too large. As the frequency difference increases, the modal interference in LFDMFP decreases, which weakens its ability to suppress periodic ambiguities. Conversely, when the frequency difference is too small, the approximations in Equation (19) apply, which leads to Equation (8) transforming into Equation (20), which represents a power spectral function at frequency ω. Consequently, the cost function of LFDMFP approaches a constant, thus rendering the method ineffective.

$$k_n(\omega + \Delta\omega) \approx k_n(\omega) + \frac{\Delta\omega}{v_{gn}(\omega)} \cdots \Rightarrow k_n(\omega + \Delta\omega) \approx k_n(\omega) \tag{19}$$

$$\Gamma(r_s; \omega, \Delta\omega) \approx A^2 |\sum_n \Phi_n^2(z_0)/k_n(\omega)k_n(\omega + \Delta\omega)| \tag{20}$$

On the other hand, the impact of the frequency difference on LFDMFP is its effect on the orthogonality of normal modes. The approximation in Equation (6) relies on the premise that the frequency difference should not be too large. Figure 4 illustrates the influence of six frequency differences (0.01, 1, 10, 50, 100, and 300 Hz) on the orthogonality with the eight normal modes. It is evident that as the frequency difference increased, the values of the off-diagonal elements grew, which compromised the orthogonality of the modes. Figure 5 shows the range estimation of the LFDMFP corresponding to the six frequency differences. Obviously, when the frequency difference was 0.01 or 1 Hz, the results in Figure 5a,b were similar to the theoretical analysis, which were manifested as the cost function changing irregularly with range. As the frequency difference increased, the LFDMFP accurately estimated the range of the sound source. When the frequency difference exceeded 100 Hz, sidelobes appeared, which affected the range estimation at low SNRs.

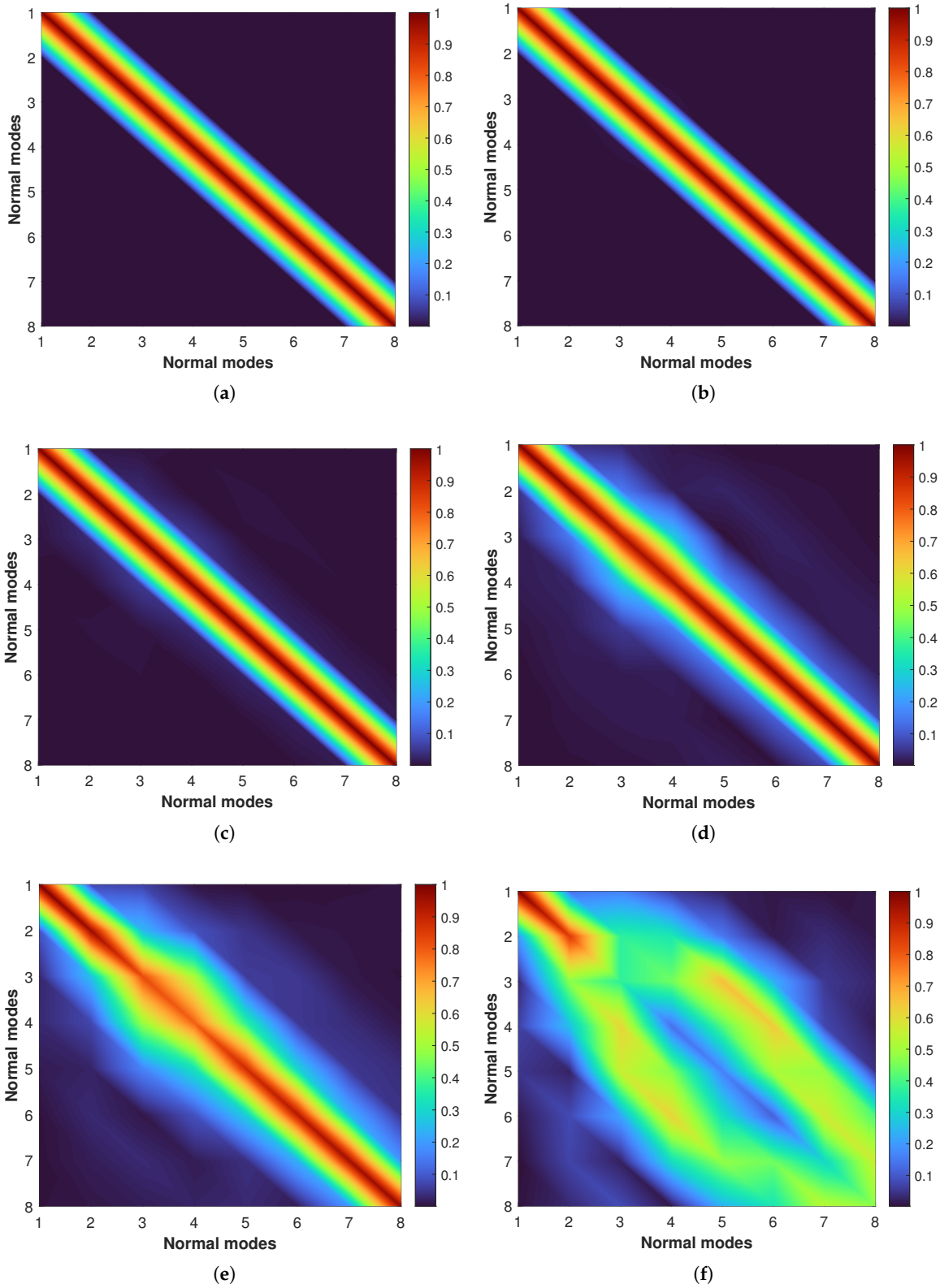


Figure 4. Orthogonality of normal modes at six different frequency differences: (**a**) 300 and 300.01 Hz, (**b**) 300 and 301 Hz, (**c**) 300 and 310 Hz, (**d**) 300 and 350 Hz, (**e**) 300 and 400 Hz, and (**f**) 300 and 600 Hz.

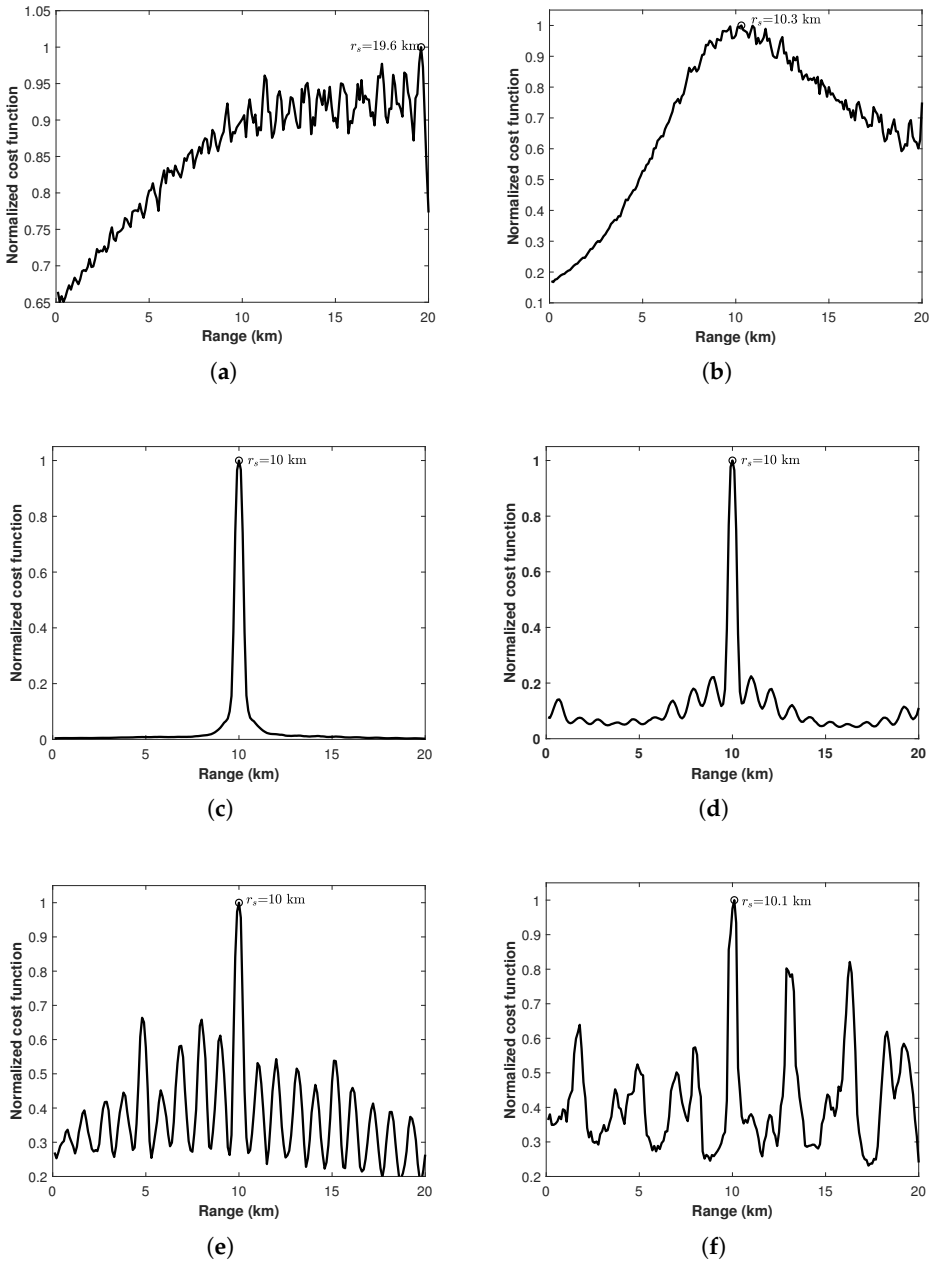


Figure 5. Range estimation of the LFDMFP at six different frequency differences: (**a**) 300 and 300.01 Hz, (**b**) 300 and 301 Hz, (**c**) 300 and 310 Hz, (**d**) 300 and 350 Hz, (**e**) 300 and 400 Hz, and (**f**) 300 and 600 Hz.

By fixing the lower frequency at 300 Hz and adjusting the frequency difference, the curve of the LFDMFP distance estimation error as a function of the frequency difference was obtained, as shown in Figure 6, where the horizontal axis represents the magnitude of the frequency difference and the vertical axis represents the error of the range estimation.

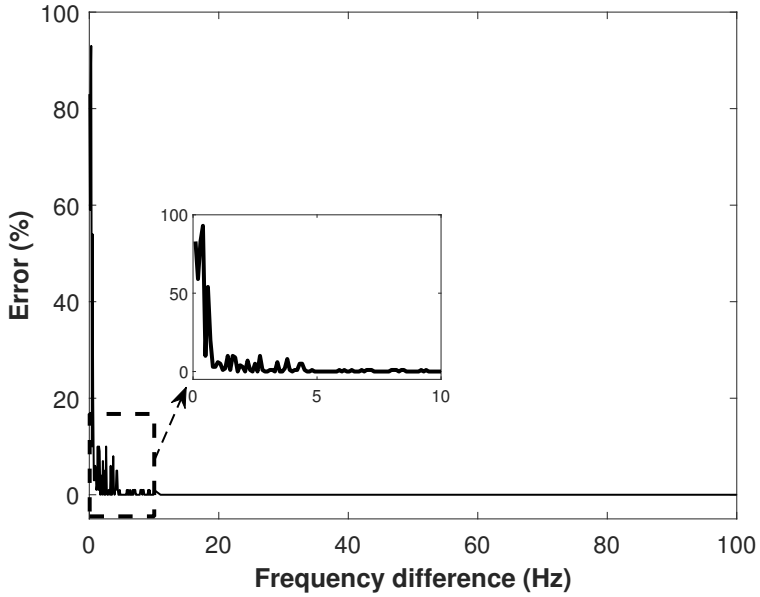


Figure 6. The influence of the frequency difference on the LFDMFP.

In this simulation, the effective range of the frequency difference was 5–100 Hz. When the frequency difference was small (less than 1 Hz), the cost function of the LFDMFP exhibited irregular variations with range, which rendered the range estimation of the source ineffective. As the frequency difference increased, the range estimation performance of the LFDMFP improved. When only considering the first two normal modes, 12 Hz was the upper limit frequency difference with only one modal interference distance within the range search interval. In this simulation, the effective mode of the normal mode was eight. As the number of normal modes increased, the interference span of the LFDMFP increased; therefore, the upper limit of the frequency difference also increased. Under high SNRs, the results shown in Figures 5 and 6 appeared. However, in actual waveguide environments, the SNR was low, and the large frequency difference led to high sidelobes, which could obscure the main lobe and cause false alarms. It is important to note that the range of frequency difference was dependent on the reference frequency and the waveguide environment. As the reference frequency increased, both the upper and lower bounds of the frequency difference range increased.

4. Experimental Results

The method presented in this paper was applied in two offshore experiments to show its effectiveness. The two experiments were the Qingdao Laoshan Bay Sound Propagation experiment conducted by the Hydroacoustic Laboratory of Ocean University of China and the Sound Propagation experiment in the South China Sea.

4.1. Sound Propagation Experiment in Laoshan Bay

The Hydroacoustic Laboratory at the Ocean University of China conducted a sound propagation experiment in August 2009 in Laoshan Bay of Qingdao. The experiment employed a dual-vessel operation mode, with the receiving ship anchored at station F to receive the transmitted signals. The receiving array was a VLA, while the launching ship emitted signals from various stations (H1–H6). Figure 7 illustrates the distribution of the experimental stations.

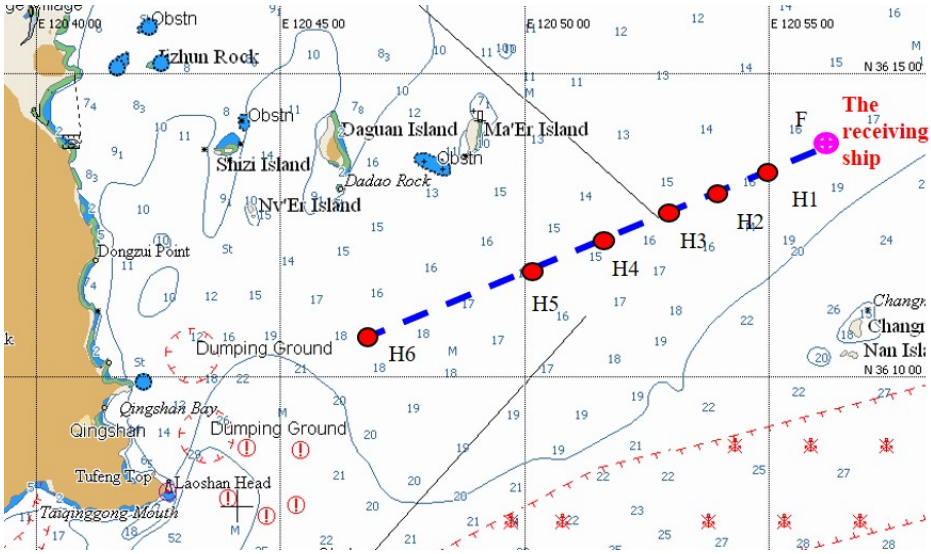
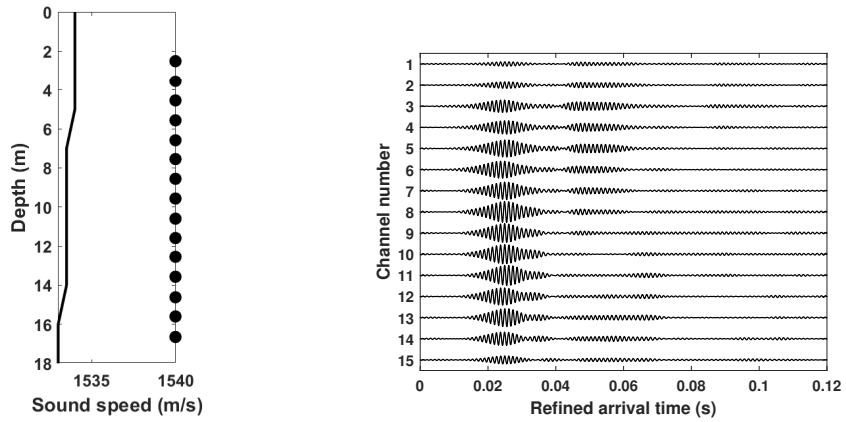


Figure 7. Source trajectory and array position.

The experimental environment could be fairly approximated to an 18 m water depth Pekeris waveguide with the sound speed profiles shown in Figure 8a. The acoustic parameters of the half-space sediment (c_b = 1596 m/s, ρ_b = 1.62 g/cm^3, α_b = 0.28 dB/λ) used here were obtained through the MFP inversion of the same data. The VLA had N = 15 sensors with a sampling rate of 12 kHz, which were uniformly spaced between 2.2 and 16.2 m. The source transmitted an M-sequence pseudorandom signal with a central frequency of 750 Hz and a bandwidth of 200 Hz. During this experiment, the depth of the source was maintained at 11 m. Figure 8b is a snapshot of the VLA-received signal and the transmitted M-sequence signal after matching filtering. Only two normal modes were observed.



(**a**) (**b**)

Figure 8. Laoshan Bay condition: (**a**) the sound speed profile of the acoustic propagation experiment and the depth of each receiver of the 15-element VLA labeled by dots; (**b**) a snapshot of the signals received by the VLA.

This section verifies the effectiveness of the presented method by presenting the results from analyzing the H3 station data. At this station, the real distance between the source and the VLA was 4.62 km. The selected frequency band ranged from 780 to 804 Hz with frequency intervals of 3 Hz. The grid used to calculate the replica field was set to a depth of 0.2 m and a distance of 10 m. The replica field was calculated every 3 Hz using the KRAKENC. Figure 9 shows the results of the two methods for the H3 station. At this station, the target source range was 4.62 km according to the GPS. Figure 9a shows the CMFP incoherent averaging results at the six frequencies, where the black box represents

the estimated range of the source. High sidelobes appeared within the range search interval, which hindered the true range estimation. The range estimation of the CMFP was 5.11 km, with an estimation error of 10.84%, which was a large deviation from the actual range. The range estimation of the LFDMFP is depicted in Figure 9b. The frequencies used for the LFDMFP were 780 and 786 Hz, and thus, a frequency difference of 6 Hz. The range estimation of the LFDMFP was 4.64 km, with an estimation error of 0.4%.

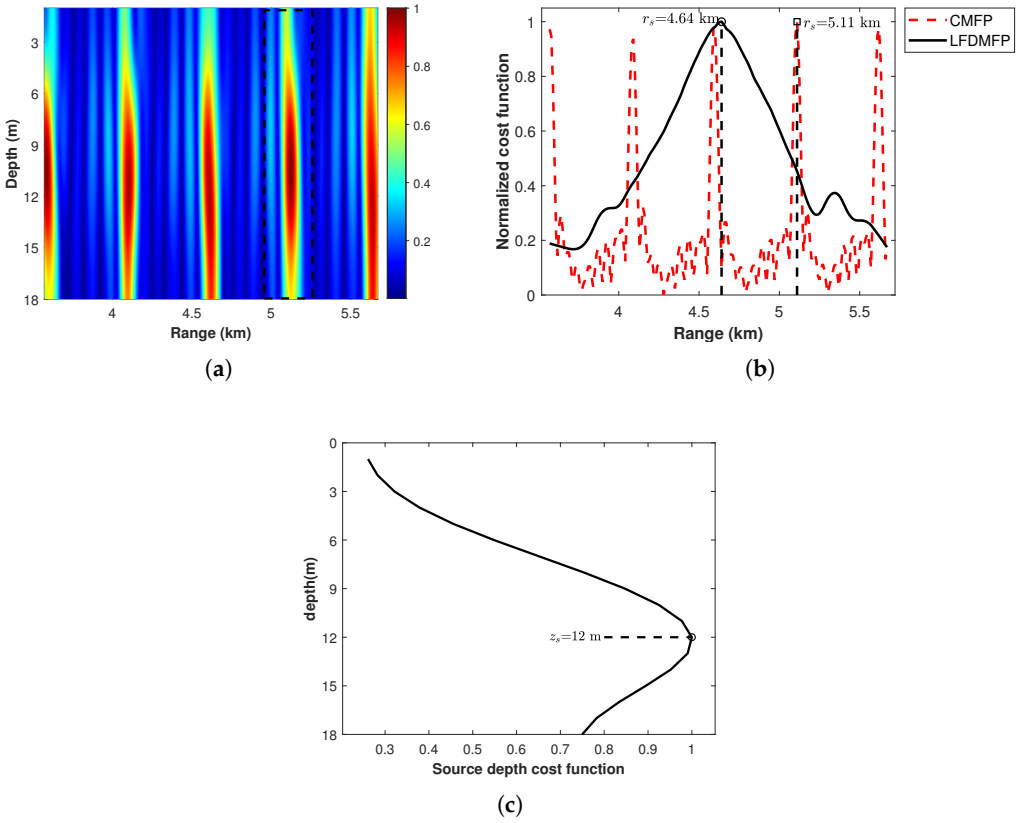


Figure 9. Localization results for the LFDMFP and CMFP: (**a**) the incoherent averaging ambiguity function of the CMFP; (**b**) the cost function of the LFDMFP at the frequency difference [780 Hz, 786 Hz]; (**c**) the depth of the source estimated by Equation (15).

Table 1 lists the dependence of the interference distance (Λ_{12}) on the frequency difference ($\Delta\omega$), which was computed using Equation (9) in the actual experiment environment. In this experiment, the interference distance λ_{12} of the first two normal modes at the frequency 780 Hz was approximated to 0.52 km (Figure 9a). Table 1 shows that $\Lambda_{12}/\lambda_{12}$ was proportional to $\omega/\Delta\omega$ and the coefficient of the proportionality was the waveguide invariant $\beta \approx 1.154$. When adjusting the frequency difference within the frequency range of the sound source signal and selecting multiple sets of data for the LFDMFP, the obtained range estimation results of the LFDMFP varied with the frequency difference, as shown in Figure 10, where the bar represents the error size corresponding to the frequency difference. According to the results in Figures 9 and 10, despite the modest resolution of the range estimation, the result was accurate and the error was less than that for the CMFP.

$\Phi_n(z_s)$ was calculated by using the sound speed profile displayed in Figure 8a, and the depth estimation was approximated to 12 m (Figure 9c).

Table 1. The dependence of Λ_{12} on the frequency difference.

$\omega, \omega + \Delta\omega$ (Hz)	Λ_{12} (km)	$\omega/\Delta\omega$	$\Lambda_{12}/\lambda_{12}$	β
780, 783	157	260	300	1.154
780, 786	77.6	130	150	1.154
780, 789	51.9	87	100	1.154
780, 792	39.3	65	75	1.154
780, 795	31.5	52	60	1.154

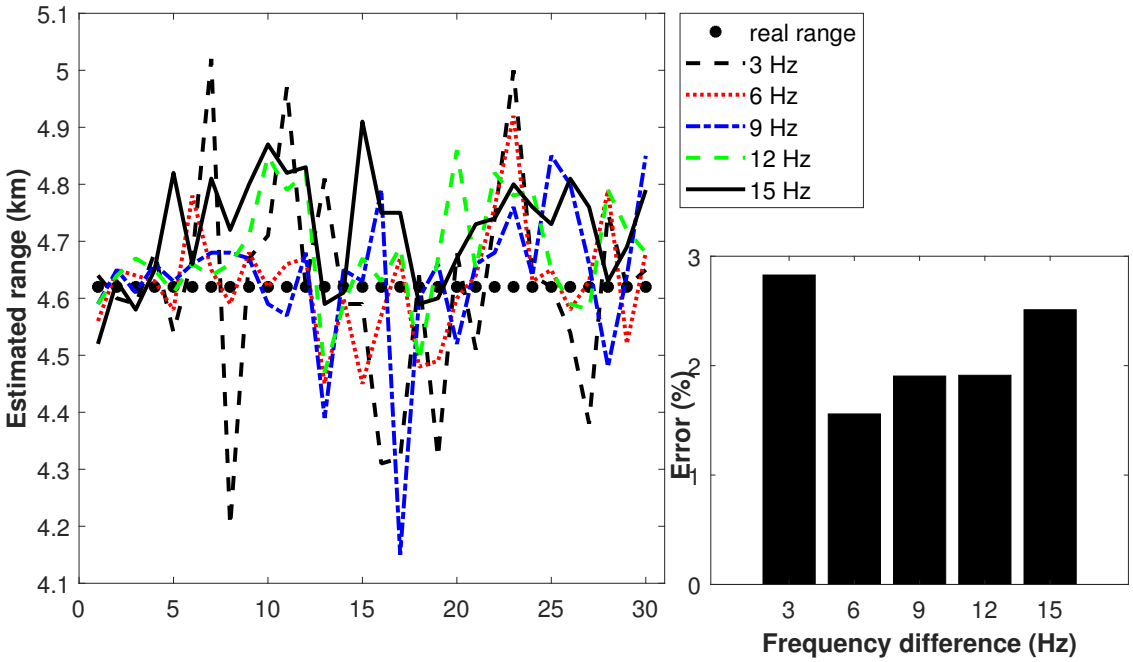
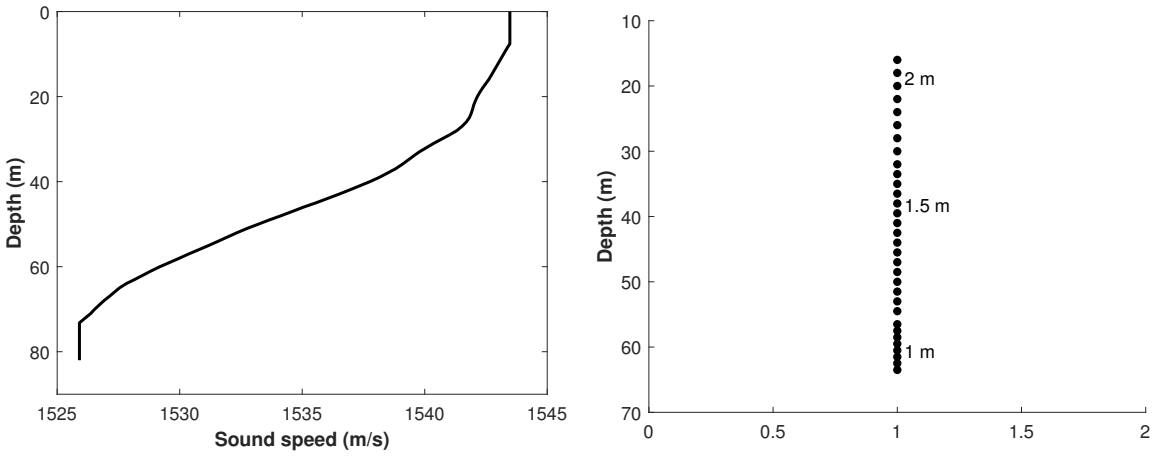


Figure 10. Range estimation and error of the LFDMFP at the different frequency differences.

4.2. Sound Propagation Experiment in the South China Sea

On 15 September 2019, an experiment that involved a towed sound source and a VLA was conducted in the South China Sea at a depth of 82 m. The sound speed of the water body in the experimental sea area is shown in Figure 11a, where the average sound speed was 1534 m/s. The sound speed, density, and attenuation of the sediment were 1573 m/s, 1.97 g/cm^3, and 0.75 dB/λ, respectively. The receiving ship was equipped with a 32-element VLA with non-uniform spacing. In the shallow water area, the array elements were spaced 2 m apart; in the bottom water area, they were 1 m apart; in the intermediate water area, they were 1.5 m apart. The specific depth of the VLA deployment is shown in Figure 11b. The launch ship towed a sound source with a central frequency of 300 Hz, which emitted a linear frequency modulation signal that ranged from 280 to 350 Hz. Throughout the towing, the sound source maintained a depth of approximately 31 m.

This section verifies the effectiveness of the LFDMFP method through experimental data from two stations with distances of 8.47 km and 13.71 km. Significant internal wave activity was observed in the latter. Within the selected frequency band (280–340 Hz), 21 frequencies with an interval of 3 Hz were selected, and the candidate source grid for replica calculation was configured in an area with a depth of 2 m and a distance of 10 m.



(**a**) (**b**)

Figure 11. The deployment depth of the VLA and sound speed in the water body in the experimental sea areas: (**a**) the sound speed profile of the acoustic propagation experiment; (**b**) the depths of the 32 VLA elements labeled by dots.

Figure 12 shows the comparison of the range estimation between the two methods when the sound source distance was 8.47 km. Figure 12a shows the incoherent averaged CMFP results at the band. The black circle marks the estimated position of the CMFP, with a range estimation and error of the CMFP of 9.16 km and 8.14%, respectively. Figure 12b shows the range estimation of the LFDMFP when the frequencies 297 and 301 Hz were used. The range estimation of the LFDMFP was 8.52 km, with an estimation error of 0.5%. By using the sound speed profile displayed in Figure 11a, the depth estimation was approximated to 34 m (Figure 12c). Figure 13 shows the range estimation of the two methods when the sound source distance was 13.71 km. The range estimations of the CMFP and LFDMFP were 12.98 and 13.60 km, respectively, with range estimation errors of 5.32 and 0.8%, respectively. The depth estimation was approximated to 24 m (Figure 13c). Unlike the processing results of the other stations, in this station, the optimal range estimation of the CMFP was not within the periodic sidelobes, which indicates that the internal wave activity significantly affected the amplitude and phase of the sound field, thereby resulting in significant differences between the estimated and actual distances of the CMFP. However, we found that the LFDMFP demonstrated strong robustness in the face of mismatch (Figure 13b). The preliminary analysis suggested that this may have stemmed from the mechanism of the LFDMFP when handling the frequency difference, which converted the phase part to $k_n(\omega + \Delta\omega) - k_n(\omega)$. Compared with the phase part $k_n(\omega)$ of the CMFP, this change significantly improved the adaptability to the horizontal wavenumber changes. In addition, the LFDMFP also alleviated the influence of the internal waves on the eigenfunctions to a certain extent during its depth integration process. This also had a positive effect on improving the robustness of the LFDMFP in the face of a mismatch. Through these mechanisms, the LFDMFP more stably handled the complex signal changes and thereby demonstrated more reliable performance in practical applications.

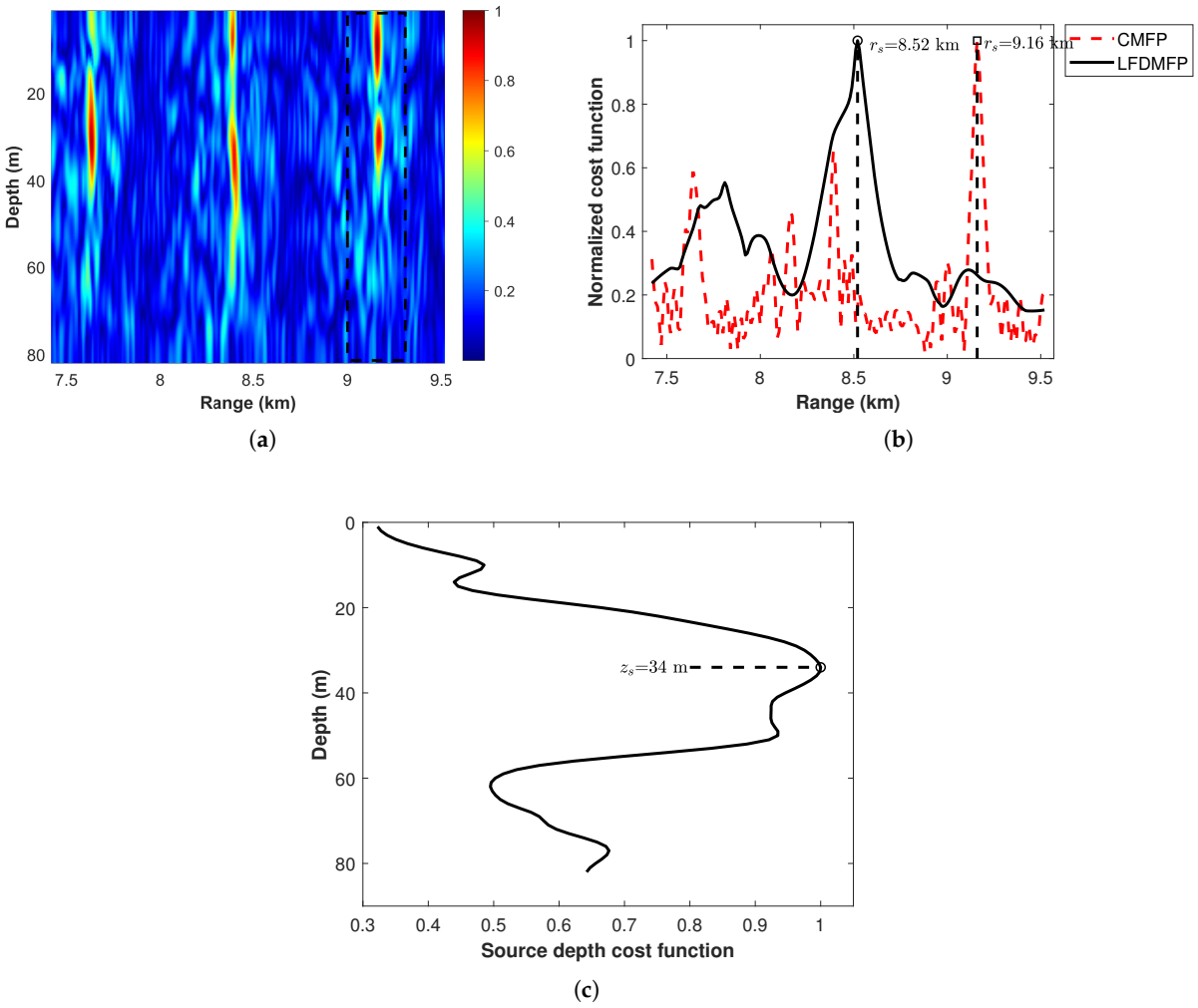


Figure 12. Localization results for the LFDMFP and CMFP: (**a**) the incoherent averaging ambiguity function of the CMFP; (**b**) the cost function of the LFDMFP at the frequency difference [297 Hz, 301 Hz]; (**c**) the depth of the source estimated using Equation (15).

Figure 14 depicts the range estimation results of the CMFP and LFDMFP over time against the actual source distance and entire range estimation error. The size of the error bar represents the relative error, which was calculated using Equation (21). Within the experimental trajectory range, the average relative error of the LFDMFP was 1.9%, which demonstrated that the LFDMFP had greater potential compared with the CMFP regarding the range estimation. Furthermore, the data from a single station position illustrated that the LFDMFP proposed in this paper exhibited robustness against environmental mismatches caused by internal waves.

$$error = |\frac{\hat{r}_s - r_0}{r_0}| \times 100\% \tag{21}$$

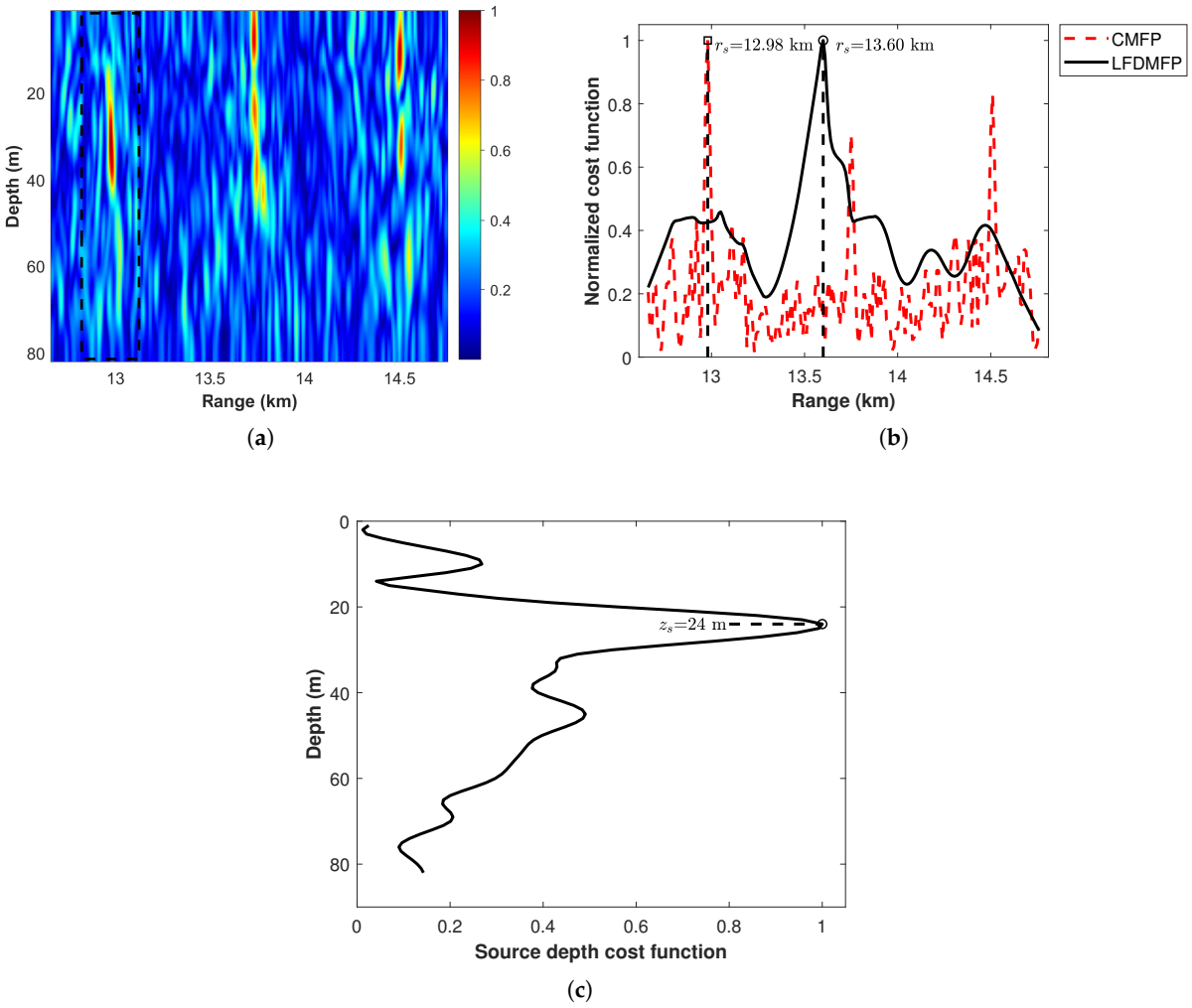


Figure 13. Localization results for the LFDMFP and CMFP: (**a**) the incoherent averaging ambiguity function of the CMFP; (**b**) the cost function of the LFDMFP at the frequency difference [300 Hz, 306 Hz]; (**c**) the depth of the source estimated using Equation (15).

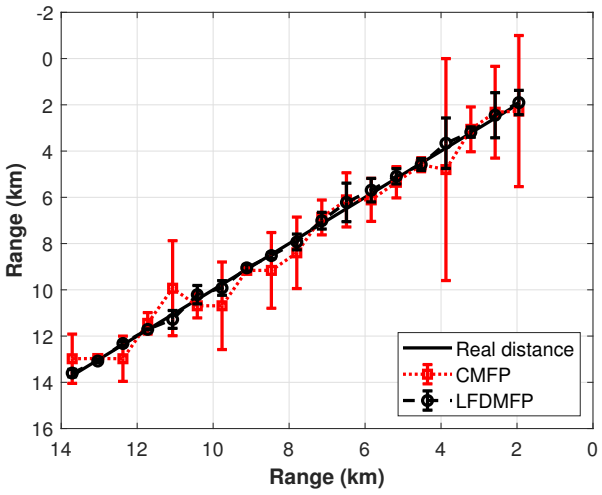


Figure 14. Comparison of range estimations and the real range for the CMFP and LFDMFP at different positions.

5. Conclusions

A novel linear array-signal-processing technique based on the normal mode theory and CMFP is presented to estimate the range of a source with only two neighbor frequency components of the radiated sound field. Its performance was investigated for source localization in simulations and experimental data. Four conclusions were drawn from this research effort: (1) The LFDMFP was successful at estimating the range in shallow water. Using measurements from the Qingdao Laoshan Bay Sound Propagation Experiment and Sound Propagation Experiment in the South China Sea, the technique's source range and depth estimations were unambiguous. The average absolute errors of the range estimation in the two experiments were 2.2 and 1.9%, respectively. (2) Due to the larger modal interference, only one peak appeared within the range search interval. Therefore, the LFDMFP could efficiently control the defect that CMFP provides multiple optimal estimations of the source range estimation. (3) The degraded resolution, which could be improved by combining the CMFP in the case where the knowledge of the actual propagation environment is known perfectly, may allow us to mesh grids with a larger range step size to reduce the computational cost of the replicas. (4) Preliminary analysis indicated that the LFDMFP had a higher tolerance for the environmental mismatch of fluctuations, which was verified by the data-processing results of the Sound Propagation Experiment in the South China Sea.

The above conclusions are positive since the simulation and experimental results validated the effectiveness of the LFDMFP. However, users need to make certain adjustments to the method based on the experimental environment and signal parameters, such as the frequency difference. Furthermore, it is hard to give a quantitative evaluation of how the frequency difference affects the range estimation. Further analysis of other factors that affect the performance of LFDMFP and a mathematical explanation of why LFDMFP is robust to environmental mismatch will be elaborated in future publications.

Author Contributions: Conceptualization, P.S. and H.W.; methodology, P.S.; software, P.S.; validation, P.S., H.W. and W.G.; formal analysis, H.W.; investigation, P.S.; resources, P.S. and H.W.; data curation, P.S.; writing—original draft preparation, P.S.; writing—review and editing, P.S. and H.W.; visualization, H.W.; supervision, W.G. and L.W.; project administration, B.S., W.G. and L.W.; funding acquisition, H.W. All authors have read and agreed to the published version of this manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant no. 11674294).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors thank Shoudong Wang and Hongxu Cui for the insightful discussions and instructive guidance.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bucker, H.P. Use of calculated sound fields and matched-field detection to locate sound sources in shallow water. *J. Acoust. Soc. Am.* **1976**, *59*, 368–373. [CrossRef]
2. Baggeroer, A.; Kuperman, W.; Mikhalevsky, P. An overview of matched field methods in ocean acoustics. *IEEE J. Ocean. Eng.* **1993**, *18*, 401–424. [CrossRef]
3. Williamson, L.; Haapaniemi, E.; Dowling, D. Sonar Performance Improvements Using First Reflections in Matched-field Processing. *Nav. Eng. J.* **2014**, *126*, 125–129.
4. Johnson, D. The application of spectral estimation methods to bearing estimation problems. *Proc. IEEE* **1982**, *70*, 1018–1028. [CrossRef]
5. Baggeroer, A.B.; Kuperman, W.A.; Schmidt, H. Matched field processing: Source localization in correlated noise as an optimum parameter estimation problem. *J. Acoust. Soc. Am.* **1988**, *83*, 571–587. [CrossRef]
6. Schmidt, H.; Baggeroer, A.B.; Kuperman, W.A.; Scheer, E.K. Environmentally tolerant beamforming for high-resolution matched field processing: Deterministic mismatch. *J. Acoust. Soc. Am.* **1990**, *88*, 1851–1862. [CrossRef]
7. Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280. [CrossRef]

8. Westwood, E.K. Broadband matched-field source localization. *J. Acoust. Soc. Am.* **1992**, *91*, 2777–2789. [CrossRef]
9. Shang, E.C. Source depth estimation in waveguides. *J. Acoust. Soc. Am.* **1985**, *77*, 1413–1418. [CrossRef]
10. Yang, T.C. A method of range and depth estimation by modal decomposition. *J. Acoust. Soc. Am.* **1987**, *82*, 1736–1745. [CrossRef]
11. Yang, T.C.; Yates, T. Matched-beam processing: A synthesis of conventional beamforming and matched-field processing. *J. Acoust. Soc. Am.* **1996**, *100*, 2852. [CrossRef]
12. Yang, T.C.; Yates, T. Matched-beam processing: Application to a horizontal line array in shallow water. *J. Acoust. Soc. Am.* **1998**, *104*, 1316–1330. [CrossRef]
13. Yan, H.; Xu, J.; Long, T.; Zhang, X. Underwater Acoustic Matched Field Imaging Based on Compressed Sensing. *Sensors* **2015**, *15*, 25577–25591. [CrossRef] [PubMed]
14. Wang, H.; Wang, N.; Gao, D.; Gao, B. Source-Space Compressive Matched Field Processing for Source Localization. *Chin. Phys. Lett.* **2016**, *33*, 044301. [CrossRef]
15. Gemba, K.L.; Hodgkiss, W.S.; Gerstoft, P. Adaptive and compressive matched field processing. *J. Acoust. Soc. Am.* **2017**, *141*, 92–103. [CrossRef] [PubMed]
16. Worthmann, B.M.; Song, H.C.; Dowling, D.R. High frequency source localization in a shallow ocean sound channel using frequency difference matched field processinga). *J. Acoust. Soc. Am.* **2015**, *138*, 3549–3562. [CrossRef]
17. Worthmann, B.M.; Song, H.C.; Dowling, D.R. Adaptive frequency-difference matched field processing for high frequency source localization in a noisy shallow ocean. *J. Acoust. Soc. Am.* **2017**, *141*, 543–556. [CrossRef]
18. Geroski, D.J.; Dowling, D.R. Robust long-range source localization in the deep ocean using phase-only matched autoproduct processinga). *J. Acoust. Soc. Am.* **2021**, *150*, 171–182. [CrossRef]
19. Park, M.; Choo, Y.; Choi, J.; Lee, K. Reformulation of frequency-difference matched-field processor for high-frequency known-source localization. *J. Acoust. Soc. Am.* **2023**, *154*, 948–967. [CrossRef]
20. Liu, M.; Niu, H.; Li, Z. Implementation of Bartlett matched-field processing using interpretable complex convolutional neural network. *JASA Express Lett.* **2023**, *3*, 026003. [CrossRef]
21. Byun, G.; Akins, F.; Gemba, K.; Song, H.; Kuperman, W. Multiple constraint matched field processing tolerant to array tilt mismatch. *J. Acoust. Soc. Am.* **2020**, *147*, 1231–1238. [CrossRef] [PubMed]
22. Frichter, G.; Byrne, C.; Feuillade, C. Sector-Focused Stability Methods for Robust Source Localization in Matched-Field Processing. *J. Acoust. Soc. Am.* **1990**, *88*, 10. [CrossRef]
23. Abadi, S.H.; Song, H.C.; Dowling, D.R. Broadband sparse-array blind deconvolution using frequency-difference beamforming. *J. Acoust. Soc. Am.* **2012**, *132*, 3018–3029. [CrossRef] [PubMed]
24. Jensen, F.B.; Kuperman, W.A.; Porter, M.B.; Schmidt, H.; Tolstoy, A. *Computational Ocean Acoustics*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 2011.
25. Tantum, S.L.; Nolte, L.W. On array design for matched-field processing. *J. Acoust. Soc. Am.* **2000**, *107*, 2101–2111. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Review

Artificial Intelligence-Based Underwater Acoustic Target Recognition: A Survey

Sheng Feng [1], Shuqing Ma [2], Xiaoqian Zhu [2,*] and Ming Yan [2]

[1] College of Computer Science, National University of Defense Technology, Changsha 410073, China; fengsh14@lzu.edu.cn
[2] College of Meteorology and Oceanography, National University of Defense Technology, Changsha 410073, China; mashuqing@nudt.edu.cn (S.M.)
* Correspondence: zhu_xiaoqian@sina.com

Abstract: Underwater acoustic target recognition has always played a pivotal role in ocean remote sensing. By analyzing and processing ship-radiated signals, it is possible to determine the type and nature of a target. Historically, traditional signal processing techniques have been employed for target recognition in underwater environments, which often exhibit limitations in accuracy and efficiency. In response to these limitations, the integration of artificial intelligence (AI) methods, particularly those leveraging machine learning and deep learning, has attracted increasing attention in recent years. Compared to traditional methods, these intelligent recognition techniques can autonomously, efficiently, and accurately identify underwater targets. This paper comprehensively reviews the contributions of intelligent techniques in underwater acoustic target recognition and outlines potential future directions, offering a forward-looking perspective on how ongoing advancements in AI can further revolutionize underwater acoustic target recognition in ocean remote sensing.

Keywords: literature review; machine learning; deep learning; ocean remote sensing; underwater target recognition

Citation: Feng, S.; Ma, S.; Zhu, X.; Yan, M. Artificial Intelligence-Based Underwater Acoustic Target Recognition: A Survey. *Remote Sens.* **2024**, *16*, 3333. https://doi.org/10.3390/rs16173333

Academic Editors: Haiyong Xu, Jingyue Lu, Lei Zhang, Marc Pinto, Farman Ali and Xuebo Zhang

Received: 7 August 2024
Revised: 5 September 2024
Accepted: 6 September 2024
Published: 8 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Ships navigating in the ocean inevitably emit sound waves. These emitted underwater signals typically contain discriminative characteristics that are critical for classifying different ships, such as the ship structure, propeller configuration, engine power, and ship condition. Common UATR tasks include vessel type classification [1–3], sonar imagery recognition [4–8], and recognition of communication signal modulation [9,10]. As an important branch of ocean remote sensing, underwater acoustic target recognition (UATR) technology has a wide range of marine applications, including marine resource development, ocean rescue, biological protection, and military applications. Its main purpose is to process and analyze sonar-received signals in order to determine the target type based on discriminative characteristics. From the perspective of pattern recognition, the task of UATR fundamentally addresses the challenge of enhancing separability among different types of ships, that is, how to pull homologous samples closer and push heterologous samples apart to construct a compact set. As shown in Figure 1, it is always possible to find separable boundaries between different ship classes for compact sets, whereas for noncompact sets, it is difficult to find an obvious separable boundary.

In fact, the inherent complexity and uncertainty of the underwater environment, including the sound sources, sea surface, water column, and seabed sediments, dictates the unique characteristics of underwater sound propagation, which poses numerous challenges for UATR. First of all, the speed of underwater sound is influenced by factors such as water temperature, density, and depth, leading to refraction where sound paths bend towards regions of lower sound speed. As illustrated in Figure 2, underwater sound signals are

characterized by multipath propagation, along with effects such as attenuation, seabed penetration, scattering, and reverberation [11].

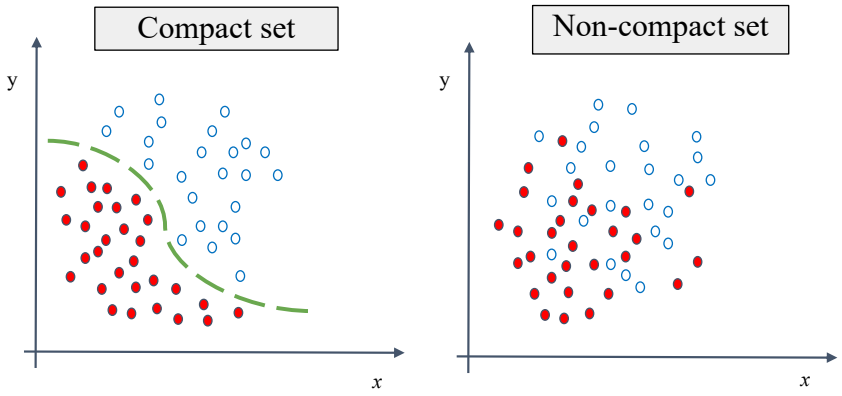


Figure 1. Schematic illustration of the separability between underwater signals from the perspective of pattern recognition.

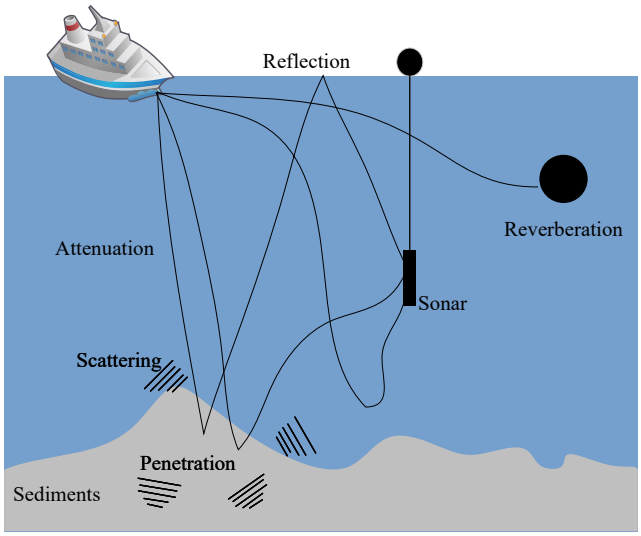


Figure 2. Schematic diagram of a typical underwater signal propagation channel.

Secondly, the unique acoustic characteristics of underwater targets are highly diverse, encompassing a variety of ship types and marine animals, each with distinctive acoustic signatures and complex sound generation mechanisms that pose challenges to recognition models. Coupled with the variability of these acoustic features against the ocean background noise, the task of UATR becomes exceptionally challenging.

In the early days, UATR methods primarily relied on experienced sonar operators to make manual determinations, including auditory and visual recognition. These operators were trained to recognize enemy submarines in ambient noise and even to distinguish marine animals from potential threats. This manual approach, combining auditory perception with visual interpretation of sonar displays, defined the early methods of UATR. However, the reliance on human interpretation not only limited the recognition speed but also introduced a significant error due to various psychological and physiological factors, which may greatly reduce the recognition accuracy. On the other hand, with the advancement of ship-radiated noise reduction, the increasing trend towards low-frequency characteristics further degrades the effectiveness of traditional methods. To address this issue, artificial intelligence (AI)-based UATR methods are rapidly developing [12], which attempt to extract meaningful knowledge from extensive training samples to find separable

boundaries between different targets and subsequently recognize new samples based on the acquired knowledge. Generally, these intelligent recognition techniques have been implemented in sonar devices, demonstrating autonomous, efficient, and accurate performance in comparison to conventional methods [3].

Figure 3 illustrates the workflow of a typical intelligent UATR system, including data collection, data preprocessing, feature extraction, and classification algorithms. Among these processes, data collection is conducted by sonar operators using hydrophone arrays to acquire acoustic data. Subsequently, data preprocessing is implemented to mitigate clutter interference and to eliminate the effects of Doppler shift [13–15]. Typical preprocessing techniques include resampling, filtering, and denoising, which contribute to obtain high-quality signal data. Importantly, the process of feature extraction and the design of classification algorithms are considered to be crucial aspects that affect the recognition performance.

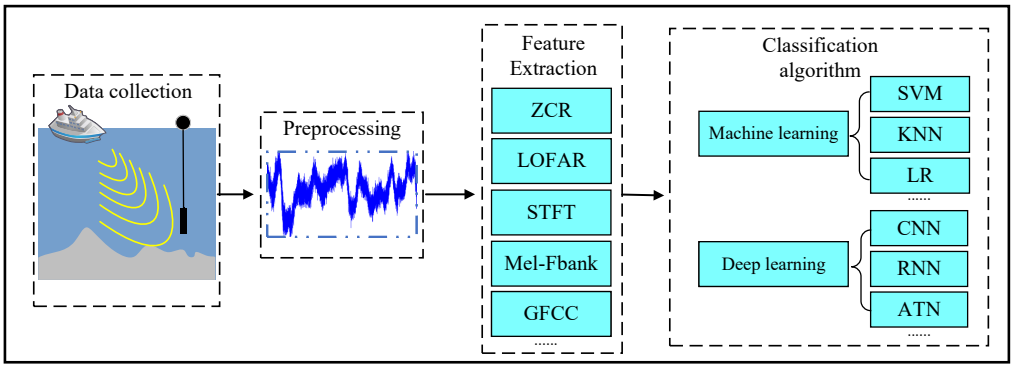


Figure 3. The workflow of a typical intelligent UATR system.

Feature extraction is typically associated with underwater acoustic signal processing techniques to extract the intrinsic physical information of the target. It is particularly crucial in addressing the second major challenge of target diversity. By employing a robust feature extraction method, it is possible to obtain discriminative features capable of distinguishing between various targets, thereby mitigating the complexity introduced by the wide range of underwater acoustic sources. In these methods, time-domain characteristics can offer straightforward acoustic representations, such as zero-crossing rate (ZCR), peak-to-peak amplitude, and waveform energy. Nonetheless, these features may not capture the time-varying characteristics that are crucial for achieving accurate UATR. To address this limitation, joint time–frequency techniques have been introduced into UATR. These techniques aim to extract discriminative features containing both temporal and spectral information, such as the short-time Fourier transform (STFT), low-frequency analysis and recording (LOFAR), Mel-filter bank (Mel-Fbank), and Gammatone Frequency Cepstral Coefficients (GFCCs).

For classification algorithms, machine learning (ML) methods were firstly introduced into the field of intelligent UATR, such as support vector machine (SVM), *K* nearest neighbor algorithm (*K*NN), hidden Markov model (HMM), and logistic regression (LR). Due to the relatively simple structure of ML models, these methods primarily utilize shallow signal features for intelligent recognition of simple targets, such as whale calls [16] and medium echoes [17]. As a subset of ML, deep learning (DL) methods are capable of learning deep representations from complex acoustic signals using deep neural network (DNNs). Composed of stacked layers, DNNs are particularly effective in automatically extracting hierarchical features from acoustic data. Recent research has extensively investigated DL-based UATR methods, demonstrating their superiority over shallow ML classifiers by leveraging high-level information inherent in acoustic signals [18,19]. Initially, early DL-based intelligent UATR systems were constrained by computational resources, particularly limited GPU memory, which made it difficult to perform fast and stable DNN training.

However, in the early 21st century, significant advancements in electronic technology were made, notably enhanced computational capabilities of GPUs. Consequently, the application of advanced DL techniques has made significant progress in intelligent UATR, thereby making substantial contributions to the field of ocean remote sensing.

In this paper, we provide a comprehensive review of two pivotal components within AI-based UATR systems: commonly utilized feature extraction methods and classification algorithms. These components represent the mainstream of ocean remote sensing due to their significant potential and impact. Furthermore, we highlight major challenges in the current field of intelligent UATR and propose potential solutions. The structure of this paper is as follows: Section 2 investigated widely utilized feature extraction methods to enhance separability in UATR. Section 3 discusses the shallow ML algorithms contributions to UATR. In Section 4, a wide range of DL applications in UATR are presented. Finally, Section 5 outlines prospective research directions aimed at proposing potential solutions to current challenges in AI-based UATR.

2. Feature Extraction Methods

As is well-known, each ship target exhibits a unique underwater acoustic pattern, with recognition performance largely dependent on feature selection. In this case, feature extraction techniques are pivotal for enhancing recognition performance. Ideally, the feature representation should be highly discriminative to provide sufficient information for classification algorithms. In such cases, it can effectively reduce intra-class variance, and increase inter-class variance, to find the compact set of acoustic signals. In particular, the waveform envelope of underwater acoustic signals, known as a kind of natural time series data, carries the inherent information of acoustic targets, providing direct insights into their temporal characteristics. Time-domain feature extraction methods from signal waveforms are closely related to the physical properties and dynamic behavior of underwater acoustic targets [20]. Deng et al. [21] proposed an integrated model-based feature extraction method that combines broadband–phase correlation and spatial distribution features within an embedded recognition system; real-time experiments demonstrated its effectiveness. In addition, Meng et al. [22] introduced a feature extraction method based on time-domain waveform structure, which achieved more than 89.5% recognition accuracy based on statistical zero wavelength and peak amplitude features using ML algorithms. Sun and Zhang [23] utilized time series chaotic dynamics to reconstruct the phase space of the ship-radiated noise. They utilized relevant dimensions and the maximum Lyapunov exponent features, thereby achieving successful UATR. While single-domain statistical feature extraction methods have shown promising results on small-scale datasets, they are also subject to limitations. Specifically, for underwater acoustic signals, time-domain analysis effectively captures time-varying characteristics but lacks frequency information. Conversely, frequency-domain analysis reveals the signal's frequency components but disregards temporal information. In AI-based UATR, researchers have increasingly focused on discriminative spectral features, which can be broadly categorized into three types: the first category is the interpretable spectral features that hold physical significance, which investigate the physical structure and motion state of the target ship. The second category deals with the problem of insufficient single-domain features by using time–frequency joint features, where time–frequency analysis techniques are used to integrate the temporal and spectral domain information of signals. The final category concerns autocoding features, which involve end-to-end learning of feature representations utilizing DNNs or other methods without manual feature engineering.

2.1. Physical Significance Feature

In marine environments, ship-radiated noise mainly includes three types: mechanical noise, hydrodynamic noise, and propeller noise. Mechanical noise is produced by the vibrations of the internal ship equipment, such as ventilation and pumps, which typically exhibit unique noise patterns with prominent line spectra. Hydrodynamic noise refers to

noise generated by the interaction between fluid flow and ship hull, which is relatively smooth and characterized by a continuous spectrum. Propeller noise constitutes the main components of the ship-radiated noise, which has a wide frequency range from 5 Hz to 100 KHz. This type of noise is notable for its significant low-frequency line spectra and continuous spectrum characteristics. Let $s(t)$ represent ship noise, given by

$$\{s(t)\} = \{x(t)\} + \sum_{i=1}^{n} \{l_i(t)\}, \tag{1}$$

where $\{x(t)\}$ denotes wide-sense stationary random processes, $l_i(t)$ denotes initially random periodic signals, and $i = 1, 2, \ldots, n$, representing n periodic signals. Then we have the power spectrum:

$$S(f) = \lim_{T \to \infty} \frac{1}{T} E\left(\left|S_{k,T}(f)\right|^2\right), \tag{2}$$

where T denotes the length of the signal segment during Fourier transformation; E denotes ensemble averaging; and k represents the segment number of the signal.

In summary, the aforementioned physical components can reveal the physical structure and motion characteristics of the ship with robustness, providing useful information of significant physical relevance. To capture this meaningful information, the LOFAR [18,24] and the DEMON spectra [25] are commonly used in UATR studies. These two types of spectrograms are illustrated in Figure 4.

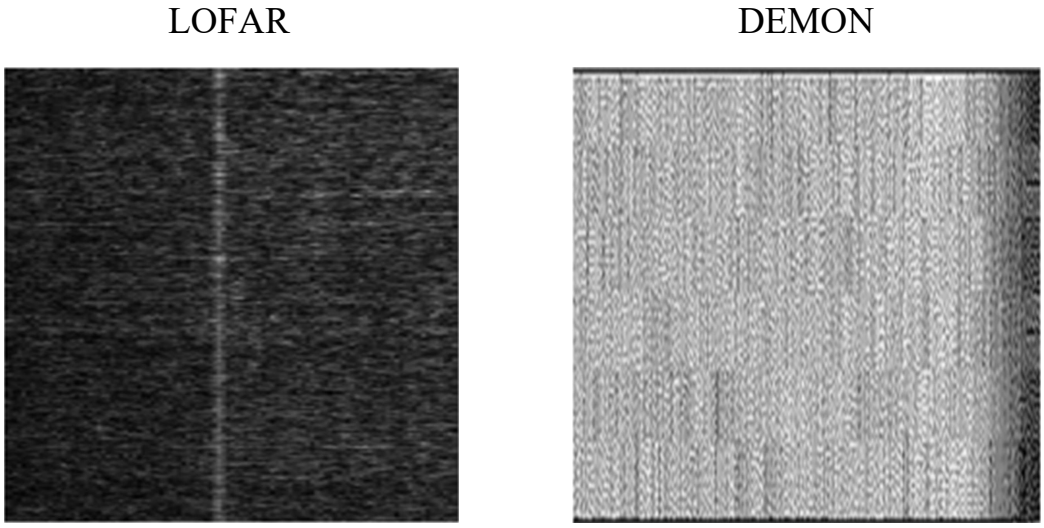


Figure 4. Physical significance feature of an underwater signal sample, including its LOFAR and DEMON [26] spectrograms.

2.1.1. LOFAR Physical Spectrum

Known for its capability to depict low-frequency line spectra, the LOFAR spectrum is particularly useful for recognizing underwater acoustic targets [26,27]. The main process to extract the LOFAR spectrum can be summarized as follows:

(1) The signal sequence is firstly framed to obtain L overlapped samples. Common framing functions include the Hanning or Hamming windows.

(2) Each framed signal undergoes normalization and centering:

$$u_j(n) = \frac{M_j(n)}{\max\left(M_j(n)\right)}, \quad 1 \leqslant j \leqslant L, \tag{3}$$

$$x_j(n) = u_j(n) - \frac{1}{L} \sum_{i=1}^{L} u_j(i). \tag{4}$$

(3) The short-time Fourier transform (STFT) is applied to signal $x_j(n)$ to obtain the LOFAR spectrum:

$$X_j(k) = FFT[x_j(n)]. \tag{5}$$

These spectra are arranged over time in a coordinate system to generate the complete LOFAR spectrogram. Due to the predominant low-frequency range of ship-radiated noise, the LOFAR spectrogram plays a critical role in analyzing the characteristics of underwater targets. The bright spectral lines in the spectrogram often correspond to the most important features of underwater acoustic targets. Jin et al. [26] utilized an adversarial generative network (GAN) to generate simulated samples based on their LOFAR spectrogram for data augmentation, which has proven its effectiveness to solve the few-shot UATR problem. In addition, Shi et al. [28] analyzed the line spectral traces on the LOFAR spectrogram and calculated the relative variance between adjacent time frames to enhance the detection sensitivity of low-intensity moving targets.

2.1.2. DEMON Physical Spectrum

In UATR, the DEMON spectrum also plays a crucial role due to its ability to extract detailed physical information from acoustic signals, such as propeller blade count and shaft frequency. This technique is specifically designed to detect modulations originating from propeller cavitation envelopes [29]. In practice, the DEMON spectrum entails the selection of the spectral band exhibiting the most pronounced time-varying modulation for envelope detection, which is subsequently followed by detailed spectral analysis to derive relevant parameters.

Specifically, the amplitude of broadband noise $n(t)$ generated by propeller cavitation is modulated by a set of sinusoidal signals. This broadband signal can be represented as

$$x(n) = \left(1 + \sum_{n=1}^{N} A_n \sin(2\pi n f_0 + \phi_n)\right) s(n) + g(n), \tag{6}$$

where A is the modulation amplitude, nf_0 is the modulation frequency, $s(n)$ denotes the broadband noise signal, and $g(t)$ represents background noise. After bandpass filtering using the coefficient $h_1(n)$, $y_1(n)$ is computed as

$$y_1(n) = \sum_{k=0}^{M-1} x(n-k) h_1(k), \tag{7}$$

Then, the Hilbert transform H is further utilized for envelope demodulation:

$$y_2(n) = |H(y_1(n))|. \tag{8}$$

After removing the DC component to obtain $y_3(n)$, subsequent downsampling and low-pass filtering are performed to yield $y_4(n)$:

$$y_4(n) = \sum_{k=0}^{P-1} y_3(dn-k) h_2(k), \tag{9}$$

where d denotes the downsampling factor and $h_2(k)$ represents the low-pass filter coefficients. Further spectral analysis using $y_4(n)$ provides power spectral density for extracting line spectra. These procedures facilitate the extraction of crucial physical parameters related to underwater targets. However, the complexity introduced by interference line spectra within DEMON spectrograms can degrade recognition performance under low signal-to-noise ratios (SNRs). To address this problem, Tong et al. [30] proposed the adaptive weighted envelope spectrum (AWES) method for better demodulation capability, which dynamically assigns weights to spectral components based on their modulation intensities. Furthermore, Li et al. emphasized that using LOFAR or DEMON alone is insufficient for

estimating arbitrary rotation speeds in target detection tasks [31]. To address this issue, they combined the LOFAR and DEMON spectrograms to enhance ship estimation under various conditions. Figure 5 illustrates an example of a fused LOFAR and DEMON spectrogram after filtering. Experimental results demonstrate that feeding the comb-filtered combined feature (LOFAR + DEMON) into the DL model achieves the highest accuracy of 98%, outperforming the individual filtered LOFAR (83%) or DEMON (94%).

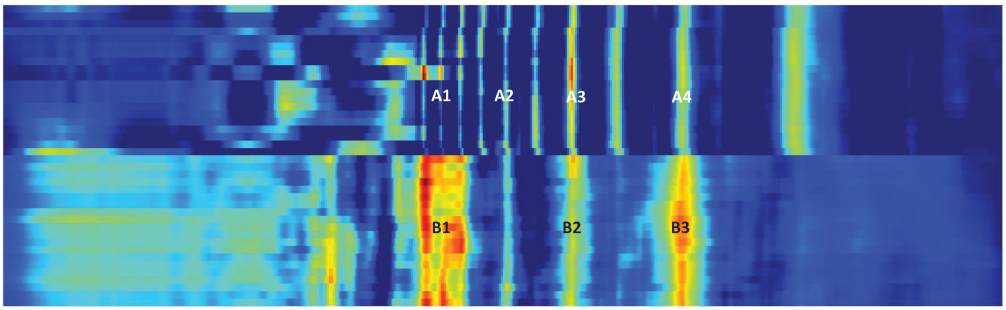


Figure 5. An example of a fused LOFAR and DEMON spectrogram with comb filtering [31]. A1–A4 and B1–B3 represent the primary spectral lines.

2.2. Joint Time–Frequency Feature

Numerous studies have shown that underwater acoustic signals exhibit nonstationary characteristics [32–34], with their frequencies varying significantly over time. Traditional Fourier transform methods only provide average information about the signal without revealing the temporal evolution of signal frequencies. To address this problem, researchers have attempted to represent the signal within a joint time–frequency domain, and have proposed a variety of time–frequency analysis techniques [35]. The objective of these techniques is to construct a joint time–frequency distribution function $F_x(t, f)$ that can well characterize the time-varying properties of the signal $x(t)$, such as the energy variation with time t and frequency f. This robust feature, used in the processing process, can provide dynamic time-varying characteristics for subsequent classification algorithms [36,37]. The commonly employed time–frequency features in UATR mainly include general time–frequency feature extraction, auditory perception features, and multidimensional features.

2.2.1. General Time–Frequency Feature

In acoustic processing, STFT is a fundamental method for analyzing underwater acoustic signals, providing a basis for the development of various time–frequency analysis methods, which assumes signal stationarity within a narrow window. Specifically, for a continuous nonstationary signal $x(t)$, given a narrow window function w sliding along the time axis, the STFT of $x(t)$ can be defined as

$$\mathrm{STFT}_x(t, \omega) = \int_{-\infty}^{\infty} x(\tau) w^*(\tau - t) e^{-j\omega\tau} d\tau, \tag{10}$$

where ω represents the angular frequency and $*$ denotes the complex conjugate. To further enrich information representation, Zhang et al. [38] proposed a multiscale STFT feature extraction method by expanding the spectrogram channel. Experimental results have illustrated its effectiveness in low-frequency information extraction, thereby improving recognition performance on real-world datasets. However, the fixed window size in STFT limits adaptive adjustment. If the window function w is too narrow, high time resolution captures snapshot characteristics, but frequency resolution is reduced due to short window segments. Conversely, if w is too wide, time domain precision is reduced, resulting in lower time resolution that fails to capture rapid signal changes. To address this issue, high-resolution time–frequency feature extraction methods have been applied to UATR [39,40]. In the early 1980s, wavelet transform was introduced as a mathematical tool for analyzing geophysical signals [41], which adapts its window size with frequency changes, making it

suitable for analyzing nonstationary signals. In continuous wavelet transform (CWT), a time dilation operator is employed, defined as the time scale. CWT maps one-dimensional square integrable functions (energy-limited signals) into a two-dimensional function of time scale a and time shift b. For a square integrable function $y(t)$, its CWT with respect to the mother wavelet $\psi(t)$ can be defined as

$$W_\psi y(a,b) = \langle y, \psi_{a,b} \rangle = \int_{-\infty}^{+\infty} y(t) \frac{1}{\sqrt{|a|}} \psi^* \cdot \left(\frac{t-b}{a} \right) \mathrm{d}t, \tag{11}$$

and the mother wavelet function $\psi(t)$ is derived from $\psi_{a,b}$ through affine transformations:

$$\psi_{a,b}(t) = [U(a,b)\psi(t)] = \frac{1}{\sqrt{|a|}} \psi \left(\frac{t-b}{a} \right). \tag{12}$$

Here, $U(a,b)$ denotes affine transformations of the $\psi(t)$, while the unitary transformation ensures energy preservation with a normalization factor of $\frac{1}{\sqrt{|a|}}$. Clearly, when $a > 1$, the wavelet function expands, thereby reducing its frequency, enabling the measurement of lower-frequency components; when $a < 1$, the wavelet function compresses, increasing its frequency and facilitating the measurement of higher-frequency components using wavelet transform. Li et al. [42] proposed a wavelet packet transform-based algorithm for extracting features from underwater radiated noise. This method decomposed signals into four wavelet packets and used energy features from each frequency band as inputs to a Radial Basis Function (RBF) network classifier. Experimental results highlighted that wavelet packet transform can enhance signal fractal characteristics during feature reduction, thus effectively accomplishing diverse recognition tasks. Rademan et al. [43] applied continuous wavelet transform to extract discriminative features from whale calls, achieving superior detection accuracy and specificity compared to STFT methods on simulated datasets. Furthermore, other high-frequency analysis methods such as Hilbert–Huang transform (HHT) [44] and Wigner–Ville distribution (WVD) [45,46] have also been employed for feature extraction. These time–frequency methods hold potential for providing discriminative features, selected based on task requirements, resource constraints, and specific advantages in various recognition scenarios. STFT, valued for its simplicity and real-time capabilities, is commonly used in real-time sonar monitoring and rapid target detection. In contrast, the wavelet transform has multiscale analysis capabilities and performs well in noisy environments with reverberation effects or with target features at different scales, effectively revealing temporal variations. For scenarios involving rapid moving target resulting in frequency modulation or Doppler effects, HHT can provide detailed analysis of target motion dynamics. Moreover, WVD has demonstrated effectiveness in applications demanding high time–frequency resolution, revealing intricate details that are difficult to achieve with other methods by properly handling cross-terms.

2.2.2. Auditory Perceptual Features

In modern ocean engineering, considerable research focuses on auditory perceptual features that closely approximate human auditory perception [47–50]. Among these features, Mel frequency-based auditory perceptual features are mostly utilized in UATR [51]. This method utilizes f_{mel} for triangular filtering and compression to derive Mel-Fbank coefficients from radiated noise. Figure 6 illustrates the response function of equi-height Mel-Fbank, composed of triangular filters symmetrically responding around their central frequencies. Specifically, filters exhibit maximal response at their center frequencies and gradually attenuate symmetrically on either side.

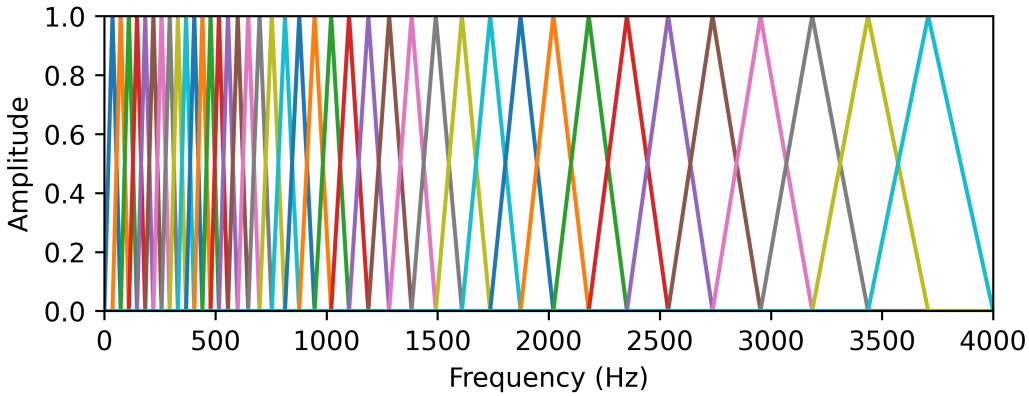


Figure 6. The Mel triangular filters implemented by Librosa package.

These frequency responses of Mel-Fbank can be expressed as

$$H(f) = \begin{cases} \frac{f_h - f}{f_h - f_0}, & f_0 < f \leqslant f_h \\ \frac{f_l - f}{f_l - f_0}, & f_l \leqslant f \leqslant f_0 \end{cases}. \tag{13}$$

Furthermore, the Mel-Fbank can be represented as

$$E(m) = \sum_{f=f_l(m)}^{f_h(m)} H_m(f)|X_n(f)|, \quad l = 1, 2, \cdots, L. \tag{14}$$

The resulting Mel-spectrogram, a two-dimensional representation of Mel-Fbank features, depicts f_{mel} frequency intensities across time frames. Due to overlapping frequency components, Mel-spectrograms are known to preserve intrinsic signal details, exhibit high feature inter-correlation, and demonstrate excellent capabilities for nonlinear feature extraction. Taking Mel-spectrograms as natural images, Liu et al. [49] integrated these auditory perceptual features of underwater targets into DNNs to learn frequency characteristics per hop group, highlighting their benefits in feature extraction under varying underwater acoustic communication scenarios. On the contrary, Tang et al. [50] argued that Mel-spectrograms differ significantly from natural images in that they contain rich time–frequency information. To efficiently handle this important information, they developed a three-dimensional spectrogram network, achieving an optimal balance between recognition performance and model complexity. Further application of Discrete Cosine Transform (DCT) on Mel-spectrograms yields Mel-Frequency Cepstral Coefficients (MFCCs) [52]. Utilizing these MFCC features extracted from ship-radiated noise, further classification of acoustic targets using ML [53] and DL techniques [54,55] has demonstrated substantial improvements in recognition performance.

The feature of Gammatone Frequency Cepstral Coefficients (GFCCs) is another widely used auditory perceptual representation in UATR [19,56,57]. Similar to MFCC features, the GFCC is obtained through DCT from a Gammatone filter bank, possessing excellent capabilities in representing signal spectral structures. The main difference between GFCCs and MFCCs lies in their design principles: GFCC models filter responses based on the cochlear frequency decomposition, whereas the MFCC relies on the f_{mel} scale. Due to their robust representation of nonlinearities and spectral envelopes, GFCC features have proven effective in describing sound signals, particularly in scenarios with low SNRs [58].

2.2.3. Multidimensional Fusion Features

To enrich feature diversity, fusion-based feature extraction methods have recently gained increasing attention in the field of UATR [59–61]. Following the concept of feature fusion, these approaches combine different time–frequency features to enrich signal rep-

resentations, thereby providing discriminative information for classification algorithms. As shown in Figure 7, these methods typically involve two primary strategies. The first strategy adopts a single-network approach with frontend feature fusion during preprocessing. The frontend fusion integrates various features into multiple input channels, similar to the RGB channels in natural images, thereby constructing a multidimensional fused feature designed for UATR. Conversely, the second strategy employs ensemble models with backend feature fusion, leveraging multiple networks to process data and classify target labels using probability vectors across multiple model backends.

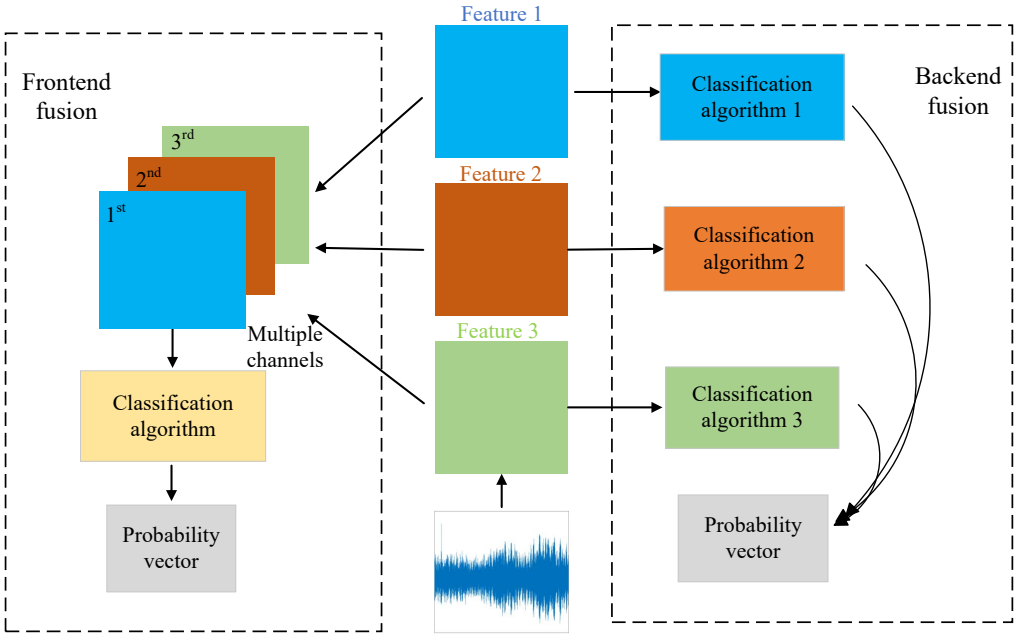


Figure 7. Two aspects of the multidimensional features.

In the frontend fusion approach, Tan et al. [62] extracted various features from underwater signals, including Amplitude Modulation Spectrogram, MFCC, Relative Spectral Transform–Perceptual Linear Prediction, GFCC, and Delta feature. These features are then fused to form AMCG-Delta features. Integration of various spectral data augmentation techniques resulted in a recognition accuracy 6.95% higher than that of the baseline method. Liu et al. [63] combined Mel-spectrograms with their first and second derivatives to form 3D Mel-spectrograms that were fed into a DNN architecture, achieving satisfactory recognition performance in multiple UATR tasks. Similarly, Wu et al. [64] also employed 3D Mel-spectrograms as input features, further combining transfer learning from ImageNet to enhance the feature applicability in underwater acoustic modalities. Their experimental findings on publicly available datasets showcased significant improvements in recognition performance, as can be seen in Figure 8.

On the other hand, backend fusion methods in underwater acoustic signal processing draw upon the principles of ensemble learning [65]. Zhang et al. [66] simultaneously extracted STFT magnitude spectrum, STFT phase spectrum, and bispectrum features of underwater signals, feeding each feature set into multiple networks and subsequently integrating their outputs to make decisions. This method takes full advantage of various time–frequency features to provide the integrated model with improved recognition accuracy and anti-noise robustness. Despite its capability to yield discriminative features, it is crucial to acknowledge that fusion-based feature extraction introduces considerable redundancy by combining multiple types of time–frequency representations. Moreover, these methods are sensitive to the selection of time–frequency feature combinations, as improper selection could result in suboptimal recognition performance.

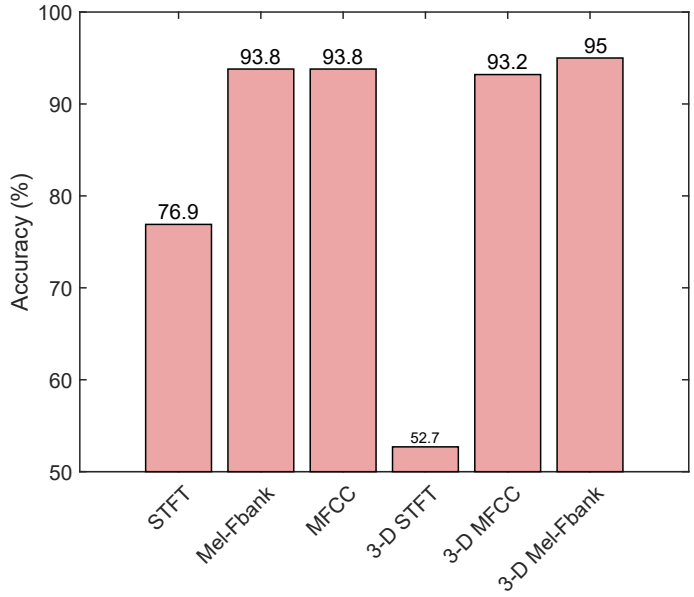


Figure 8. Accuracy comparison using various feature extraction methods with ResNet18 on the Shipsear dataset, as adapted from Wu et al. [64].

2.3. Autoencoding Feature

Manual feature extraction methods require complicated feature engineering, which involves numerous underwater acoustic processing techniques that are computationally intensive and inefficient. In contrast, autoencoding feature extraction methods have recently gained popularity in UATR due to their ability to automatically extract features [67,68], which is straightforward and efficient to achieve end-to-end UATR. AEs are the primary autoencoding feature extraction method in UATR [69,70]. In the unsupervised learning paradigm, AEs are originally designed for data compression, denoising, and sparse representation [71], which encode high-dimensional data into low-dimensional latent representations with the goal of reducing redundancy and capturing implicit patterns.

As shown in Figure 9, an AE consists of an encoder and a decoder, which together map an input x to an output x' while requiring that x' and x are sufficiently close, $x \approx x'$. The input can be a signal waveform or a spectrum. In the time domain, Dong et al. [72] designed a Bidirectional Denoising Autoencoder (BDAE) to compare the differences between original and denoised signals, thereby learning anti-noise robust characteristics of underwater acoustic signals. Experimental results demonstrated the effectiveness of BDAE features particularly in low SNRs. Li et al. [73] designed deep convolutional filter banks to decompose and merge ship-radiated noise, extracting deep time–frequency features through fully connected layers. Their research suggested that setting more filter banks can achieve better classification performance than advanced methods. From a spectral perspective, Luo et al. [74] used a Restricted Boltzmann Machine (RBM) for automatic encoding of power and demodulation spectra. Their method facilitated hierarchical feature extraction from underwater acoustic signals without the need for task-specific supervision. Experimental results showed that the proposed method can obtain discriminative signal representations to achieve effective UATR.

Moreover, self-supervised learning (SSL) methods can effectively extract automatic encoding features through well-designed upstream tasks. Generally, this learning paradigm can be categorized into generative and contrastive learning based on the upstream task. Contrastive learning focuses on learning meaningful representations by comparing differences between samples, whereas generative methods emphasize extracting features by leveraging the inherent structure of the data themselves. Combined with contrastive learning, Sun et al. [75] proposed the contrastive learning Underwater Acoustic Target Recognition (CCU) model for feature extraction. CCU dynamically extracts discriminative

features from diverse data inputs, achieving superior recognition performance compared to AEs. Furthermore, Wang et al. [76] proposed an Acoustic-embedding Memory Unit Modified Space AE (ASAE) for both generative and contrastive learning. This enhancement of space AE serves as an upstream task to efficiently extract high-level semantic information from acoustic target signals. Experimental results demonstrated that ASAE spectrogram features can achieve better recognition performance than mainstream time–frequency features.

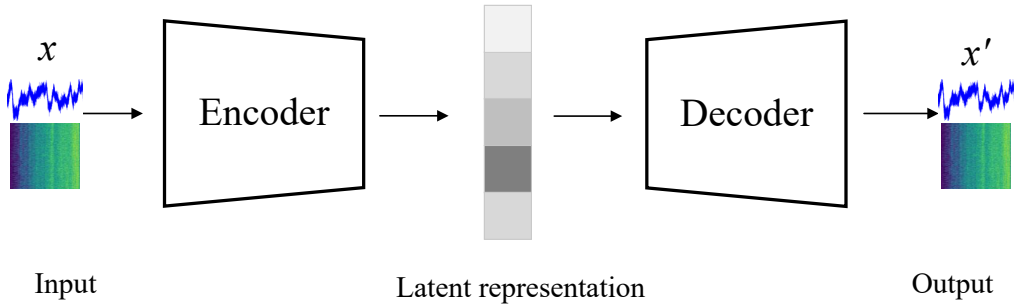


Figure 9. The basic architechture of AE.

While autoencoding features have shown success in UATR, these features essentially lack a close connection to physical mechanisms and target characteristics due to insufficient data preprocessing. Consequently, these features are considered difficult to provide intuitive physical explanations for, which are critical factors for model interpretability.

3. Machine Learning-Based Recognition Methods

Due to the simplicity, efficiency, and ease of implementation, ML methods have become widely utilized for intelligent UATR [77–79]. Their standard paradigm involves feeding extracted features into shallow ML algorithms, which then find classification boundaries based on feature similarity, effectively dividing compact sets in high-dimensional space. Table 1 provides an overview of representative ML-based UATR methodologies.

Among these ML methods, distance-based methods such as SVMs [80] and KNN [81] are well-known for their effectiveness in discovering decision boundaries within the feature space, which is crucial for recognizing underwater acoustic signals. As the most popular ML methods for UATR, SVMs employ kernel functions to map signal features into high-dimensional spaces, where separable boundaries between data points can be always found, ensuring different category samples lie on opposite sides of these boundaries while maintaining sufficient margin distance, thereby addressing the task of UATR. Many studies demonstrated that combining an SVM with individual features can achieve high classification results [82–86]. However, these studies did not explore the performance of multidimensional feature approaches. Liu et al. [87] demonstrated that using multiple features can lead to better results on real-world acoustic datasets.

Table 1. Representative works of ML-based UATR methods. Note that the default metric is recognition accuracy. Adjusted Rand index (ARI) is a widely used clustering evaluation metric.

Reference	Feature	Method	Dataset	Result
Li et al. [82]	wavelet analysis	SVM	three-category dataset	90.12%
Yang et al. [88]	feature selection	SVM	UCI sonar dataset [89]	81%
Moura et al. [83]	LOFAR	SVM	four-category dataset	76.73%
Sherin et al. [84]	MFCC	SVM	four-category dataset	74.28%
Wang et al. [85]	STFT	SVM	four-category dataset	88.64% (10 dB)

Table 1. *Cont.*

Reference	Feature	Method	Dataset	Result
Yao et al. [86]	STFT	SVM	two-category dataset	90%
Liu et al. [87]	multidimensional fusion feature	SVM	4-category dataset	97%
Choi et al. [90]	cross-spectral density matrix	SVM	simulation dataset	97.98%
Wei et al. [91]	MFCC	SVM	five-category dataset	93%
Chen et al. [92]	spectral ridge	SVM	four-category whale call dataset	99.415%
Yaman et al. [93]	512-dimensional feature vector	SVM	five-category propeller dataset	99.8%
Saffari et al. [94]	FFT	*K*NN	ten-category ship simulation dataset	98.26%
Li et al. [95]	complex multiscale diffuse entropy	*K*NN	four-category dataset	96.25%
Alvaro et al. [81]	-	*K*NN	five-category dataset	98.04%
Jin et al. [96]	eigenmode function	*K*NN	Shipsear dataset [1]	95%
Mohammed et al. [97]	GFCC	HMM	ten-category dataset	89%
You et al. [98]	GFCC	HMM	simulation dataset	90% (8 dB)
Yang et al. [17]	mutual information	LR	UCI sonar dataset [89]	94.7%
Seo et al. [99]	FFT	LR	simulation dataset	77.43%
Yang et al. [100]	auditory cortical representation	LR	3-category dataset	100%
K et al. [101]	mutual information	decision tree	UCI sonar dataset [89]	95%
Yaman et al. [93]	512-dimensional feature vector	decision tree	five-category dataset	99%
Yu et al. [102]	covariance matrix	decision tree	two-category dataset	99.2%
Zhou et al. [103]	multicorrelation coefficient	random forest	South China Sea dataset	93.83%
Choi et al. [90]	cross-spectral covariance matrix	random forest	two simulation datasets	96.83%
Chen et al. [92]	spectral ridge	random forest	four-category whale call dataset	99.69%
Wang et al. [104]	MFCC	GMM	four-category dataset	ARI 77.97
Sabara et al. [105]	-	GMM	SUBECO dataset [105]	74%
Yang et al. [106]	MFCC	GMM	five-category dataset	-
Yang et al. [106]	MFCC	fuzzy clustering	five-category dataset	-
Agersted et al. [107]	intensity spectrum	hierarchical clustering	Norwegian Institute of Marine Research	Best clusters = 7

*K*NN method predicts new samples by voting rule with the k nearest training data points. Common distance metrics $d(x, y)$ used to measure the similarity, such as Euclidean, Manhattan and Minkowski distances. Li et al. [95] introduced refined composite multiscale dispersion entropy (RCMDE) features for ship-radiated noise and utilized *K*NN for target recognition. Comparative experiments on 4-category ship dataset demonstrated that RCMDE-*K*NN can effectively capture the implicit pattern of underwater signals, which achieved a recognition accuracy at 96.25%. Jin et al. [96] extracted intrinsic mode function (IMF) feature based on target signal center frequency and energy intensity. This feature is then fed into *K*NN algorithm for nearshore ship recognition. Experimental results on the Shipsear dataset [1] demonstrated its effectiveness with a 95% recognition accuracy. However, the strength of these distance-based methods comes with the trade-off of potentially high computational demands due to the need to compute multiple distances [83,84]. To reduce spatial complexity, Yang et al. [88] proposed a novel ensemble SVM approach through weighted sampling and feature selection, named as the WSFSelect-SVME. This method integrates multiple SVMs classifiers to obtain robust recognition performance in real-world datasets. Additionally, Wang et al. [85] explored the use of DL algorithms for feature extraction, followed by SVM for chromatic feature classification. Experimental results

on noisy datasets demonstrated its effectiveness, which achieved an average recognition rate of 94.92%.

On the other hand, HMM [97] and GMM [104] are two representative probability-based models derived from natural language processing (NLP), which can provide rich probabilistic explanations on the acoustic data. Specifically, HMM is utilized to characterize the temporal sequence of underwater acoustic signals, capturing dynamic sound variations over time through state transition and observation probabilities [97,98]. On the other hand, GMM models the static features of underwater acoustic signals, emphasizing the statistical properties of various frequency components using Probability Density Function modeling [106]. Despite their success in underwater acoustic target recognition, HMMs and GMMs are are acknowledged for their challenging training processes and sensitivity to parameter selection. These characteristics can significantly influence the practicality of deploying these models in ocean applications. To address this problem, various optimization methods, such as genetic algorithm [98], particle swarm algorithm [28] have been integrated into these models to reduce the sensitivity, which have achieved promising results.

Additionally, tree-based methods shine for their ability to provide a visual representation of the decision-making process. In the field of UATR, decision trees and random forest are commonly used tree-based methods. Specifically, decision trees work by recursively partitioning the data space based on individual acoustic features. Its architecture is intuitive: internal nodes represent features, branches signify decision rules, and leaf nodes indicate the outcomes, which has made certain achievements using various acoustic features [93,102]. Expanding on this foundation, random forests construct multiple trees, each contributing to a collective prediction, which can improve the generalization and reduce the risk of overfitting. The randomness introduced in the selection of data subsets and features for each tree is important to the robust UATR performance [90,103]. However, to avoid overfitting, a careful balance of model complexity is required for this ensemble method.

Lastly, regression methods, such as linear regression (LR) [17] has been proven its effectiveness in UATR. Yang et al. [17] conducted preprocessing of sonar-collected underwater signals to isolate critical attributes such as signal intensity, frequency, and reflection angles. By integrating feature selection techniques, they successfully performed binary classification tasks on rocks and mines using LR models. Seo et al. [99] also employed LR models for UATR. They initially used LR to filter target signals from strong noise environments and further trained the model using simulated data to identify real-world samples, which achieved efficient UATR under limited real data conditions. It is noteworthy that while this method alleviates data requirements with simulated data, the fidelity to actual marine environments remains a challenge, impacting performance and requiring substantial time investment for data generation. Despite the efficiency and simplicity of regression methods, their performance can be limited by the linear assumption, which may lead to suboptimal results in the presence of complex, nonlinear datasets where they might yield suboptimal results.

Overall, the successful application of various ML methods demonstrates their potential in handling underwater acoustic data, which further drives intelligent development in UATR. These advances have provided a strong theoretical foundation for DL-based methods, while also providing valuable insights into underwater data processing. Nonetheless, ML algorithms typically apply limited linear or nonlinear transformations on acoustic data, capturing shallow internal structures. In scenarios with limited samples and constrained computational resources, their ability to represent complex underwater signals is diminished, which limits their generalization and recognition accuracy.

4. Deep Learning-Based Recognition Methods

The DL-based intelligent UATR methods effectively utilize various neural networks to classify underwater targets. By mimicking the neural structure of the human brain, DL methods rely on layered architectures to learn deep abstract patterns within data. Based

on learned universal principles, these methods analyze, reason, and predict outcomes for unseen samples, as illustrated in the standard prediction framework shown in Figure 10.

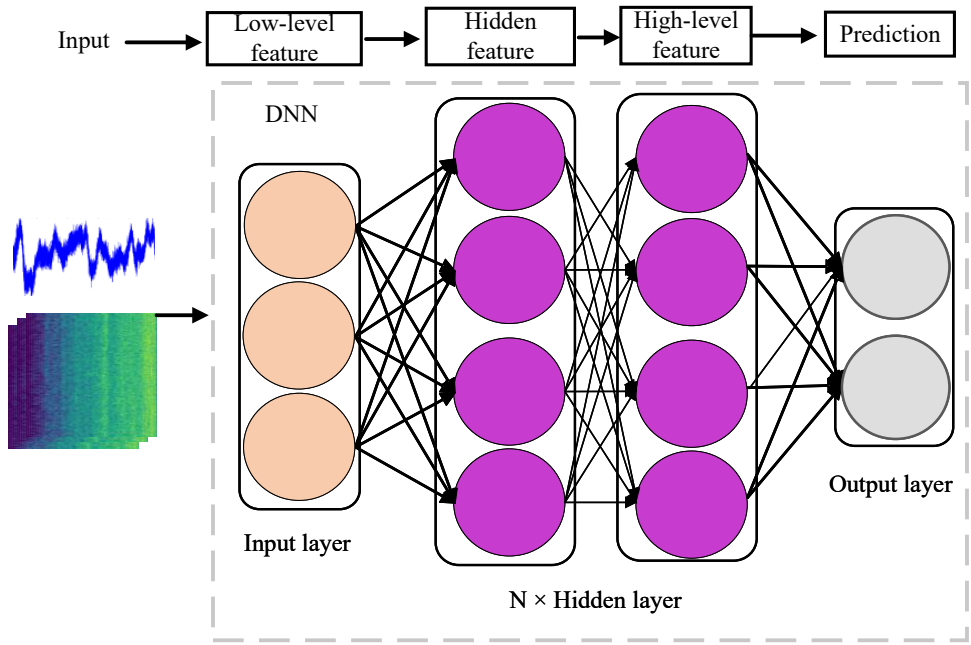


Figure 10. A standard prediction framework of DL-based UATR methods.

Within this paradigm, as DNN progress through layers, input features transform from low-level to high-level representations after passing through multiple hidden layers. These high-level features are finally integrated through the output layer to make the final prediction. Due to their superior signal processing capabilities, DL-based methodologies have emerged as a compelling alternative to conventional ML techniques, and have become a prominent area of current research [18,108].

Figure 11 depicts the advantages and disadvantages of representative DL-based UATR methods. Among these methods, shallow neural networks (SNNs) are firstly applied for UATR, such as multilayer perceptions (MLP) [109,110] and Restricted Boltzmann Machines (RBMs) [111,112], which achieved certain successes in underwater signal processing. However, their shallow architecture limited the ability to discern intricate patterns. The complexity of underwater environments and the instability of data necessitate the use of DNNs, particularly under conditions of few-shot and low SNRs.

Recurrent neural networks (RNNs), convolutional neural networks (CNNs), attention neural networks (ATNs), and Transformers have risen as pivotal DNN-based methodologies for intelligent UATR, demonstrating significant effectiveness in practical scenarios. Other DL technologies, such as generative adversarial networks (GANs) [26,113], transfer learning [114], and SSL [115,116], have opened new frontiers for intelligent UATR, showcasing enormous development potential. Importantly, the data learning processes in DL often involve substantial amounts of underwater acoustic signals, making the partitioning of datasets a critical factor that influences performance. Two of the most extensively studied datasets within this field are Shipsear [1] and Deepship [2]. The partitioning of these datasets is typically conducted using two principal methods. The first method is known as the random partitioning approach. This approach can potentially lead to a misleadingly optimistic assessment of model performance. This can occur if a data segment that is temporally subsequent is present in the training set, while an earlier segment is allocated to the test set. Such a scenario can lead to an overestimation of recognition performance, as the model may inadvertently benefit from future information that is not available in

real-world conditions. The accuracy comparison of widely utilized DNNs employing random partitioning is depicted in Figure 12a.

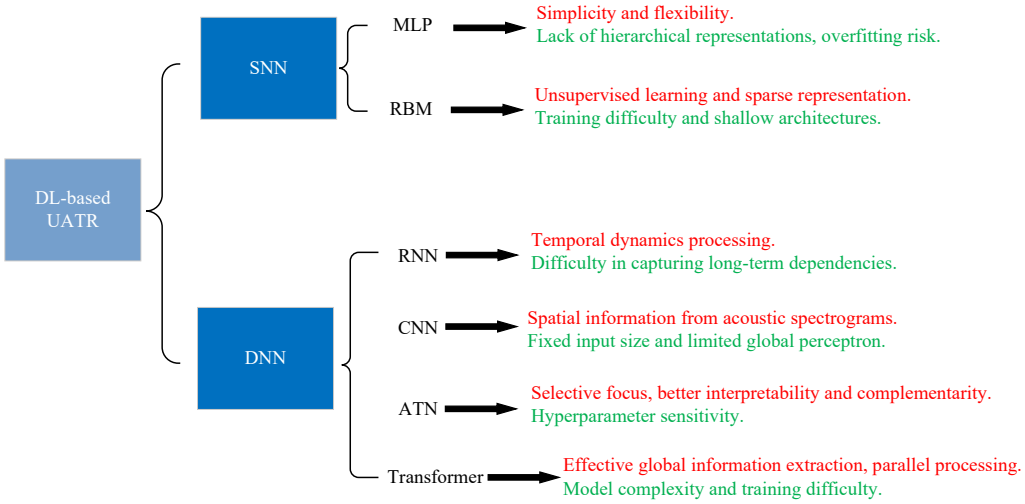


Figure 11. Representative DL neural networks in the field of intelligent UATR. The advantages and disadvantages of each method are marked in red and green, respectively.

The second method, known as causal partitioning, provides a more reasonable approach to dataset segmentation. Specifically, this technique segments audio files in accordance with their temporal sequence, ensuring that the model is not exposed to data from the future. In this way, it effectively prevents data leakage and provides a more accurate assessment of the prediction performance. The recognition accuracy comparison of DNNs leveraging causal partitioning is presented in Figure 12b. Causal partitioning respects the temporal integrity of the data, which is essential for training models that can generalize well to new, unseen data while maintaining the temporal causality inherent in time-series data. Overall, in the field of UATR, the mainstream recognition models are evolving from CNNs to Transformers. As can be seen from Figure 12, the advanced Transformer models have achieved superior recognition performance across both partitioning methods compared to CNN models. Moreover, the continuous integration of various cutting-edge computer techniques on Transformer architecture, including SSL [117,118], is expected to further improve recognition capabilities.

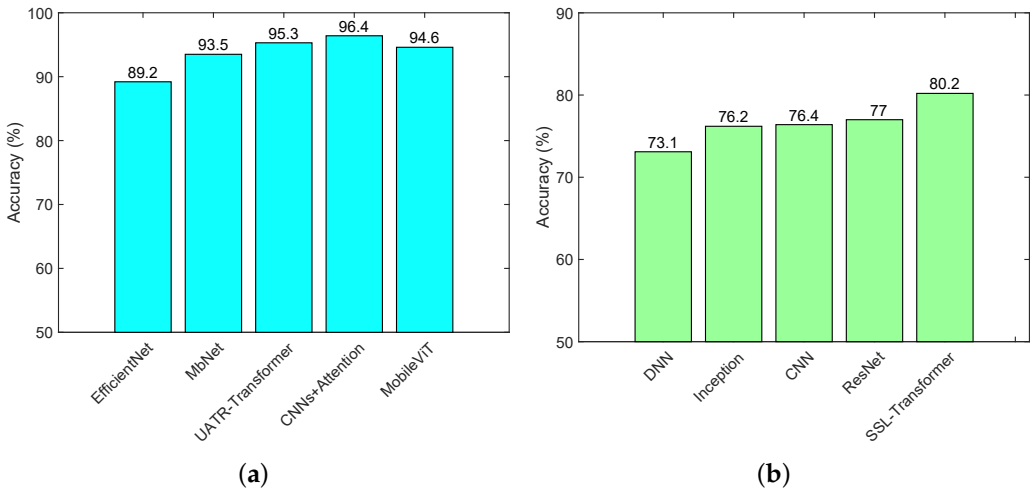


Figure 12. Accuracy comparison with different DNNs on the Deepship dataset. (**a**) Random partition, adapted from Zhou et al. [119], (**b**) causal partition, adapted from Irfan et al. [2] and Xu et al. [117].

4.1. RNN

In the field of DL, the RNN is a type of DNN capable of processing continuous data such as text [120] and speech [121]. Its fundamental concept involves utilizing internal recurrent connections to retain information from previous inputs, which are then integrated into current outputs. By leveraging these recursive connections, RNN-based UATR methods can directly handle temporal waveform data through memory functions that dynamically leverage temporal correlations, enabling accurate target classification [122–124]. In these methods, long short-term memory (LSTM) is one of the most commonly used RNN architectures [125,126]. Figure 13 presents the structure of the computational units in LSTM. LSTM leverages three gating mechanisms (input gate, forget gate, and output gate) that facilitate dynamic utilization of long-term dependencies to process underwater acoustic signals, which can address the problem of gradient vanishing and exploding commonly encountered in traditional RNNs. Specifically, within LSTM, gates controlled by activation functions σ (typically sigmoid functions in input and forget gates) regulate the information flow. Initially, the model input x_t is selectively processed by the forget gate in conjunction with the previous hidden state s_{t-1} to determine which information to retain or discard from the cell state C_{t-1}.

$$f_t = \sigma\Big(W_f \cdot [s_{t-1}, x_t] + b_f\Big). \tag{15}$$

Subsequently, the input gate determines update values i_t, creating a new candidate cell state C'_t via tanh function, updated alongside i_t.

$$\begin{aligned} i_t &= \sigma(W_i \cdot [s_{t-1}, x_t] + b_i) \\ C'_t &= \tanh(W_C \cdot [s_{t-1}, x_t] + b_C). \end{aligned} \tag{16}$$

Historic information is retained via multiplication by f_t to form new cell state C_t.

$$C_t = f_t * C_{t-1} + i_t * C'_t. \tag{17}$$

Finally, the output gate, through a sigmoid layer, determines which information within the cell state should be passed to the next hidden state. The processed cell state C_t, after tanh function, is multiplied by the output of the sigmoid output gate, ultimately determining information passed to the subsequent hidden state.

$$\begin{aligned} o_t &= \sigma(W_o[s_{t-1}, x_t] + b_o) \\ s_t &= o_t * \tanh(C_t). \end{aligned} \tag{18}$$

Yu et al. [127] leveraged the robust memory selection capability of LSTMs to process instantaneous features extracted from underwater communication signals, achieving recognition accuracy exceeding 80%, which demonstrates its effectiveness in modulation pattern recognition. Moreover, Yang et al. [128] employed pretrained LSTMs and softmax classifiers to classify ship-radiated noise in severe underwater communication environments. Furthermore, Zhang et al. [125] exploited backend fusion with multiple LSTMs to process various acoustic features; the final feature maps are integrated into softmax layers to make predictions, which achieved superior recognition performance over single-feature approaches.

Gated recurrent units (GRUs), another variant of RNNs, offer a simplified architecture compared to LSTM. Particularly, GRUs merge cell states and hidden states into a unified hidden state and incorporate two gates (update and reset gates) to process acoustic signals [129], which contain fewer parameters, making GRUs suitable for deployment in real-time recognition systems. Qi et al. [130] utilized GRUs to extract fine-grained features from acoustic spectrograms, achieving satisfactory recognition performance with real ocean data. Moreover, Wang et al. [131] proposed a hybrid time-series network for modulation signal recognition in complex underwater communication environments. By integrating

bidirectional GRUs to enhance the hidden representations of acoustic signals, the proposed method demonstrated robust recognition capability under severe interference.

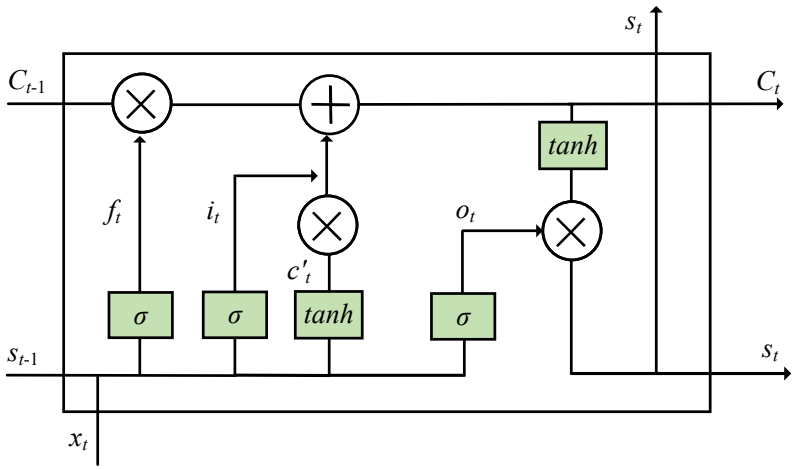


Figure 13. The computational units in LSTM.

To enhance feature representations, a widely adopted strategy integrates a CNN-based local feature extractor in front of RNNs [132]. Kamal et al. [133] proposed a hybrid CNN-LSTM model for recognizing underwater signals with real ocean data. The proposed model employed convolutional filters to learn sufficient time–frequency representations, subsequently extracting discriminative features via 2D convolutions, which achieved an impressive recognition accuracy of 95.2%. Moreover, Qi et al. [134] introduced an intelligent UATR method that combines LSTM with 1D CNN, where the LSTM module was utilized to capture phase and frequency characteristics from acoustic signals. Experimental results demonstrated its effectiveness to achieve superior recognition performance compared to a single network.

4.2. CNN

Inspired by the notable success of CNNs in handling natural images, early researchers directly applied side-scan and forward-scan sonar image data inputs to CNNs for UAT and have made certain achievements [135,136]. Subsequently, researchers began to explore the direct application of signal waveform to CNNs for UATR [19,137]. In comparison to processing time-domain waveforms, the sliding window in CNN is particularly suitable for extracting discriminative features from high-resolution time–frequency spectrograms [138]. In practice, the representation of time–frequency spectrograms can effectively capture stable energy distribution characteristics, which have shown superior recognition performance compared to waveform inputs [108,139]. In this case, the integration of CNNs with time–frequency spectrograms has been considered as the baseline method in the field of UATR [140,141]. ResNet [142] stands as a common CNN architecture for UATR, with its structure illustrated in Figure 14. Its main innovation, the residual block, consists of two consistently sized convolutional layers with a skip connection that adds input directly to the output of the block, formulated as $H(x) = F(x) + x$. This residual adding operation helps ResNet mitigate gradient vanishing because it allows gradients to propagate directly through to earlier layers, thus maintaining information as the network goes deeper. Additionally, hierarchical convolution and pooling layers are employed in standard ResNets to reduce resolution. Building upon the architecture of ResNet, researchers developed DenseNet [143], which incorporates dense connections between all layers for feature reuse and extraction efficiency. Both ResNet and DenseNet have been extensively researched for their advantages in extracting detailed texture features and facilitating spatial information flow, establishing their significance in UATR [144–146].

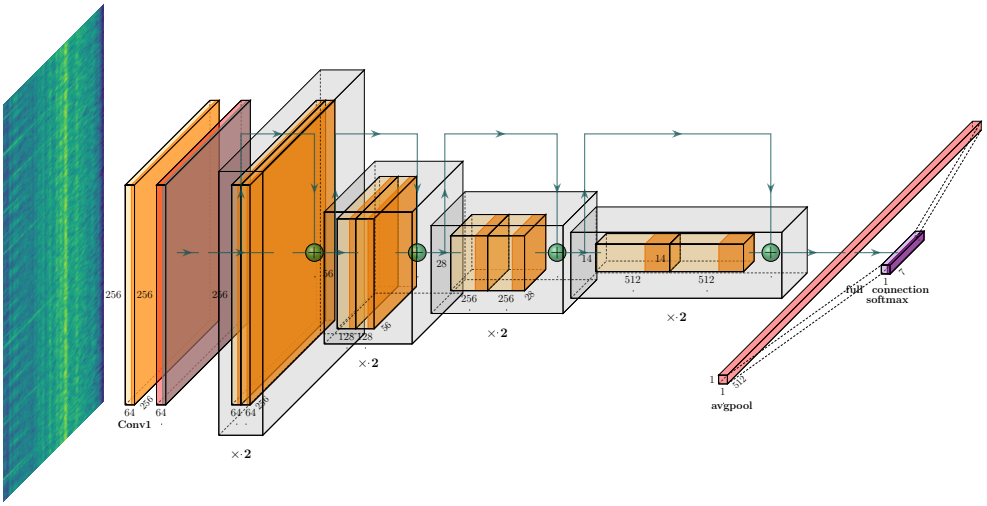


Figure 14. The UATR framework based on ResNet18, which commonly accepts the acoustic spectrograms as model input.

To improve efficiency, various convolution schemes have been further developed in CNNs. Depth-wise convolution (DW) [147] has been integrated into lightweight CNN architectures to increase feature discriminability while reducing computational costs [148]. Hu et al. [149] constructed CNNs using DW and dilated convolution mechanisms for passive UATR. The experimental results demonstrated that it outperformed traditional CNNs by extracting discriminative features. Using an Inception convolution block [150], Zheng et al. [151] applied the STFT features of underwater acoustic signals to GoogleNet [152] for shallow-water underwater target detection, demonstrating improved detection performance over energy detectors. Moreover, Cai et al. [153] leveraged Xception, a variant of Inception networks [147] optimized with DW convolutions, to develop a multibranch DNN capable of recognizing underwater targets. The proposed approach integrated Xception for detecting line spectral pair features, alongside other branches including MobileNet [154] and DenseNet, which achieved average recognition accuracies exceeding 95% across six classes of marine biological sounds. Considering hardware constraints and real-time requirements, a variety of lightweight CNNs [155–157] have been introduced with robust feature processing capabilities in real-time UATR tasks.

4.3. ATN

Underwater target signals essentially represent natural time series in which critical frequency information is essential for target classification. Attention mechanisms can effectively focus on these critical frequencies, thereby potentially enhancing feature processing capabilities. Xiao et al. [158] developed an interpretable ATN-based UATR method using STFT spectrograms, demonstrating significant potential to apply attention into UATR. By combining attention mechanisms with multiscale CNNs [119], ATNs can obtain superior recognition performance compared to baseline CNN and RNN models. Particularly, scalable attention modules, such as channel attention mechanism (CAM) [159] and spatial attention mechanism (SAM) [160], have been effectively integrated with CNNs to improve the feature extraction from signal spectrograms [161,162].

Bidirectional attention [163], a special co-attention and self-attention fusion technique, is designed to comprehensively capture semantic and structural information between different regions within model inputs. Liu et al. [164] combined bidirectional attention with Inception, dynamically weighting the receptive field to eliminate irrelevant backgrounds, thereby achieving satisfactory recognition accuracy. In addition, Wang et al. [161] introduced AMNet, a model combining multibranch backbone networks with convolutional attention modules. These modules selectively weight global information, thus adaptively

filtering key information within acoustic spectrograms to potentially enhance classification performance.

In summary, ATNs exhibit substantial potential for robust signal representation capabilities to enhance recognition accuracy in the field of UATR, which makes them promising research directions in intelligent UATR.

4.4. Transformer

The Transformer, originally proposed by Google for machine translation in natural language processing tasks [165], can be viewed as a specialized form of ATN. This model primarily utilizes a multihead self-attention (MHSA) to describe correlations across multiple subspaces. The MHSA is a novel attention mechanism that significantly improves model representation and computational efficiency. As depicted in Figure 15, MHSAs simultaneously compute and integrate semantic information from H distinct heads on query Q, key K, and value V sets, thereby outperforming conventional attention mechanisms.

To leverage its robust data processing capabilities, Feng et al. [141] introduced UATR-Transformer. Specifically, the proposed model employs hierarchical tokenization strategies to derive computational units from acoustic Mel-spectrograms. Further feature transformation through an MHSA demonstrates that UATR-Transformer can achieve recognition performance comparable to state-of-the-art CNNs due to its both global and local information perception. Inspired by Audio Spectrogram Transformer (AST) in speech recognition [166], Li et al. [167] adapted its structure into UATR, proposing the Spectrogram Transformer Model (STM), which also outperforms CNNs in UATR tasks. The Swin Transformer [168], a specialized variant utilizing shifted windows for hierarchical feature extraction, was introduced by Wu et al. [169] in UATR as the time–frequency Swin-Transformer (TFST). TFST exhibits superior performance over both CNNs and traditional Transformers in learning discriminative features from moving targets on two real ocean datasets. Despite their satisfactory recognition accuracy, a primary drawback of Transformers is the high complexity of MHSAs, which impacts their real-time performance. To address this problem, Yao et al. [170] combined MobileNet with Transformer architecture to design MobileViT, which effectively integrates global and local information perception of underwater acoustic signals with low model complexity, validated through empirical dataset evaluations. SSL-Transformers exhibit robust generalization capabilities, making them powerful feature learners for underwater acoustic signals. To this end, several studies have explored SSL-Transformer methods for UATR [118,171], focusing on upstream tasks to learn generalized representations of underwater acoustic signals without label information, thereby addressing the challenge of limited data in UATR. While Transformers have demonstrated promising results in UATR, challenges of complex training requirements necessitate the integration of diverse training methods, including data augmentation [172] and knowledge distillation [173].

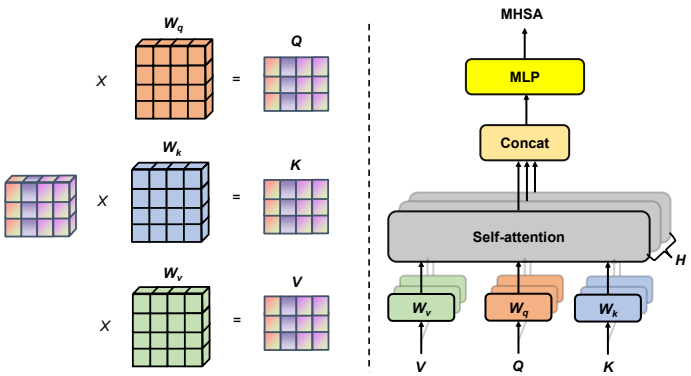


Figure 15. The MHSA mechanism in Transformer. The left side describes the self-attention mechanism, and the right side describes the MHSA. W_q, W_k, W_v comprise the learnable projection matrix.

5. Challenges and Future Prospects

As previously mentioned, the rapid development of AI has greatly promoted the application of data-driven methods in the field of UATR. In particular, the current DL methods have fully exploited the flexibility and scalability of DNNs to design advanced recognition systems. However, intelligent UATR still faces several serious challenges, which can be summarized as follows: the complex recognition condition issue, poor interpretability, weak generalization, and low robustness under adversarial attacks.

5.1. Complex Recognition Condition Issue

Recognizing underwater acoustic targets in conditions of low SNR and few-shot scenarios indeed poses a significant challenge in the field of underwater acoustics. As aforementioned, the multiple interferences and complex channels present in the ocean environment significantly increase the difficulty of UATR. Recognition tasks under low SNRs are further complicated by these challenges. The low SNR makes it extremely difficult to distinguish the weak signals from background noise, thereby complicating the model's ability to learn deep, discriminative features. To address this issue, many researchers have been dedicated to developing robust recognition methods, particularly through the use of well-designed DL models. Zhou et al. [174] introduced a cross-attention fusion joint training framework that synergizes denoising and recognition models for robust UATR. This framework has been shown to effectively tackle the UATR task in noisy environments by enhancing the SNR and improving feature extraction. Similarly, Li et al. [175] proposed a robust UATR method inspired by auditory passive attention mechanisms. Their approach considered three auditory passive attention loss functions to achieve robust UATR under low SNR conditions, achieving an average F1-score of 67.43% (SNR = -18 dB) and a notably high score of 90.79% (SNR = 6 dB). Additionally, several data preprocessing techniques, such as dual-path denoising [176] and adaptive filter techniques [177], can be also beneficial in addressing the challenges posed by low SNR conditions.

Due to the inherent challenges in acquiring high-quality underwater acoustic data, the task of few-shot UATR has become another prevalent research focus. Yang et al. [178] utilized a Siamese network that consists of 1D convolution and an LSTM for few-shot recognition. Experimental results based on sea trial data demonstrated its effectiveness. However, this method is prone to overfitting, potentially leading to a performance degradation. To address this issue, SSL methods have been increasingly adopted [179,180], aiming to enhance model generalizability and robustness with minimal reliance on labeled data. Common SSL techniques used in UATR can be broadly categorized into contrastive SSL [179,180] and generative SSL methods [117,171]. As depicted in Figure 16, contrastive SSL focuses on learning robust representations by comparing original data against its transformed versions, aiming to distinguish between positive and negative pairs to minimize the contrastive loss. On the other hand, generative SSL tasks are designed to capture the intrinsic structure of the data, enabling the model to generate new samples that closely resemble the original sample. In UATR, a generative model can be trained to reconstruct missing parts of a sonar signal or spectrogram and then optimized by minimizing the generative loss.

Figure 17 illustrated the accuracy comparison of few-shot models based on the Shipsear dataset. Among them, the contrastive SSL-based methods, CDCF achieves better results than conventional few-shot models, which demonstrates that SSL can help DL models to learn discriminative features from scarce acoustic data, thereby improving recognition accuracy in few-shot UATR tasks.

For future directions to address the issue of complex recognition conditions, further leveraging more powerful meta-learning approaches [181,182] with various feature extraction methods could be a promising strategy. This method is expected to help DL models to rapidly learn discriminative features in complex recognition environments.

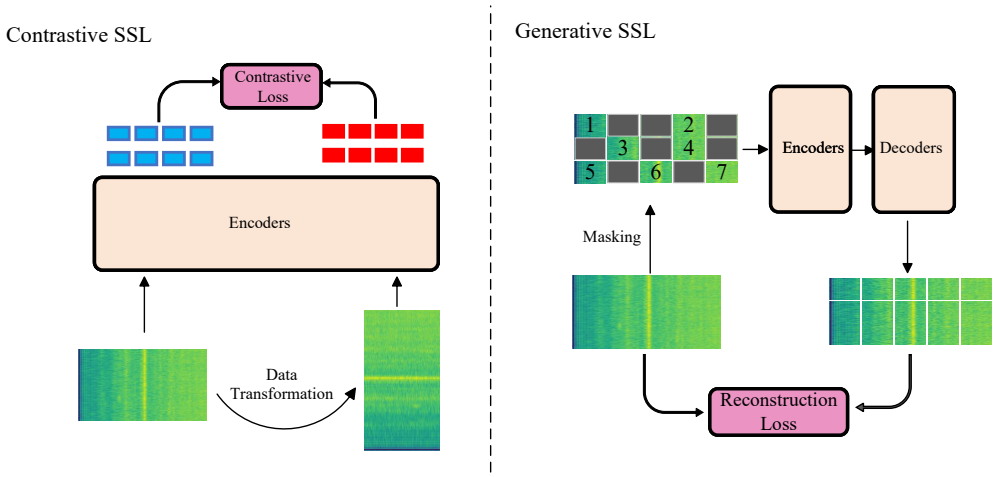


Figure 16. Contrastive and generative SSL methods for UATR.

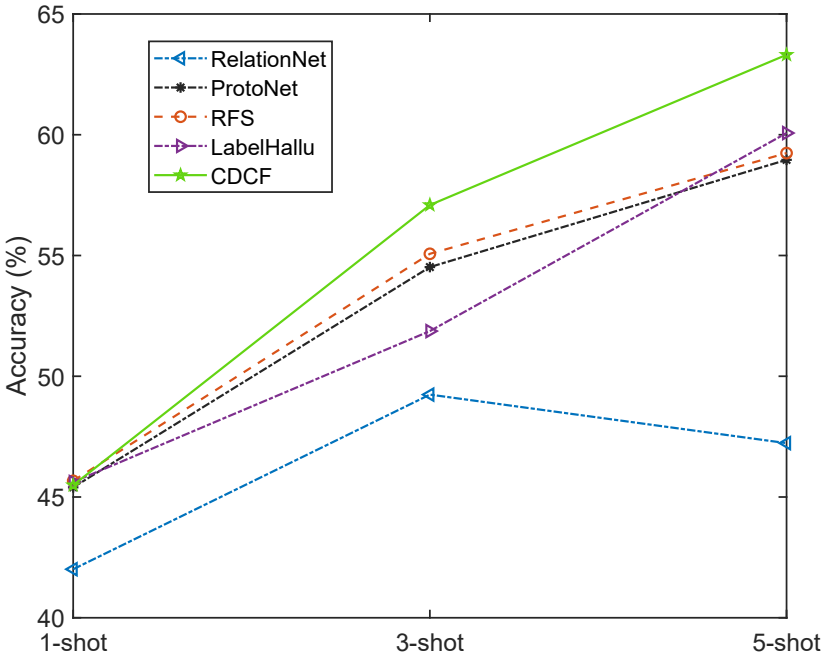


Figure 17. Accuracy comparison on three few-shot tasks based on Shipsear dataset, as adapted from Cui et al. [179].

5.2. Interpretability Problem

The challenge of the interpretability problem comes from the DL itself as a "black box" [183] and the complex mechanism of underwater acoustic signals. Currently, it remains unclear which discriminative features neural networks actually learn and whether temporal or frequency structures play crucial roles in the recognition process for all these models. In the field of computer vision, while visualizing feature maps partially explains how models recognize specific object contours by capturing texture and shape, these methods fail to fully reveal the inner workings of the models [38,139,184]. Even in the domain of speech recognition, which shares similarities with UATR tasks [185,186], existing interpretability studies have not provided a clear physical mechanism explanation due to lacking specific semantic information in acoustic signals.

On the other hand, UATR is closely related to underwater physics. The complex acoustic propagation channels and dynamic marine environments have substantial influ-

ence on acoustic signal transmission, which leads to nonstationary, non-Gaussianity, and nonlinearity characteristics. These complexities pose challenges in extracting crucial information within the underwater signals. As a result, DL models may encounter difficulties in effectively perceiving important physical information during decision making, which affects their interpretability.

To address this challenge, it is crucial to keep studying feature map visualization in UATR, such as Grad-CAM [146,187], t-distributed stochastic neighbor embedding (T-SNE) reduction [180,188,189], and attention [158] mechanisms. Among these methods, Grad-CAM and attention mechanisms are more inclined to explore interpretability from the underwater acoustic spectrogram, while t-SNE is more focused on learning the data distribution to enhance the interpretability, as can be seen in Figure 18. Generally speaking, these methods are mostly driven by image-based approaches without considering intrinsic properties. As a result, forthcoming interpretability research will focus on the perspective of features rather than simply treating acoustic features as natural images. The introduction of graph processing may be a promising solution because graphs are well suited to represent complex relationships [190]. By modeling signal features or signal samples as graphs, it is possible to form connected graphs that reveal the correlations between regions, thereby enhancing model interpretability.

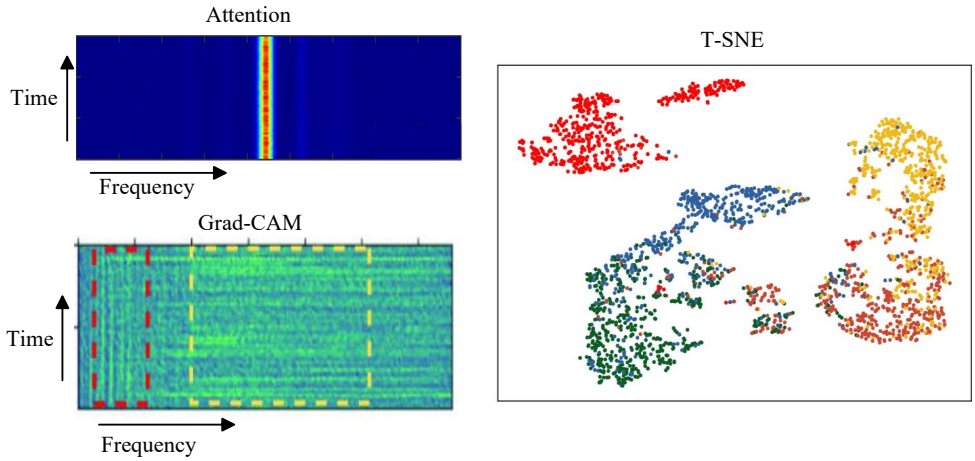


Figure 18. Interpretable methods in intelligent UATR.

5.3. Generalization Issue

In a standard intelligent UATR paradigm, both training and testing datasets typically come from the same sea areas and ship types. However, real ocean recognition scenarios often involve unseen sea areas and ships, leading to significant differences between data distributions known as environment mismatch. In such cases, the model's generalization ability to new data can be reduced, leading to a significant decrease in accuracy. To address this problem, data augmentation [63,114,191] and transfer learning [167,192] techniques are introduced into this field. However, due to the vast differences in actual marine environments, the volume of data used is still insufficient to leverage the advantages of data-driven learning. Consequently, these methods have not yet been able to consistently deliver high recognition accuracy across different sea areas.

The recent advancements of universal large models may further enhance the generalization of recognition models. Large DL models that have hundreds of millions of parameters, such as ChatGPT [193], Whisper [194], Pangu-Weather [195], and Xihe-ocean forecasting model [196], have achieved considerable success in their respective fields and demonstrate strong generalization capabilities even in zero-shot tasks. However, due to insufficient acoustic data in the field of underwater acoustic signals, it is impractical to directly train a universal recognition model to classify underwater signals. To address this challenge, advanced transfer learning techniques may benefit from universal model

training [197], which involves aligning acoustic data with upstream tasks. Additionally, combining these techniques with efficient fine-tuning methods, such as Lora [198], QLora [199], and Adapter Tuning [200], holds potential for resolving the generalization problem arising from environmental mismatches.

5.4. Adversarial Robustness Challenge

Numerous studies have demonstrated the inherent fragility of AI methods. Specifically, well-designed perturbations have been shown to significantly degrade model performance, thereby reducing their reliability [201,202]. In practice, these perturbations create adversarial samples that cause neural networks to learn discontinuous mappings. Such adversarial samples can significantly confuse the backpropagation process of DL models, resulting in unexpected misclassifications. In safety-critical domains, the presence of adversarial samples poses a substantial threat to various intelligent applications, potentially leading to severe consequences [203]. Take autonomous driving as an example. Adversarial samples generated by altering light brightness or modifying traffic signal indicators can cause in-vehicle intelligent devices to malfunction, ultimately resulting in accidents [204]. UATR also belongs to the safety-critical area with significant military implications [205]. Adversarial attacks on intelligent recognition systems could lead to unexpected and severe outcomes with the misclassification of ships or inadequate detection of critical underwater targets.

The impact of adversarial attacks on intelligent recognition systems is illustrated in Figure 19. Taking the label data of motorboats as an example, both the waveform signals and the time–frequency spectrograms are subjected to small perturbations when under adversarial attacks, resulting in imperceptible adversarial data. These adversarial data are essentially indistinguishable from the original data, yet they lead to the classifier misclassifying motorboat samples as oil tankers. Such failures have the potential to impact strategic decisions, carrying serious implications for military operations and national security.

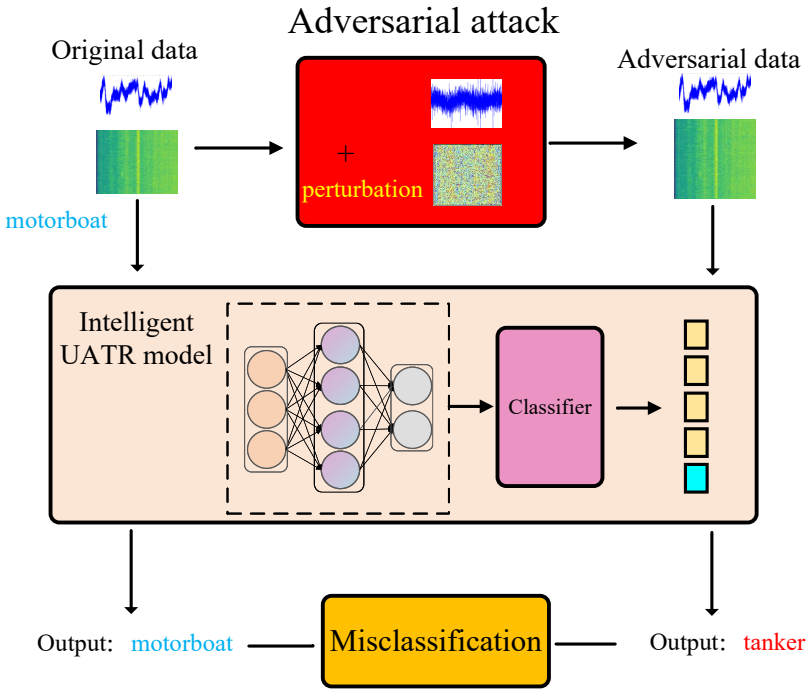


Figure 19. Misclassification caused by adversarial attacks on intelligent UATR systems.

In [135], Ma et al. investigated the Lambertian-based adversarial attacks in the field of underwater side-scan sonar image classification, which demonstrated higher attack success rate with other adversarial attacks. Moreover, Feng et al. [206] evaluated the adversarial robustness of the advanced intelligent UATR systems based on acoustic spectrogram inputs. Experimental findings indicate that both CNNs and Transformers significantly degrade the

recognition performance in the presence of adversarial perturbations. However, it is worth noted that their study is based on data from the receiver side without considering the transmitter side. Since the UATR task is closely related to the transmitter, acoustic channel, and receiver, the best practice is to design the adversarial attack algorithm from the transmitter to attack the intelligent recognition system at the receiver. Such an approach is limited by the scarcity of underwater acoustic samples and the privacy of ship physical parameters.

For adversarial defense that shares insights, it is more appropriate to investigate it from the receiver side, since in practice the receiver may not clearly know the attack type and source. A major future direction in this area is to investigate intelligent recognition systems under strong adversarial attacks. Potential solutions include adversarial training [207], adversarial sample detection [208], and defensive distillation [209].

6. Conclusions

With the rapid progress in AI, AI-based UATR has seen significant maturation for effective ocean remote sensing. To further enhance the understanding of this field, this paper provides a comprehensive review of both traditional and state-of-the-art feature extraction and classification methods in AI-based UATR, highlighting their respective strengths and potential shortcomings. In addition, the current prevalent issues of the complex recognition condition, generalizability, interpretability, and adversarial robustness are thoroughly discussed. To address the above challenges, innovative research directions and suggested methodologies that hold promise are proposed for future research. These directions will address current challenges and are expected to become significant research areas in UATR. As these technologies evolve, they promise to revolutionize ocean remote sensing, offering new insights and capabilities for underwater exploration and monitoring.

Author Contributions: Conceptualization, S.F. and S.M.; investigation, S.F. and S.M.; project administration, S.M.; resources, M.Y.; supervision, S.M. and M.Y.; visualization, S.F.; writing—original draft, S.F.; writing—review and editing, S.F., S.M., X.Z. and M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Defense Fundamental Scientific Research Program under Grant Number JCKY2020550C011.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Santos-Domínguez, D.; Torres-Guijarro, S.; Cardenal-López, A.; Pena-Gimenez, A. ShipsEar: An underwater vessel noise database. *Appl. Acoust.* **2016**, *113*, 64–69. [CrossRef]
2. Irfan, M.; Jiangbin, Z.; Ali, S.; Iqbal, M.; Masood, Z.; Hamid, U. DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. *Expert Syst. Appl.* **2021**, *183*, 115270. [CrossRef]
3. Niu, H.; Li, X.; Zhang, Y.; Xu, J. Advances and applications of machine learning in underwater acoustics. *Intell. Mar. Technol. Syst.* **2023**, *1*, 8. [CrossRef]
4. Zhang, X.; Yang, P.; Wang, Y.; Shen, W.; Yang, J.; Wang, J.; Ye, K.; Zhou, M.; Sun, H. A Novel Multireceiver SAS RD Processor. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–11. [CrossRef]
5. Zhang, X.; Yang, P.; Wang, Y.; Shen, W.; Yang, J.; Ye, K.; Zhou, M.; Sun, H. LBF-Based CS Algorithm for Multireceiver SAS. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
6. Zhang, H.; Shafiq, M.O. Survey of transformers and towards ensemble learning using transformers for natural language processing. *J. Big Data* **2024**, *11*, 25. [CrossRef]
7. Zhang, X.; Yang, P.; Sun, H. Frequency-domain multireceiver synthetic aperture sonar imagery with Chebyshev polynomials. *Electron. Lett.* **2022**, *58*, 995–998. [CrossRef]
8. Zhang, X.; Yang, P.; Feng, X.; Sun, H. Efficient imaging method for multireceiver SAS. *IET Radar Sonar Navig.* **2022**, *16*, 1470–1483. [CrossRef]
9. Zhang, R.; He, C.; Jing, L.; Zhou, C.; Long, C.; Li, J. A Modulation Recognition System for Underwater Acoustic Communication Signals Based on Higher-Order Cumulants and Deep Learning. *J. Mar. Sci. Eng.* **2023**, *11*, 1632. [CrossRef]
10. Jiang, Z.; Zhang, J.; Wang, T.; Wang, H. Modulation recognition of underwater acoustic communication signals based on neural architecture search. *Appl. Acoust.* **2024**, *225*, 110155. [CrossRef]

11. Wu, J.; Chen, Y.; Jia, B.; Li, G.; Zhang, Y.; Yong, J. Optimal design of emission waveform for acoustic scattering test under multipath interference. In Proceedings of the 2020 5th International Conference on Communication, Image and Signal Processing (CCISP), Chengdu, China, 13–15 November 2020; pp. 102–106. [CrossRef]
12. Sumithra, G.; Ajay, N.; Neeraja, N.; Adityaraj, K. Hybrid Acoustic System for Underwater Target Detection and Tracking. *Int. J. Appl. Comput. Math.* **2023**, *9*, 149. [CrossRef]
13. Zhu, J.; Xie, Z.; Jiang, N.; Song, Y.; Han, S.; Liu, W.; Huang, X. Delay-Doppler Map Shaping through Oversampled Complementary Sets for High-Speed Target Detection. *Remote Sens.* **2024**, *16*, 2898. [CrossRef]
14. Zhu, J.; Song, Y.; Jiang, N.; Xie, Z.; Fan, C.; Huang, X. Enhanced Doppler Resolution and Sidelobe Suppression Performance for Golay Complementary Waveforms. *Remote Sens.* **2023**, *15*, 2452. [CrossRef]
15. Yoo, K.B.; Edelmann, G.F. Low complexity multipath and Doppler compensation for direct-sequence spread spectrum signals in underwater acoustic communication. *Appl. Acoust.* **2021**, *180*, 108094. [CrossRef]
16. Klionskii, D.M.; Kaplun, D.I.; Voznesensky, A.S.; Romanov, S.A.; Levina, A.B.; Bogaevskiy, D.V.; Geppener, V.V.; Razmochaeva, N.V. Solution of the Problem of Classification of Hydroacoustic Signals Based on Harmonious Wavelets and Machine Learning. *Pattern Recognit. Image Anal.* **2020**, *30*, 480–488. [CrossRef]
17. Quraishi, S.J.; Singh, M.; Prasad, S.K.; Arora, K.; Pathak, S.; Singh, A. A Machine Learning Approach to Rock and Mine Classification in Sonar Systems Using Logistic Regression. In Proceedings of the 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 1–3 November 2023; pp. 462–468. [CrossRef]
18. Wang, P.; Peng, Y. Research on Feature Extraction and Recognition Method of Underwater Acoustic Target Based on Deep Convolutional Network. In Proceedings of the 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China, 25–27 August 2020; pp. 863–868. [CrossRef]
19. Doan, V.S.; Huynh-The, T.; Kim, D.S. Underwater Acoustic Target Classification Based on Dense Convolutional Neural Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
20. Meng, Q.; Yang, S. A wave structure based method for recognition of marine acoustic target signals. *J. Acoust. Soc. Am.* **2015**, *137*, 2242. [CrossRef]
21. Deng, J.; Yang, X.; Liu, L.; Shi, L.; Li, Y.; Yang, Y. Real-Time Underwater Acoustic Homing Weapon Target Recognition Based on a Stacking Technique of Ensemble Learning. *J. Mar. Sci. Eng.* **2023**, *11*, 2305. [CrossRef]
22. Meng, Q.; Yang, S.; Piao, S. The classification of underwater acoustic target signals based on wave structure and support vector machine. *J. Acoust. Soc. Am.* **2014**, *136*, 2265. [CrossRef]
23. Sun, Y.; Zhang, X. Analysis of Chaotic Characteristics of Ship Radiated Noise Signals with Different Data Lengths. In Proceedings of the OCEANS 2022, Chennai, VA, USA, 17–20 October 2022; pp. 1–7. [CrossRef]
24. van Haarlem, M. LOFAR: The Low Frequency Array. *Eas Publ. Ser.* **2005**, *15*, 431–444. [CrossRef]
25. Chung, K.W.; Sutin, A.; Sedunov, A.; Bruno, M.S. DEMON Acoustic Ship Signature Measurements in an Urban Harbor. *Adv. Acoust. Vib.* **2011**, *2011*, 952798. [CrossRef]
26. Jin, G.; Liu, F.; Wu, H.; Song, Q. Deep learning-based framework for expansion, recognition and classification of underwater acoustic signal. *J. Exp. Theor. Artif. Intell.* **2019**, *32*, 205–218. [CrossRef]
27. Chen, J.; Han, B.; Ma, X.; Zhang, J. Underwater Target Recognition Based on Multi-Decision LOFAR Spectrum Enhancement: A Deep-Learning Approach. *Future Internet* **2021**, *13*, 265. [CrossRef]
28. Shi, Y.; Piao, S.; Guo, J. Line spectrum detection and motion parameters estimation for underwater moving target. *J. Phys. Conf. Ser.* **2024**, *2718*, 012090. [CrossRef]
29. Pollara, A.; Sutin, A.; Salloum, H. Improvement of the Detection of Envelope Modulation on Noise (DEMON) and its application to small boats. In Proceedings of the OCEANS 2016 MTS/IEEE, Monterey, CA, USA, 19–23 September 2016; pp. 1–10. [CrossRef]
30. Tong, W.; Wu, K.; Wang, H.; Cao, L.; Huang, B.; Wu, D.; Antoni, J. Adaptive Weighted Envelope Spectrum: A robust spectral quantity for passive acoustic detection of underwater propeller based on spectral coherence. *Mech. Syst. Signal Process.* **2024**, *212*, 111265. [CrossRef]
31. Li, L.; Song, S.; Feng, X. Combined LOFAR and DEMON Spectrums for Simultaneous Underwater Acoustic Object Counting and F0 Estimation. *J. Mar. Sci. Eng.* **2022**, *10*, 1565. [CrossRef]
32. Yan, J.; Sun, H.; Chen, H.; Junejo, N.U.R.; Cheng, E. Resonance-Based Time-Frequency Manifold for Feature Extraction of Ship-Radiated Noise. *Sensors* **2018**, *18*, 936. [CrossRef] [PubMed]
33. Cao, X.; Togneri, R.; Zhang, X.; Yu, Y. Convolutional Neural Network with Second-Order Pooling for Underwater Target Classification. *IEEE Sens. J.* **2019**, *19*, 3058–3066. [CrossRef]
34. Li, J.; Wang, B.; Cui, X.; Li, S.; Liu, J. Underwater Acoustic Target Recognition Based on Attention Residual Network. *Entropy* **2022**, *24*, 1657. [CrossRef]
35. Gabor, D. Theory of communication. *J. Inst. Electr. Eng. Part I Gen.* **1946**, *94*, 58. [CrossRef]
36. Ioup, J.W. Time-frequency analysis for acoustics education and for listening to whales in the Gulf of Mexico. *J. Acoust. Soc. Am.* **2013**, *134*, 4124. [CrossRef]
37. Luo, X.; Zhang, M.; Liu, T.; Huang, M.; Xu, X. An Underwater Acoustic Target Recognition Method Based on Spectrograms with Different Resolutions. *J. Mar. Sci. Eng.* **2021**, *9*, 1246. [CrossRef]
38. Zhang, Y.; Zeng, Q. MSLEFC: A low-frequency focused underwater acoustic signal classification and analysis system. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106333. [CrossRef]

39. Yang, M.; Li, X.; Yang, Y.; Meng, X. Characteristic analysis of underwater acoustic scattering echoes in the wavelet transform domain. *J. Mar. Sci. Appl.* **2017**, *16*, 93–101. [CrossRef]
40. Jing, L.; Zheng, T.; He, C.; Yin, H. Iterative adaptive frequency-domain equalization based on sliding window strategy over time-varying underwater acoustic channels. *JASA Express Lett.* **2021**, *1*, 076002. [CrossRef] [PubMed]
41. Morlet, J.; Arens, G.; Fourgeau, E.; Glard, D. Wave propagation and sampling theory—Part I: Complex signal and scattering in multilayered media. *Geophysics* **1982**, *47*, 203–221. [CrossRef]
42. Xin-xin, L.; Shi-e, Y.; Ming, Y. Feature extraction from underwater signals using wavelet packet transform. In Proceedings of the 2008 International Conference on Neural Networks and Signal Processing, Nanjing, China, 7–11 June 2008; pp. 400–405. [CrossRef]
43. Rademan, M.; Versfeld, D.; du Preez, J. Soft-output signal detection for cetacean vocalizations using spectral entropy, k-means clustering and the continuous wavelet transform. *Ecol. Inform.* **2023**, *74*, 101990. [CrossRef]
44. Han, Z.; Zhang, X.; Yan, B.; Qiao, L.; Wang, Z. The time-frequency analysis of the acoustic signal produced in underwater discharges based on Variational Mode Decomposition and Hilbert–Huang Transform. *Sci. Rep.* **2023**, *13*, 22. [CrossRef]
45. Choo, Y.S.; Byun, S.H.; Kim, S.M.; Lee, K. Target detection in pseudo Wigner-Ville distribution of underwater beamformed signals. *J. Acoust. Soc. Am.* **2019**, *146*, 2960. [CrossRef]
46. Wu, Y.; Li, X.; Wang, Y. Extraction and classification of acoustic scattering from underwater target based on Wigner-Ville distribution. *Appl. Acoust.* **2018**, *138*, 52–59. [CrossRef]
47. Wang, S.; Zeng, X. Robust underwater noise targets classification using auditory inspired time–frequency analysis. *Appl. Acoust.* **2014**, *78*, 68–76. [CrossRef]
48. Domingos, L.C.F.; Santos, P.E.; Skelton, P.S.M.; Brinkworth, R.S.A.; Sammut, K. An Investigation of Preprocessing Filters and Deep Learning Methods for Vessel Type Classification with Underwater Acoustic Data. *IEEE Access* **2022**, *10*, 117582–117596. [CrossRef]
49. Liu, Y.; Zhao, Y.; Gerstoft, P.; Zhou, F.; Qiao, G.; Yin, J. Deep transfer learning-based variable Doppler underwater acoustic communications. *J. Acoust. Soc. Am.* **2023**, *154*, 232–244. [CrossRef] [PubMed]
50. Tang, N.; Zhou, F.; Wang, Y.; Zhang, H.; Lyu, T.; Wang, Z.; Chang, L. Differential treatment for time and frequency dimensions in mel-spectrograms: An efficient 3D Spectrogram network for underwater acoustic target classification. *Ocean Eng.* **2023**, *287*, 115863. [CrossRef]
51. Abdul, Z.K.; Al-Talabani, A.K. Mel Frequency Cepstral Coefficient and its Applications: A Review. *IEEE Access* **2022**, *10*, 122136–122158. [CrossRef]
52. Zhang, Y.; Xu, K.; Wan, J. Rubost Feature for Underwater Targets Recognition Using Power-Normalized Cepstral Coefficients. In Proceedings of the 2018 14th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 12–16 August 2018; pp. 90–93. [CrossRef]
53. Tong, Y.; Zhang, X.; Ge, Y. Classification and Recognition of Underwater Target Based on MFCC Feature Extraction. In Proceedings of the 2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Macau, China, 21–24 August 2020; pp. 1–4. [CrossRef]
54. Ren, J.; Huang, Z.; Li, C.; Guo, X.; Xu, J. Feature Analysis of Passive Underwater Targets Recognition Based on Deep Neural Network. In Proceedings of the OCEANS 2019, Marseille, France, 17–20 June 2019; pp. 1–5. [CrossRef]
55. Hu, F.; Fan, J.; Kong, Y.; Zhang, L.; Guan, X.; Yu, Y. A Deep Learning Method for Ship-Radiated Noise Recognition Based on MFCC Feature. In Proceedings of the 2023 7th International Conference on Transportation Information and Safety (ICTIS), Xian, China, 4–6 August 2023; pp. 1328–1335. [CrossRef]
56. Lian, Z.; Xu, K.; Wan, J.; Li, G. Underwater acoustic target classification based on modified GFCC features. In Proceedings of the 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 25–26 March 2017; pp. 258–262. [CrossRef]
57. Cao, Y.; Yan, J.; Sun, K.; Luo, X. Hydroacoustic Target Detection Based on Improved GFCC and Lightweight Neural Network. In Proceedings of the 2023 42nd Chinese Control Conference (CCC), Tianjin, China, 24–26 July 2023; pp. 6239–6243. [CrossRef]
58. Liu, G.K. Evaluating Gammatone Frequency Cepstral Coefficients with Neural Networks for Emotion Recognition from Speech. *arXiv* **2018**, arXiv:1806.09010.
59. Song, K.; Wang, N.; Zhang, Y. An Improved Deep Canonical Correlation Fusion Method for Underwater Multisource Data. *IEEE Access* **2020**, *8*, 146300–146307. [CrossRef]
60. Hong, F.; Liu, C.; Guo, L.; Chen, F.; Feng, H. Underwater Acoustic Target Recognition with a Residual Network and the Optimized Feature Extraction Method. *Appl. Sci.* **2021**, *11*, 1442. [CrossRef]
61. Chen, Z.; Tang, J.; Qiu, H.; Chen, M. MGFGNet: An automatic underwater acoustic target recognition method based on the multi-gradient flow global feature enhancement network. *Front. Mar. Sci.* **2023**, *10*, 1306229. [CrossRef]
62. Tan, J.; Pan, X. Underwater acoustic target recognition based on convolutional neural network and multi-feature fusion. In Proceedings of the Third International Conference on Computer Vision and Pattern Analysis (ICCPA 2023), Hangzhou, China, 31 March–2 April 2023; Shen, L., Zhong, G., Eds.; International Society for Optics and Photonics, SPIE; Volume 12754, p. 1275432. [CrossRef]
63. Liu, F.; Shen, T.; Luo, Z.; Zhao, D.; Guo, S. Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation. *Appl. Acoust.* **2021**, *178*, 107989. [CrossRef]

64. Wu, J.; Li, P.; Wang, Y.; Lan, Q.; Xiao, W.; Wang, Z. VFR: The Underwater Acoustic Target Recognition Using Cross-Domain Pre-Training with FBank Fusion Features. *J. Mar. Sci. Eng.* **2023**, *11*, 263. [CrossRef]
65. Yang, Y.; Lv, H.; Chen, N. A Survey on ensemble learning under the era of deep learning. *Artif. Intell. Rev.* **2023**, *56*, 5545–5589. [CrossRef]
66. Zhang, Q.; Da, L.; Zhang, Y.; Hu, Y. Integrated neural networks based on feature fusion for underwater target recognition. *Appl. Acoust.* **2021**, *182*, 108261. [CrossRef]
67. Luo, X.; Feng, Y. An Underwater Acoustic Target Recognition Method Based on Restricted Boltzmann Machine. *Sensors* **2020**, *20*, 5399. [CrossRef]
68. Nie, L.; Zhang, Y.; Wang, H. Classification of underwater soundscapes using raw hydroacoustic signals. *J. Acoust. Soc. Am.* **2023**, *154*, A304. [CrossRef]
69. Chen, Y.; Shang, J. Underwater Target Recognition Method Based on Convolution Autoencoder. In Proceedings of the 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Chongqing, China, 11–13 December 2019; pp. 1–5. [CrossRef]
70. Khishe, M. DRW-AE: A Deep Recurrent-Wavelet Autoencoder for Underwater Target Recognition. *IEEE J. Ocean. Eng.* **2022**, *47*, 1083–1098. [CrossRef]
71. Berahmand, K.; Daneshfar, F.; Salehi, E.S.; Li, Y.; Xu, Y. Autoencoders and their applications in machine learning: A survey. *Artif. Intell. Rev.* **2024**, *57*, 28. [CrossRef]
72. Dong, Y.; Shen, X.; Wang, H. Bidirectional Denoising Autoencoders-Based Robust Representation Learning for Underwater Acoustic Target Signal Denoising. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–8. [CrossRef]
73. Li, J.; Yang, H.; Shen, S.; Xu, G. The Learned Multi-scale Deep Filters for Underwater Acoustic Target Modeling and Recognition. In Proceedings of the OCEANS 2019, Marseille, France, 17–20 June 2019; pp. 1–4. [CrossRef]
74. Luo, X.; Feng, Y.; Zhang, M. An Underwater Acoustic Target Recognition Method Based on Combined Feature with Automatic Coding and Reconstruction. *IEEE Access* **2021**, *9*, 63841–63854. [CrossRef]
75. Sun, B.; Luo, X. Underwater acoustic target recognition based on automatic feature and contrastive coding. *IET Radar Sonar Navig.* **2023**, *17*, 1277–1285. [CrossRef]
76. Wang, X.; Meng, J.; Liu, Y.; Zhan, G.; Tian, Z. Self-supervised acoustic representation learning via acoustic-embedding memory unit modified space autoencoder for underwater target recognition. *J. Acoust. Soc. Am.* **2022**, *152*, 2905–2915. [CrossRef] [PubMed]
77. Gomez, B.; Kadri, U. Earthquake source characterization by machine learning algorithms applied to acoustic signals. *Sci. Rep.* **2021**, *11*, 23062. [CrossRef] [PubMed]
78. Zelada Leon, A.; Huvenne, V.A.; Benoist, N.M.; Ferguson, M.; Bett, B.J.; Wynn, R.B. Assessing the Repeatability of Automated Seafloor Classification Algorithms, with Application in Marine Protected Area Monitoring. *Remote Sens.* **2020**, *12*, 1572. [CrossRef]
79. Harakawa, R.; Ogawa, T.; Haseyama, M.; Akamatsu, T. Automatic detection of fish sounds based on multi-stage classification including logistic regression via adaptive feature weighting. *J. Acoust. Soc. Am.* **2018**, *144*, 2709–2718. [CrossRef]
80. Xinhua, Z.; Zhenbo, L.; Chunyu, K. Underwater acoustic targets classification using support vector machine. In Proceedings of the International Conference on Neural Networks and Signal Processing, Nanjing, China, 14–17 December 2003; Volume 2, pp. 932–935. [CrossRef]
81. Alvaro, A.; Schwock, F.; Ragland, J.; Abadi, S. Ship detection from passive underwater acoustic recordings using machine learning. *J. Acoust. Soc. Am.* **2021**, *150*, A124. [CrossRef]
82. Li, H.; Cheng, Y.; Dai, W.; Li, Z. A method based on wavelet packets-fractal and SVM for underwater acoustic signals recognition. In Proceedings of the 2014 12th International Conference on Signal Processing (ICSP), HangZhou, China, 19–23 October 2014; pp. 2169–2173. [CrossRef]
83. de Moura, N.N.; de Seixas, J.M. Novelty detection in passive SONAR systems using support vector machines. In Proceedings of the 2015 Latin America Congress on Computational Intelligence (LA-CCI), Curitiba, Brazil, 13–16 October 2015; pp. 1–6. [CrossRef]
84. Sherin, B.M.; Supriya, M.H. Selection and parameter optimization of SVM kernel function for underwater target classification. In Proceedings of the 2015 IEEE Underwater Technology (UT), Chennai, India, 23–25 February 2015; pp. 1–5. [CrossRef]
85. Wang, B.; Wu, C.; Zhu, Y.; Zhang, M.; Li, H.; Zhang, W. Ship Radiated Noise Recognition Technology Based on ML-DS Decision Fusion. *Comput. Intell. Neurosci.* **2021**, *2021*, 8901565. [CrossRef]
86. Yao, Q.; Jiang, J.; Chen, G.; Li, Z.; Yao, Z.; Lu, Y.; Hou, X.; Fu, X.; Duan, F. Recognition method for underwater imitation whistle communication signals by slope distribution. *Appl. Acoust.* **2023**, *211*, 109531. [CrossRef]
87. Liu, F.; Li, G.; Yang, H. Application of multi-algorithm mixed feature extraction model in underwater acoustic signal. *Ocean Eng.* **2024**, *296*, 116959. [CrossRef]
88. Yang, H.; Gan, A.; Chen, H.; Pan, Y.; Tang, J.; Li, J. Underwater acoustic target recognition using SVM ensemble via weighted sample and feature selection. In Proceedings of the 2016 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 12–16 January 2016; pp. 522–527. [CrossRef]
89. Fisher, R.A. *Iris*; UCI Machine Learning Repository: Espoo, Finland, 1988. [CrossRef]
90. Choi, J.; Choo, Y.; Lee, K. Acoustic Classification of Surface and Underwater Vessels in the Ocean Using Supervised Machine Learning. *Sensors* **2019**, *19*, 3492. [CrossRef]

91. Wei, M.; Chen, K.; Lin, Y.; Cheng, E. Recognition of behavior state of Penaeus vannamei based on passive acoustic technology. *Front. Mar. Sci.* **2022**, *9*, 973284. [CrossRef]
92. Chen, H.; Sun, H.; Junejo, N.U.R.; Yang, G.; Qi, J. Whale Vocalization Classification Using Feature Extraction with Resonance Sparse Signal Decomposition and Ridge Extraction. *IEEE Access* **2019**, *7*, 136358–136368. [CrossRef]
93. Yaman, O.; Tuncer, T.; Tasar, B. DES-Pat: A novel DES pattern-based propeller recognition method using underwater acoustical sounds. *Appl. Acoust.* **2021**, *175*, 107859. [CrossRef]
94. Saffari, A.; Zahiri, S.H.; Khishe, M. Automatic recognition of sonar targets using feature selection in micro-Doppler signature. *Def. Technol.* **2023**, *20*, 58–71. [CrossRef]
95. Li, Y.X.; Jiao, S.B.; Geng, B.; Zhang, Q.; Zhang, Y.M. A comparative study of four nonlinear dynamic methods and their applications in classification of ship-radiated noise. *Def. Technol.* **2022**, *18*, 183–193. [CrossRef]
96. Jin, S.Y.; Su, Y.; Guo, C.J.; Fan, Y.X.; Tao, Z.Y. Offshore ship recognition based on center frequency projection of improved EMD and KNN algorithm. *Mech. Syst. Signal Process.* **2023**, *189*, 110076. [CrossRef]
97. Mohammed, S.K.; Hariharan, S.M.; Kamal, S. A GTCC-Based Underwater HMM Target Classifier with Fading Channel Compensation. *J. Sens.* **2018**, *2018*, 6593037:1–6593037:14. [CrossRef]
98. You, H.; Byun, S.H.; Choo, Y. Underwater Acoustic Signal Detection Using Calibrated Hidden Markov Model with Multiple Measurements. *Sensors* **2022**, *22*, 5088. [CrossRef]
99. Seo, Y.; On, B.; Im, S.; Shim, T.; Seo, I. Underwater Cylindrical Object Detection Using the Spectral Features of Active Sonar Signals with Logistic Regression Models. *Appl. Sci.* **2018**, *8*, 116. [CrossRef]
100. Yang, L.; Chen, K. Performance and strategy comparisons of human listeners and logistic regression in discriminating underwater targets. *J. Acoust. Soc. Am.* **2015**, *138*, 3138–3147. [CrossRef]
101. K, S.; R, K.; Kumar, P.S.; V, R.; Lakshmi, G. Rock/Mine Classification Using Supervised Machine Learning Algorithms. In Proceedings of the 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India, 27–28 January 2023; pp. 177–184. [CrossRef]
102. Yu, Q.; Zhang, W.; Zhu, M.; Shi, J.; Liu, Y.; Liu, S. Surface and Underwater Acoustic Source Recognition Using Array Feature Extraction Based on Machine Learning. *J. Phys. Conf. Ser.* **2024**, *2718*, 012100. [CrossRef]
103. Zhou, X.; Yang, K. A denoising representation framework for underwater acoustic signal recognition. *J. Acoust. Soc. Am.* **2020**, *147*, EL377–EL383. [CrossRef] [PubMed]
104. Wang, Q.; Wang, L.; Zeng, X.; Zhao, L. An Improved Deep Clustering Model for Underwater Acoustical Targets. *Neural Process. Lett.* **2018**, *48*, 1633–1644. [CrossRef]
105. Sabara, R.; Soares, C.; Zabel, F.; Oliveira, J.V.; Jesus, S.M. Automatic Acoustic Target Detection and Classification off the Coast of Portugal. In Proceedings of the Global Oceans 2020: Singapore—U.S. Gulf Coast, Biloxi, MS, USA, 5–30 October 2020; pp. 1–9. [CrossRef]
106. Yang, K.; Zhou, X. Unsupervised Classification of Hydrophone Signals with an Improved Mel-Frequency Cepstral Coefficient Based on Measured Data Analysis. *IEEE Access* **2019**, *7*, 124937–124947. [CrossRef]
107. Agersted, M.D.; Khodabandeloo, B.; Liu, Y.; Melle, W.; Klevjer, T.A. Application of an unsupervised clustering algorithm on in situ broadband acoustic data to identify different mesopelagic target types. *Ices J. Mar. Sci.* **2021**, *78*, 2907–2921. [CrossRef]
108. Luo, X.; Chen, L.; Zhou, H.; Cao, H. A Survey of Underwater Acoustic Target Recognition Methods Based on Machine Learning. *J. Mar. Sci. Eng.* **2023**, *11*, 384. [CrossRef]
109. Baran, R.H.; Coughlan, J.M. Neural network for passive acoustic discrimination between surface and submarine targets. In Proceedings of the Automatic Object Recognition, San Francisco, CA, USA, 1 August 1991; Sadjadi, F.A., Ed.; International Society for Optics and Photonics, SPIE; Volume 1471, pp. 164–176. [CrossRef]
110. Khotanzad.; Lu.; Srinath. Target detection using a neural network based passive sonar system. In Proceedings of the International 1989 Joint Conference on Neural Networks, Washington, DC, USA, 18–22 June 1989; Volume 1, pp. 335–340. [CrossRef]
111. Filho, W. Preprocessing passive sonar signals for neural classification. *IET Radar Sonar Navig.* **2011**, *5*, 605–612. [CrossRef]
112. Yue, H.; Zhang, L.; Wang, D.; Wang, Y.; Lu, Z. The Classification of Underwater Acoustic Targets Based on Deep Learning Methods. In Proceedings of the 2017 2nd International Conference on Control, Automation and Artificial Intelligence (CAAI 2017), Sanya, China, 25–26 June 2017; pp. 526–529. [CrossRef]
113. Zhao, J.; Wang, S.; Jia, X.; Gao, Y.; Zhu, W.; Ma, F.; Liu, Q. Underwater target perception algorithm based on pressure sequence generative adversarial network. *Ocean Eng.* **2023**, *286*, 115547. [CrossRef]
114. Li, D.; Liu, F.; Shen, T.; Chen, L.; Zhao, D. Data augmentation method for underwater acoustic target recognition based on underwater acoustic channel modeling and transfer learning. *Appl. Acoust.* **2023**, *208*, 109344. [CrossRef]
115. Yang, J.; Yan, S.; Zeng, D.; Tan, G. Self-supervised learning minimax entropy domain adaptation for the underwater target recognition. *Appl. Acoust.* **2024**, *216*, 109725. [CrossRef]
116. Wang, X.; Wu, P.; Li, B.; Zhan, G.; Liu, J.; Liu, Z. A self-supervised dual-channel self-attention acoustic encoder for underwater acoustic target recognition. *Ocean Eng.* **2024**, *299*, 117305. [CrossRef]
117. Xu, K.; Xu, Q.; You, K.; Zhu, B.; Feng, M.; Feng, D.; Liu, B. Self-supervised learning–based underwater acoustical signal classification via mask modeling. *J. Acoust. Soc. Am.* **2023**, *154*, 5–15. [CrossRef] [PubMed]
118. You, K.; Xu, K.; Feng, M.; Zhu, B. Underwater acoustic classification using masked modeling-based swin transformer. *J. Acoust. Soc. Am.* **2022**, *152*, A296. [CrossRef]

119. Zhou, A.; Li, X.; Zhang, W.; Zhao, C.; Ren, K.; Ma, Y.; Song, J. An attention-based multi-scale convolution network for intelligent underwater acoustic signal recognition. *Ocean Eng.* **2023**, *287*, 115784. [CrossRef]
120. Schoene, A.M.; Turner, A.P.; De Mel, G.; Dethlefs, N. Hierarchical Multiscale Recurrent Neural Networks for Detecting Suicide Notes. *IEEE Trans. Affect. Comput.* **2023**, *14*, 153–164. [CrossRef]
121. Bansal, A.; Garg, N.K. Robust technique for environmental sound classification using convolutional recurrent neural network. *Multimed. Tools Appl.* **2023**, *83*, 54755–54772. [CrossRef]
122. Wang, Y.; Wu, H.; Zhang, J.; Gao, Z.; Wang, J.; Yu, P.S.; Long, M. PredRNN: A Recurrent Neural Network for Spatiotemporal Predictive Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 2208–2225. [CrossRef]
123. Zhang, X.; Zhong, C.; Zhang, J.; Wang, T.; Ng, W.W. Robust recurrent neural networks for time series forecasting. *Neurocomputing* **2023**, *526*, 143–157. [CrossRef]
124. Hewamalage, H.; Bergmeir, C.; Bandara, K. Recurrent Neural Networks for Time Series Forecasting: Current status and future directions. *Int. J. Forecast.* **2021**, *37*, 388–427. [CrossRef]
125. Zhang, S.; Xing, S. Intelligent Recognition of Underwater Acoustic Target Noise by Multi-Feature Fusion. In Proceedings of the 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 8–9 December 2018; Volume 1, pp. 212–215. [CrossRef]
126. Zhang, S.; Wang, C.; Sun, Q. Underwater Target Noise Recognition and Classification Technology based on Multi-Classes Feature Fusion. *JNWPU* **2020**, *38*, 366–376. [CrossRef]
127. Yu, X.; Li, L.; Yin, J.; Shao, M.; Han, X. Modulation Pattern Recognition of Non-cooperative Underwater Acoustic Communication Signals Based on LSTM Network. In Proceedings of the 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Chongqing, China, 11–13 December 2019; pp. 1–5. [CrossRef]
128. Yang, H.; Xu, G.; Yi, S.; Li, Y. A New Cooperative Deep Learning Method for Underwater Acoustic Target Recognition. In Proceedings of the OCEANS 2019, Marseille, France, 17–20 June 2019; pp. 1–4. [CrossRef]
129. Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, Doha, Qatar, 25–29 October 2014; A meeting of SIGDAT, a Special Interest Group of the ACL; Moschitti, A., Pang, B., Daelemans, W., Eds.; ACL: Edinburgh, UK, 2014; pp. 1724–1734. [CrossRef]
130. Qi, P.; Yin, G.; Zhang, L. Underwater acoustic target recognition using RCRNN and wavelet-auditory feature. *Multimed. Tools Appl.* **2023**, *83*, 47295–47317. [CrossRef]
131. Wang, Y.; Zhang, H.; Xu, L.; Cao, C.; Gulliver, T.A. Adoption of hybrid time series neural network in the underwater acoustic signal modulation identification. *J. Frankl. Inst.* **2020**, *357*, 13906–13922. [CrossRef]
132. Liu, Y.; Li, X.; Yang, L.; Bian, G.; Yu, H. A CNN-Transformer Hybrid Recognition Approach for sEMG-Based Dynamic Gesture Prediction. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–16. [CrossRef]
133. Kamal, S.; Satheesh Chandran, C.; Supriya, M. Passive sonar automated target classifier for shallow waters using end-to-end learnable deep convolutional LSTMs. *Eng. Sci. Technol. Int. J.* **2021**, *24*, 860–871. [CrossRef]
134. Qi, P.; Sun, J.; Long, Y.; Zhang, L.; Tianye. Underwater Acoustic Target Recognition with Fusion Feature. In *Neural Information Processing, Proceedings of the 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, 8–12December 2021*; Proceedings, Part I 28; Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 609–620.
135. Ma, Q.; Jiang, L.; Yu, W. Lambertian-based adversarial attacks on deep-learning-based underwater side-scan sonar image classification. *Pattern Recognit.* **2023**, *138*, 109363. [CrossRef]
136. Steiniger, Y.; Kraus, D.; Meisen, T. Survey on deep learning based computer vision for sonar imagery. *Eng. Appl. Artif. Intell.* **2022**, *114*, 105157. [CrossRef]
137. Wang, Y.; Jin, Y.; Zhang, H.; Lu, Q.; Cao, C.; Sang, Z.; Sun, M. Underwater Communication Signal Recognition Using Sequence Convolutional Network. *IEEE Access* **2021**, *9*, 46886–46899. [CrossRef]
138. Xiaoping, S.; Jinsheng, C.; Yuan, G. A New Deep Learning Method for Underwater Target Recognition Based on One-Dimensional Time-Domain Signals. In Proceedings of the 2021 OES China Ocean Acoustics (COA), Harbin, China, 14–17 July 2021; pp. 1048–1051. [CrossRef]
139. Zhu, P.; Zhang, Y.; Huang, Y.; Zhao, C.; Zhao, K.; Zhou, F. Underwater acoustic target recognition based on spectrum component analysis of ship radiated noise. *Appl. Acoust.* **2023**, *211*, 109552. [CrossRef]
140. Dong, Y.; Shen, X.; Jiang, Z.; Wang, H. Recognition of imbalanced underwater acoustic datasets with exponentially weighted cross-entropy loss. *Appl. Acoust.* **2021**, *174*, 107740. [CrossRef]
141. Feng, S.; Zhu, X. A Transformer-Based Deep Learning Network for Underwater Acoustic Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
142. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
143. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]

144. Yao, Y.; Zeng, X.; Wang, H.; Liu, J. Research on Underwater Acoustic Target Recognition Method Based on DenseNet. In Proceedings of the 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Xi'an, China, 15–17 July 2022; pp. 114–118. [CrossRef]
145. Hong, F.; Liu, C.; Guo, L.; Chen, F.; Feng, H. Underwater Acoustic Target Recognition with ResNet18 on ShipsEar Dataset. In Proceedings of the 2021 IEEE 4th International Conference on Electronics Technology (ICET), Chengdu, China, 7–10 May 2021; pp. 1240–1244. [CrossRef]
146. Sun, Q.; Wang, K. Underwater single-channel acoustic signal multitarget recognition using convolutional neural networks. *J. Acoust. Soc. Am.* **2022**, *151*, 2245–2254. [CrossRef]
147. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
148. Mehta, S.; Rastegari, M.; Shapiro, L.; Hajishirzi, H. ESPNetv2: A Light-Weight, Power Efficient, and General Purpose Convolutional Neural Network. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9182–9192. [CrossRef]
149. Hu, G.; Wang, K.; Liu, L. Underwater Acoustic Target Recognition Based on Depthwise Separable Convolution Neural Networks. *Sensors* **2021**, *21*, 1429. [CrossRef] [PubMed]
150. Miao, Y.; Zakharov, Y.V.; Sun, H.; Li, J.; Wang, J. Underwater Acoustic Signal Classification Based on Sparse Time–Frequency Representation and Deep Learning. *IEEE J. Ocean. Eng.* **2021**, *46*, 952–962. [CrossRef]
151. Zheng, Y.; Gong, Q.; Zhang, S. Time-Frequency Feature-Based Underwater Target Detection with Deep Neural Network in Shallow Sea. *J. Phys. Conf. Ser.* **2021**, *1756*, 012006. [CrossRef]
152. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
153. Cai, W.; Zhu, J.; Zhang, M.; Yang, Y. A Parallel Classification Model for Marine Mammal Sounds Based on Multi-Dimensional Feature Extraction and Data Augmentation. *Sensors* **2022**, *22*, 7443. [CrossRef]
154. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:abs/1704.04861.
155. Yan, C.; Yu, Y.; Yan, S.; Yao, T.; Yang, C.; Liu, L.; Pan, G. Underwater target recognition using a lightweight asymmetric convolutional neural network. In Proceedings of the 17th International Conference on Underwater Networks & Systems, WUWNet '23, Shenzhen, China, 24–26 November 2023; Association for Computing Machinery: New York, NY, USA, 2024. [CrossRef]
156. Tian, G.; Haiyang, Y.; Haiyan, W.; Fan, W.; Xiao, C. CA_MobileNetV2 for Underwater Acoustic Target Recognition. In Proceedings of the 2023 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Zhengzhou, China, 14–17 November 2023; pp. 1–5. [CrossRef]
157. Jiang, Z.; Zhao, C.; Wang, H. Classification of Underwater Target Based on S-ResNet and Modified DCGAN Models. *Sensors* **2022**, *22*, 2293. [CrossRef]
158. Xiao, X.; Wang, W.; Ren, Q.; Gerstoft, P.; Ma, L. Underwater acoustic target recognition using attention-based deep neural network. *JASA Express Lett.* **2021**, *1*, 106001. [CrossRef]
159. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
160. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An Empirical Study of Spatial Attention Mechanisms in Deep Networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6687–6696. [CrossRef]
161. Wang, B.; Zhang, W.; Zhu, Y.; Wu, C.; Zhang, S. An Underwater Acoustic Target Recognition Method Based on AMNet. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
162. Zhu, M.; Zhang, X.; Jiang, Y.; Wang, K.; Su, B.; Wang, T. Hybrid Underwater Acoustic Signal Multi-Target Recognition Based on DenseNet-LSTM with Attention Mechanism. In Proceedings of the 2023 Chinese Intelligent Automation Conference, Nanjing, China, 2–5 October 2023; Deng, Z., Ed.; Springer Nature: Singapore, 2023; pp. 728–738.
163. Seo, M.J.; Kembhavi, A.; Farhadi, A.; Hajishirzi, H. Bidirectional Attention Flow for Machine Comprehension. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
164. Liu, C.; Hong, F.; Feng, H.; Hu, M. Underwater Acoustic Target Recognition Based on Dual Attention Networks and Multiresolution Convolutional Neural Networks. In Proceedings of the OCEANS 2021: San Diego—Porto, San Diego, CA, USA, 20–23 September 2021; pp. 1–5. [CrossRef]
165. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. *arXiv* **2017**, arXiv:1706.03762v7.
166. Gong, Y.; Chung, Y.A.; Glass, J. AST: Audio Spectrogram Transformer. In Proceedings of the Proc. Interspeech 2021, Brno, Czechia, 30 August–3 September 2021; pp. 571–575. [CrossRef]
167. Li, P.; Wu, J.; Wang, Y.; Lan, Q.; Xiao, W. STM: Spectrogram Transformer Model for Underwater Acoustic Target Recognition. *J. Mar. Sci. Eng.* **2022**, *10*, 1428. [CrossRef]

168. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [CrossRef]
169. Wu, F.; Yao, H.; Wang, H. Recognizing the State of Motion by Ship-Radiated Noise Using Time-Frequency Swin-Transformer. *IEEE J. Ocean. Eng.* **2024**, *49*, 1–12. [CrossRef]
170. Yao, H.; Gao, T.; Wang, Y.; Wang, H.; Chen, X. Mobile_ViT: Underwater Acoustic Target Recognition Method Based on Local–Global Feature Fusion. *J. Mar. Sci. Eng.* **2024**, *12*, 589. [CrossRef]
171. Feng, S.; Zhu, X.; Ma, S. Masking Hierarchical Tokens for Underwater Acoustic Target Recognition with Self-Supervised Learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 1365–1379. [CrossRef]
172. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In Proceedings of the Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019; Kubin, G., Kacic, Z., Eds.; ISCA: Copenhagen, Denmark, 2019; pp. 2613–2617. [CrossRef]
173. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge Distillation: A Survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
174. Zhou, A.; Li, X.; Zhang, W.; Li, D.; Deng, K.; Ren, K.; Song, J. A Novel Cross-Attention Fusion-Based Joint Training Framework for Robust Underwater Acoustic Signal Recognition. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
175. Li, J.; Yang, H. Deep learning method with auditory passive attention for underwater acoustic target recognition under the condition of ship interference. *Ocean Eng.* **2024**, *302*, 117674. [CrossRef]
176. Song, Y.; Liu, F.; Shen, T. A novel noise reduction technique for underwater acoustic signals based on dual-path recurrent neural network. *IET Commun.* **2023**, *17*, 135–144. [CrossRef]
177. Yang, H.; Li, L.; Li, G. A New Denoising Method for Underwater Acoustic Signal. *IEEE Access* **2020**, *8*, 201874–201888. [CrossRef]
178. Yang, H.; Liu, M.; Zhang, S.; Zheng, R.; Dong, S. Few-shot Underwater Acoustic Target Recognition Based on Siamese Network. In Proceedings of the 2023 42nd Chinese Control Conference (CCC), Tianjin, China, 24–26 July 2023; pp. 8252–8257. [CrossRef]
179. Cui, X.; He, Z.; Xue, Y.; Tang, K.; Zhu, P.; Han, J. Cross-Domain Contrastive Learning-Based Few-Shot Underwater Acoustic Target Recognition. *J. Mar. Sci. Eng.* **2024**, *12*, 264. [CrossRef]
180. Tian, S.; Bai, D.; Zhou, J.; Fu, Y.; Chen, D. Few-shot learning for joint model in underwater acoustic target recognition. *Sci. Rep.* **2023**, *13*, 17502. [CrossRef]
181. Zheng, S.; Mai, S.; Sun, Y.; Hu, H.; Yang, Y. Subgraph-Aware Few-Shot Inductive Link Prediction Via Meta-Learning. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 6512–6517. [CrossRef]
182. Ma, R.; Li, S.; Zhang, B.; Fang, L.; Li, Z. Flexible and Generalized Real Photograph Denoising Exploiting Dual Meta Attention. *IEEE Trans. Cybern.* **2023**, *53*, 6395–6407. [CrossRef] [PubMed]
183. Liang, Y.; Li, S.; Yan, C.; Li, M.; Jiang, C. Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing* **2021**, *419*, 168–182. [CrossRef]
184. Gao, Y.; Mosalam, K.M. Deep learning visual interpretation of structural damage images. *J. Build. Eng.* **2022**, *60*, 105144. [CrossRef]
185. Agrawal, P.; Ganapathy, S. Interpretable Representation Learning for Speech and Audio Signals Based on Relevance Weighting. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2823–2836. [CrossRef]
186. Montavon, G.; Samek, W.; Müller, K.R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **2018**, *73*, 1–15. [CrossRef]
187. Xie, Y.; Ren, J.; Xu, J. Guiding the underwater acoustic target recognition with interpretable contrastive learning. In Proceedings of the OCEANS 2023, Limerick, Ireland, 5–8 June 2023; pp. 1–6. [CrossRef]
188. Chen, Y.; Du, S.; Quan, H. Feature Analysis and Optimization of Underwater Target Radiated Noise Based on t-SNE. In Proceedings of the 2018 10th International Conference on Wireless Communications and Signal Processing (WCSP), Hangzhou, China, 18–20 October 2018; pp. 1–5. [CrossRef]
189. Xu, Y.; Kong, X.; Cai, Z. Cross-validation strategy for performance evaluation of machine learning algorithms in underwater acoustic target recognition. *Ocean Eng.* **2024**, *299*, 117236. [CrossRef]
190. Han, K.; Wang, Y.; Guo, J.; Tang, Y.; Wu, E. Vision GNN: An Image is Worth Graph of Nodes. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 8291–8303.
191. Xu, J.; Xie, Y.; Wang, W. Underwater acoustic target recognition based on smoothness-inducing regularization and spectrogram-based data augmentation. *Ocean Eng.* **2023**, *281*, 114926. [CrossRef]
192. Li, D.; Liu, F.; Shen, T.; Chen, L.; Yang, X.; Zhao, D. Generalizable Underwater Acoustic Target Recognition Using Feature Extraction Module of Neural Network. *Appl. Sci.* **2022**, *12*, 10804. [CrossRef]
193. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual Conference, 6–12 December 2020; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.

194. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning. JMLR.org, Honolulu, HI, USA, 23–29 July 2023. ICML'23.
195. Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **2023**, *619*, 533–538. [CrossRef]
196. Wang, X.; Wang, R.; Hu, N.; Wang, P.; Huo, P.; Wang, G.; Wang, H.; Wang, S.; Zhu, J.; Xu, J.; et al. XiHe: A Data-Driven Model for Global Ocean Eddy-Resolving Forecasting. *arXiv* **2024**, arXiv:2402.02995.
197. Sun, W.; Yan, R.; Jin, R.; Zhao, R.; Chen, Z. FedAlign: Federated Model Alignment via Data-Free Knowledge Distillation for Machine Fault Diagnosis. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–12. [CrossRef]
198. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
199. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. In Proceedings of the Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, 10–16 December 2023; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.
200. Lu, J.; Jin, F.; Zhang, J. Adapter Tuning with Task-Aware Attention Mechanism. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, 4–10 June 2023; IEEE: New York, NY, USA, 2023; pp. 1–5. [CrossRef]
201. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015
202. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
203. Liu, X.; Xie, L.; Wang, Y.; Zou, J.; Xiong, J.; Ying, Z.; Vasilakos, A.V. Privacy and security issues in deep learning: A survey. *IEEE Access* **2020**, *9*, 4566–4593. [CrossRef]
204. Modas, A.; Sanchez-Matilla, R.; Frossard, P.; Cavallaro, A. Toward Robust Sensing for Autonomous Vehicles: An Adversarial Perspective. *IEEE Signal Process. Mag.* **2020**, *37*, 14–23. [CrossRef]
205. Svenmarck, P.; Luotsinen, L.; Nilsson, M.; Schubert, J. Possibilities and challenges for artificial intelligence in military applications. In Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting, Bordeaux, France, 30 May–1 June 2018; pp. 1–16.
206. Feng, S.; Zhu, X.; Ma, S.; Lan, Q. Adversarial Attacks in Underwater Acoustic Target Recognition with Deep Learning Models. *Remote Sens.* **2023**, *15*, 5386. [CrossRef]
207. Bai, T.; Luo, J.; Zhao, J.; Wen, B.; Wang, Q. Recent Advances in Adversarial Training for Adversarial Robustness. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21; International Joint Conferences on Artificial Intelligence Organization, Montreal, QC, Canada, 19–27 August 2021; pp. 4312–4321. [CrossRef]
208. Aldahdooh, A.; Hamidouche, W.; Fezza, S.A.; Déforges, O. Adversarial example detection for DNN models: A review and experimental comparison. *Artif. Intell. Rev.* **2022**, *55*, 4403–4462. [CrossRef]
209. Çatak, F.Ö.; Kuzlu, M.; Çatak, E.; Cali, U.; Guler, O. Defensive Distillation-Based Adversarial Attack Mitigation Method for Channel Estimation Using Deep Learning Models in Next-Generation Wireless Networks. *IEEE Access* **2022**, *10*, 98191–98203. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

Target Motion Parameters Estimation by Full-Plane Hyperbola-Warping Transform with a Single Hydrophone

Yuzheng Li, Bo Gao *, Zhuo Chen, Yueqi Yu, Zhennan Wang and Dazhi Gao

Department of Marine Technology, Ocean University of China, Qingdao 266100, China; liyuzheng@stu.ouc.edu.cn (Y.L.); chenzhuo3286@stu.ouc.edu.cn (Z.C.); yyq9011@stu.ouc.edu.cn (Y.Y.); wangzhennan@stu.ouc.edu.cn (Z.W.); dzgao@ouc.edu.cn (D.G.)

* Correspondence: gaobo@ouc.edu.cn

Abstract: In this paper, to counteract the sensitivity of the traditional Hough transform to noise and the fluctuations in parameter estimation, we propose a hyperbolic warping transform that integrates all interference fringes in the time–frequency domain to accurately estimate the motion parameters of a single hydrophone. This method can accurately estimate the target motion parameters, including the time of closest point of approach (t_{CPA}), the ratio of the nearest distance to the speed ($b = r_{CPA}/v$), and the waveguide invariant (β). The two algorithms are compared by simulation and sea trial experiments. Hyperbola-warping improves the noise immunity performance by 10 dB in simulation experiments, increases the detection range by 20% in sea trial experiments, and demonstrates that the method proposed in this paper has better noise resistance and practicality.

Keywords: single hydrophone; shallow-water acoustic field interference structure; hyperbola-warping transform; target motion parameter estimation

Citation: Li, Y.; Gao, B.; Chen, Z.; Yu, Y.; Wang, Z.; Gao, D. Target Motion Parameters Estimation by Full-Plane Hyperbola-Warping Transform with a Single Hydrophone. *Remote Sens.* **2024**, *16*, 3307. https://doi.org/10.3390/rs16173307

Academic Editor: Marc Pinto

Received: 31 July 2024
Revised: 31 August 2024
Accepted: 2 September 2024
Published: 5 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

When an acoustic signal propagates in the shallow-water waveguide, each normal mode interferes with the other to form a stable interference structure. The interference structure contains a large amount of target information and marine environment information. Therefore, how to use the interference structure to obtain target parameters [1,2] and marine environment inversion [3,4] has become an important research content in the field of underwater acoustics in recent years. Li et al. [5] proposed an extraction and phase compensation method based on the time–frequency spectrogram, and the acoustic data obtained can be used for source ranging. Song et al. [6], observing a sudden change in the waveguide invariant from the time–frequency spectrum data in sea trial experiments, proposes a new model that accounts for the source's movement in a gradually varying range and azimuth within a non-radial waveguide. Song et al. [7] presented a method for estimating the range of a distant ship in shallow water, utilizing a vertical array and a guide ship at a known range near the array, by integrating four distinct approaches: blind deconvolution, waveguide invariant, virtual receiver (VR), and array invariant.

Compared with the three sub-array method [8–10], target motion analysis (TMA) [11,12], matching field [13] and other methods [14,15], single hydrophone target parameter estimation does not require a large hydrophone array and a large number of sound field calculations, so it has become a research hotspot for scholars at home and abroad.

Current common methods for underwater target parameter estimation using a single hydrophone primarily rely on array invariance, waveguide invariance, and warping transforms. The array-invariant method leverages the multipath and dispersion properties of the waveguide, utilizing a guided sound source to passively estimate sound source parameters by integrating beamformed output sound intensity with array invariants. Wang et al. [16] has extended the generalized array invariant to deep-sea environments where the waveguide invariant varies with distance and has validated the feasibility of deep-sea source

distance estimation based on the array invariant with experimental data. Gihoon et al. [17] has expanded the array invariant to fully incorporate the angle dependence of the waveguide invariant, developing what is referred to as the adaptive array invariant, which, in theory, enables the provision of a perfect range estimate without the need to constrain the propagation angle. Wang et al. [18] have proposed a study that significantly improves localization accuracy by employing two-dimensional (2D) deconvolution, which achieves superior beam-time migration over conventional plane-wave beamforming, with simulations demonstrating a substantial reduction in range estimation error for the array-invariant method based on 2D deconvolution compared to existing techniques. This method has fewer requirements on environmental parameters and simple calculations. However, the array-invariant method has a large error for the estimation of long-distance sound source parameters [19].

In our context, the signal is recorded on a single hydrophone as a function of time. When the target approaches close to and then recedes from the hydrophone, a hyperbolic interference fringe structure is generated in the time–frequency domain. Therefore, the waveguide invariant method exploits the intrinsic relationship between sound field distance–frequency and interference structure characteristics in the shallow-water waveguide for passive target estimation. The core approach involves deriving the sound intensity interference fringe equation based on shallow-water sound field interference theory and employing image processing techniques, like the Hough transform, to estimate motion parameters. The Hough transform is often used to estimate the motion parameters of a target, including the time of closest point of approach (t_{CPA}), the ratio of the nearest distance to the speed (r_{CPA}/v) and other parameters [20,21]. Zhao et al. [22] proposed a passive localization method for vertically moving targets based on interference patterns detected by a single hydrophone, which mainly includes deriving target trajectory equations, designing a homomorphic filter, and using the Hough transform to extract fringe information. However, the Hough transform can only deal with a single stripe. When the target signal is affected by noise or the continuity of the interference pattern is disrupted, the estimation results can be significantly impacted.

Warping transform is essentially to resample the signal in the time domain or frequency domain according to the dispersion characteristics of normal waves and transform the complex non-stationary acoustic propagation signal into a quasi-single-frequency signal with a specific frequency or an instantaneous pulse signal with a specific time delay [23,24], so as to realize the application of passive sound source detection [25], geoacoustic parameter inversion [26] and bubble pulsation elimination [27]. Mao et al. [28] has broadened the application of the Doppler-warping transform, which was traditionally only applicable to sound sources moving in uniform linear motion, to now include a wider array of sound source trajectories. Gao et al. [27] proposed the Doppler-warping transformation to linearize the phase of the Doppler signal and, using this method, deduces the warping operator to devise an algorithm for estimating target motion parameters. Li et al. [29] refined the Warping transform formula to better adapt it for the separation and extraction of normal modes in broadband pulse sound pressure signals within non-ideal waveguides. Wang et al. [30] proposed a passive ranging method for impulsive sound sources in shallow-water environments, which leverages the warping transformation of the received signal's energy density function. However, at this stage, the signal studied by the warping transform is mainly a pulse signal. If the signal is not transient, such as the broadband continuous noise of the ship, then the method cannot be applied.

In contrast to conventional warping transform methods, this paper introduces a novel approach that applies the warping transform to all interference fringes in the time–frequency domain. This is carried out in accordance with the motion model of the target to address broadband continuous noise, such as that produced by ships. The proposed algorithm is a single hydrophone target motion parameter estimation method based on the full-plane Hyperbola-warping transform. The method presented in this paper utilizes all the information from the interference plane, whereas the Hough transform

method only employs information from a single interference bright fringe on the interference plane. It is capable of estimating the waveguide invariant (β), the time of the target's closest point of approach (t_{CPA}), and the ratio of the closest point distance to the target's speed ($b = {}^{r_{CPA}}/_{v}$). Compared to the traditional Hough transform, the proposed algorithm demonstrates superior performance in processing both simulation and sea trial data.

The letter is organized as follows: the parameter searching algorithm is introduced in Section 2. To validate the proposed method, a simulation is given in Section 3, and the results of the sea trial are shown in Section 4. Finally, a conclusion is given in Section 5.

2. Motion Parameter Estimation

2.1. Theoretical Analysis

In a range-independent waveguide, consider a target sound source positioned at a depth z_S with an amplitude $A(\omega)$, where ω represents the angular frequency. This source moves linearly at a constant velocity v from point S_A to point S_B. A hydrophone is situated at point S_C, with a depth of z, and is separated from the sound source by a distance r. As the target navigates toward point S_D, it reaches the closest proximity to the hydrophone. The time at which this minimum distance occurs is defined as the nearest point moment, denoted by t_{CPA}. Correspondingly, the minimum distance itself is referred to as the nearest point distance, denoted by r_{CPA}. A visual representation of this motion model is provided in Figure 1.

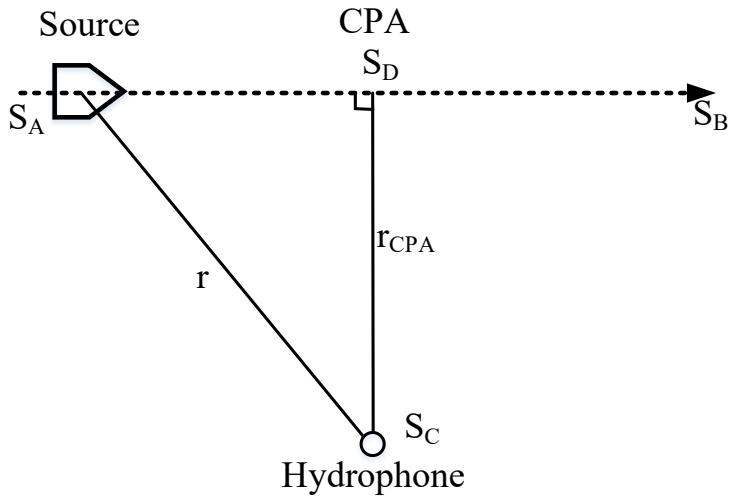


Figure 1. Geometric model of target navigation.

As per reference [31], the sound pressure received by the hydrophone at a distance from the source is given by:

$$p(r,\omega) = A(\omega)\sum_{m=1}^{M} B_m(z_s,z)\exp[-i(\omega t - k_m r + \pi/4)] \tag{1}$$

$$B_m(z_s,z) = \left(\frac{2\pi}{k_m r}\right)^{1/2}\Psi_m(z_s)\Psi_m(z) \tag{2}$$

where $\Psi_m(z)$ represents the mode depth function of the mth mode, k_m is the horizontal wave number of the mth mode, m is the number of modes, and the sound intensity $I(r,\omega)$ received by the hydrophone can be expressed as:

$$I(r,\omega) \equiv \langle pp^* \rangle = A^2(\omega)\left(\sum_m B_m^2 + 2\sum_{\substack{n,m \\ n\neq m}} B_n B_m \cos(\Delta k_{mn}(\omega)r)\right) \tag{3}$$

where p^* denotes the complex conjugate of the sound pressure, and Δk_{mn} represents the horizontal wave number difference between the m-th and n-th modes. The first term $\sum_m B_m^2$ in Equation (3) is non-coherent and slowly varying, while the second term $\sum_{\substack{n,m \\ n \neq m}} B_n B_m \cos(\Delta k_{mn}(\omega) r)$ is coherent, which will cause bright and dark stripes in the distance–frequency diagram.

Chuprov et al. [32] initially defined the concept of waveguide invariant β based on the slope $\partial\omega/\partial r$ of the interference fringe, with the definition given by:

$$\beta = \frac{r}{\omega}\frac{\partial \omega}{\partial r} \tag{4}$$

When the target moves at a constant speed, the radial velocity between the target and the hydrophone changes, and the interference structure is shown in Figure 2.

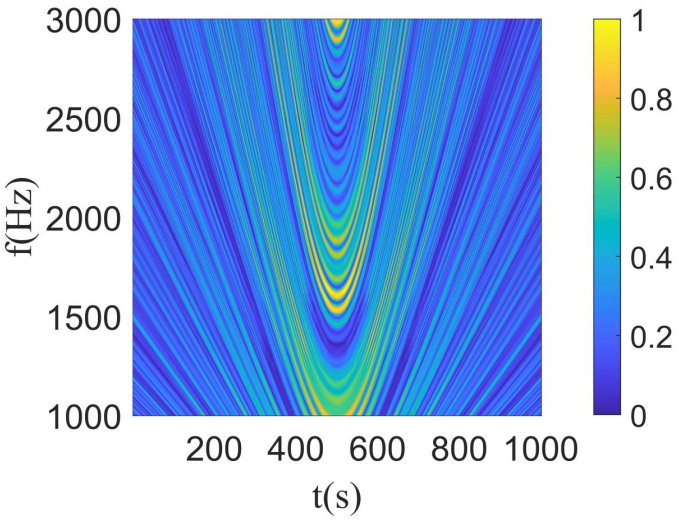


Figure 2. Time–frequency plane interference fringes of sound intensity.

According to the trajectory shown in Figure 1, the distance between the target and the hydrophone does not change linearly, so the interference fringes of the simulated time–frequency spectrum (Figure 2) show a hyperbolic shape. According to the trajectory of the target, the distance r between the target and the hydrophone at time t can be expressed as:

$$r = \sqrt{r_{CPA}^2 + v^2(t - t_{CPA})^2} \tag{5}$$

According to the principle of the waveguide invariant, we substitute Equation (5) into Equation (4) to obtain:

$$\frac{d\omega}{\omega} = \beta\frac{dr}{r} = \frac{\beta d(\sqrt{r_{CPA}^2 + v^2(t - t_{CPA})^2})}{\sqrt{r_{CPA}^2 + v^2(t - t_{CPA})^2}} \tag{6}$$

In Equation (6), we perform a simultaneous integration with respect to both $d\omega/\omega$ and $\frac{\beta d(\sqrt{r_{CPA}^2+v^2(t-t_{CPA})^2})}{\sqrt{r_{CPA}^2+v^2(t-t_{CPA})^2}}$, from which Equation (7) can be derived.

$$\frac{\omega}{\omega_0} = [1 + (\frac{v}{r_{CPA}})^2(t - t_{CPA})^2]^{\frac{\beta}{2}} \tag{7}$$

where ω_0 represents the frequency at $t = t_{CPA}$ on an interference fringe.

Therefore, for the same interference fringe, the relationship between time and frequency can be described by Equation (7), or for a known frequency range and time window range, the sound intensity can be transformed as follows:

$$I_R(f,t) = I(f^*,t) = I(f[1+\frac{(t-t_{CPA})^2}{b^2}]^{\frac{\beta}{2}},t) \tag{8}$$

Then, we define the Hyperbola-warping operator as:

$$h_w = f^* = f[1+\frac{(t-t_{CPA})^2}{b^2}]^{\frac{\beta}{2}} \tag{9}$$

Finally, through the transformation of Equation (8), it is equivalent to extracting the sound intensity $I(f, t_{CPA})$ located on the same interference fringe before the transformation at t time, which is expressed as $I(f^*,t_{CPA})$. Therefore, in the interference image I_R after the warping transformation of the sound intensity spectrum by the Hyperbola-warping operator in Equation (9), the fluctuation structure of the sound intensity at each time on the frequency is the same as that of the sound intensity at a time t_{CPA} on the frequency.

To clarify the explanation of the algorithm and provide a comprehensive understanding, we have integrated a simulation example. The simulation is conducted in a simplified shallow-water environment with horizontal waveguide invariance, as depicted in Figure 3a. A target at a depth z of 2 m travels at a speed of 5 m/s along the path illustrated in Figure 1. The hydrophone, positioned at a depth z_S of 7 m, remains stationary. The closest approach time t_{CPA} between the hydrophone and the target is 500 s, with the nearest distance r being 1000 m, as represented in Figure 3b. The sound pressure field is computed using the normal mode model Kraken [33], as visualized in Figure 3c.

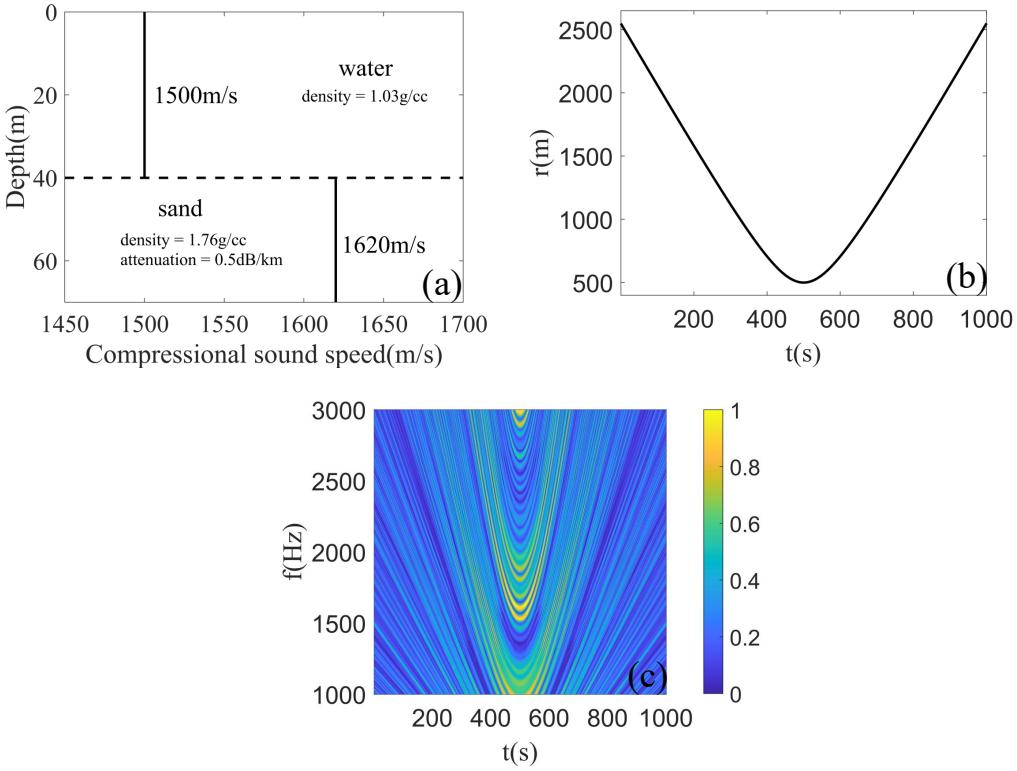


Figure 3. (**a**) Simulation parameters. (**b**) Distance between the hydrophone and the target. (**c**) Time–frequency spectrum of the signal received by the hydrophone.

According to the algorithm, when the target parameters b and t_{CPA} obtain the correct values in the interference image I_R after the Hyperbola-warping transformation, the

fluctuation structure of the sound intensity at each moment is the same as the fluctuation structure of the sound intensity at the t_{CPA} moment. When $b = r_{CPA}/v = 500/5 = 100$ s, $t_{CPA} = 500$ s and $\beta = 1$, the interference image will be 'straightened' after the transformation of Equation (8).

To further clarify the algorithm, we select the data from 400 s to 600 s in Figure 2, as depicted in Figure 4a. When the parameter values are correct ($b = 100$ s, $t_{CPA} = 500$ s, $\beta = 1$), the transformed time–frequency spectrum is shown in Figure 4b. Each bright (dark) fringe in the figure is linear, which is completely consistent with the fluctuation structure at t_{CPA} time. When the parameter value is incorrect (for example, $b = 125$ s, $t_{CPA} = 470$ s, $\beta = 0.95$), the transformed time–frequency spectrum is shown in Figure 4c. The bright (dark) stripes in the figure show different bending shapes, which are quite different from the undulating structure at t_{CPA} time.

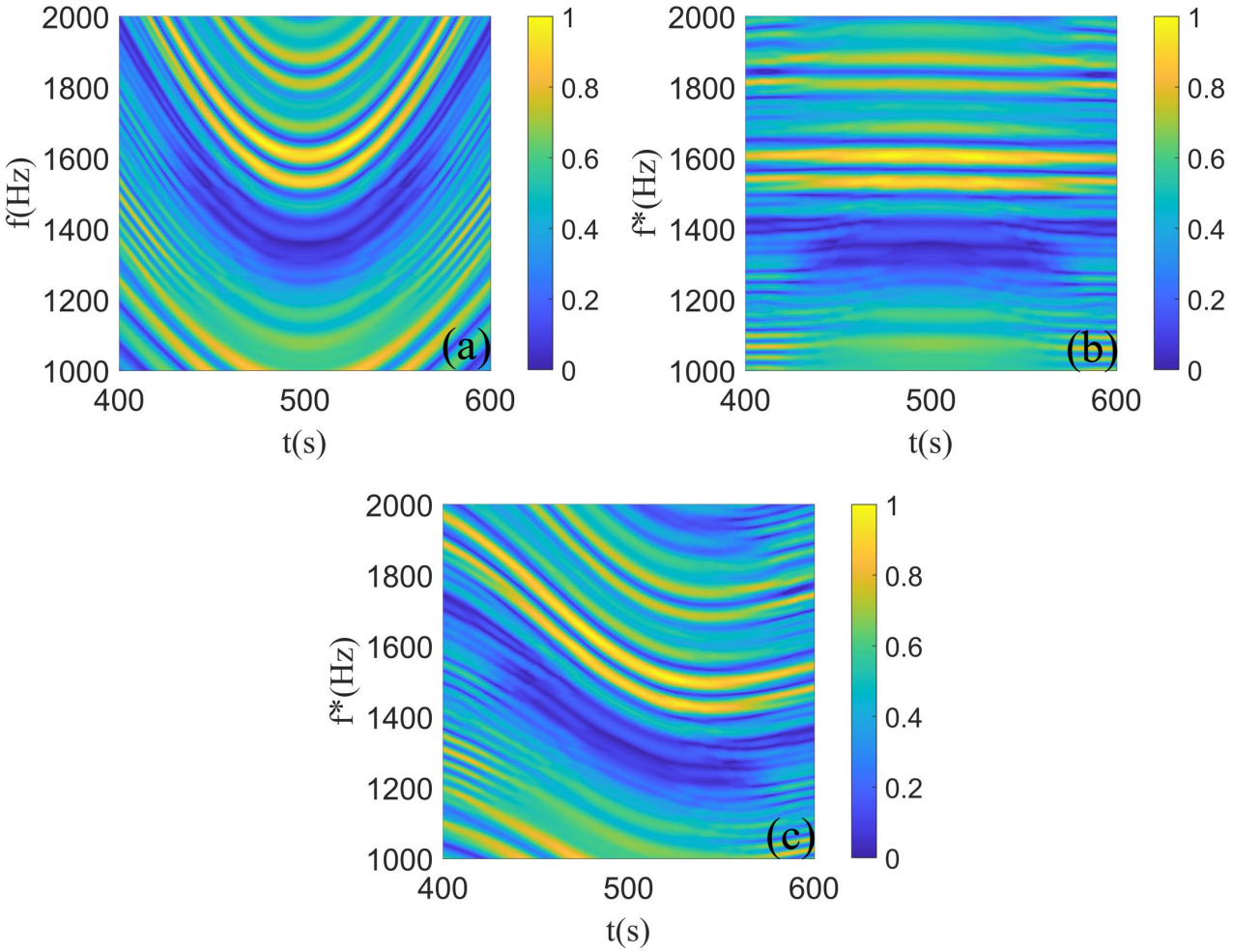


Figure 4. (**a**) The selected part of the time–frequency spectrogram. (**b**) The resampling results of the time–frequency spectrogram using the correct motion parameters. (**c**) The resampling results of the time–frequency spectrogram using the wrong motion parameters.

To further compare the effect of the transformation, we accumulate the time–frequency spectrum along the time axis according to Equation (10) (considering the influence of the entire frequency band on the result, greatly improving the anti-noise ability), and obtain the average value of the frequency band signal in the time period ($\bar{I}_R(f, b_m, t_{CPAn}, \beta_q)$).

$$\bar{I}_R(f, b_m, t_{CPAn}, \beta) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} I_R(f, t, b_m, t_{CPAn}, \beta_q) dt \tag{10}$$

Then the $\bar{I}_R(f, b_m, t_{CPAn}, \beta_q)$ is compared with the sound intensity change ($\bar{I}_R(f, b_m, t_{CPA}, \beta_q)$) the moment of t_{CPA}, and the results are shown in Figure 5.

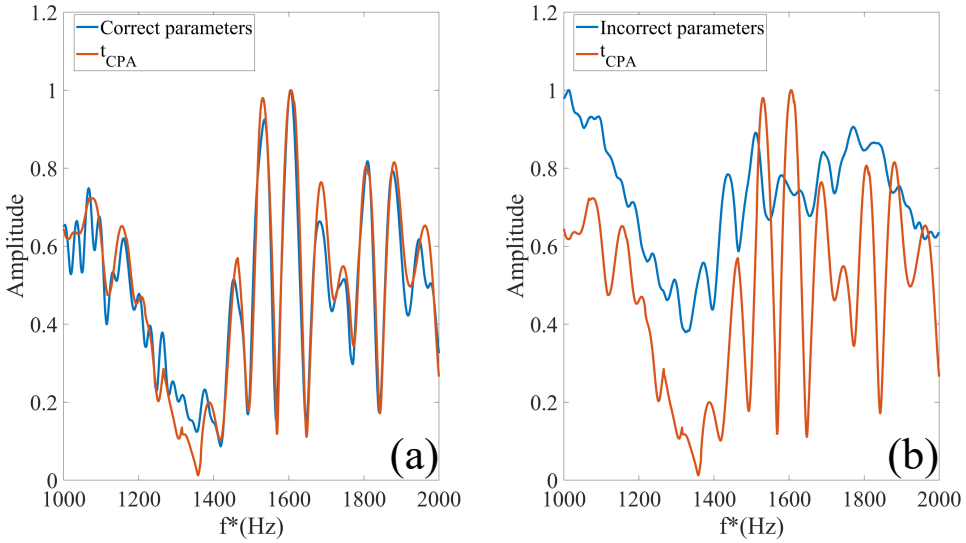


Figure 5. (**a**) When the parameters are correct, the comparison results of $\bar{I}_R(f)$ and $I(f, t_{CPA})$ are obtained. (**b**) When the parameters are wrong, the comparison results of $\bar{I}_R(f)$ and $I(f, t_{CPA})$ are obtained.

In Figure 5, we can see that under the premise of correct parameter values, the fluctuation structure of the converted time–frequency spectrogram at each moment is basically the same as that of the t_{CPA} moment, which makes the $\bar{I}_R(f)$ waveform change more intense and the periodicity more obvious. Under the premise of wrong parameter values, due to the different fluctuation structures of the converted time–frequency spectrogram at each moment, after cumulative averaging, the $\bar{I}_R(f)$ waveform changes more gently and the periodicity becomes more blurred. Therefore, we calculate the spectrum $Y(\Omega)$ of $\bar{I}_R(f)$ when the parameters are correct and wrong according to Equation (11).

$$Y(\Omega) = \int \bar{I}_R(\omega') e^{-j\Omega\omega'} d\omega', \omega' = \left(\frac{1}{2\pi f}\right)^{-\frac{1}{\beta}} \tag{11}$$

Here, Ω represents the corresponding frequency. The changes in the spectrum in the two cases are compared, as shown in Figure 6.

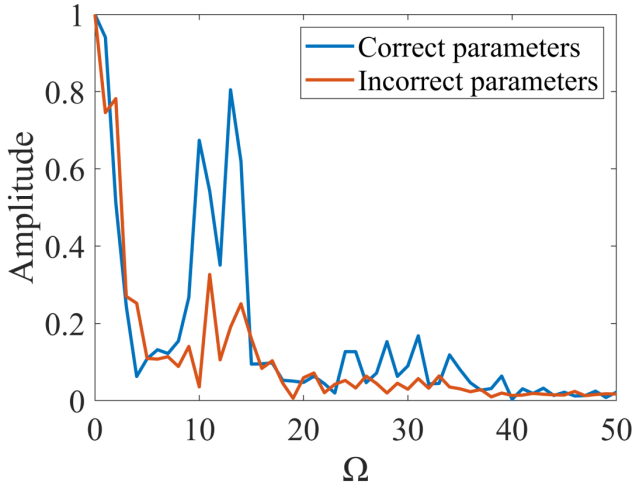


Figure 6. The spectrum analysis results of $\bar{I}_R(f)$ when the parameters are correct and wrong, respectively.

In Figure 6, we can clearly see the difference between the spectrum $Y(\Omega)$ when the parameters are correct and wrong. When the parameters are correct, the spectrum $Y(\Omega)$ waveform changes more violently, there are more obvious peaks, and the standard deviation of the signal is relatively large. When the parameters are wrong, the spectrum $Y(\Omega)$ waveform changes more gently, and the standard deviation of the signal is relatively small. Therefore, we can choose Equation (12) as the cost function to determine whether the parameter value is correct.

$$L = \sigma(Y(\Omega)) \tag{12}$$

2.2. Algorithm Description

Based on the previous description, we can summarize the whole processing flow of the algorithm as follows:

1. The received signal is transformed by a short-time Fourier transform to obtain its time–frequency spectrum $I(f,t)$, and the time window $[t_1, t_2]$ and frequency window $[f_1, f_2]$ for parameter estimation are selected in the time–frequency spectrum.
2. Define the search grids for parameters b, t_{CPA} and β, denoted by $\{b_1, b_2, \ldots, b_M\}$, $\{t_{CPA_1}, t_{CPA_2}, \ldots, t_{CPA_N}\}$ and $\{\beta_1, \beta_2, \ldots, \beta_Q\}$.
3. Select any parameter $\{b_m, t_{CPA_n}, \beta_q\}(b_m \in \{b_1, b_2, \ldots, b_M\}, t_{CPA_n} \in \{t_{CPA_1}, t_{CPA_2}, \ldots, t_{CPA_N}\}, \beta_q \in \{\beta_1, \beta_2, \ldots, \beta_Q\})$. The Hyperbola-warping transform is applied to the sound intensity in the $t \in [t_1, t_2]$ range to obtain $I_R(f, t, b_m, t_{CPAn}, \beta_q)$.
4. Calculate the average value of $I_R(f, t, b_m, t_{CPAn}, \beta_q)$ over the time window, denoted as $\overline{I}_R(f, b_m, t_{CPAn}, \beta_q)$

$$\overline{I}_R(f, b_m, t_{CPAn}, \beta_q) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} I_R(f, t, b_m, t_{CPAn}, \beta_q) dt \tag{13}$$

5. Calculate the spectrum of $\overline{I}_R(f, b_m, t_{CPAn}, \beta_q)$, denoted as $Y(\Omega, b_m, t_{CPAn}, \beta_q)$.
6. Define the standard deviation of the cost function as $Y(\Omega, b_m, t_{CPAn}, \beta_q)$, denoted by $L(b_m, t_{CPA_n}, \beta_q)$.

$$L(b_m, t_{CPA_n}, \beta_q) = \sigma(Y(\Omega, b_m, t_{CPAn}, \beta_q)) \tag{14}$$

7. Repeat steps 3–6 until all parameter combinations in the search grid are traversed. When the cost function $L(b_m, t_{CPA_n}, \beta_q)$ reaches the maximum value, the corresponding parameter combination is the estimated motion parameter, denoted by $\{\widehat{b}, \widehat{t}_{CPA}, \widehat{\beta}\}$.

$$\{\widehat{b}, \widehat{t}_{CPA}, \widehat{\beta}\} = \underset{b_m, t_{CPAn}, \beta_q}{\operatorname{argmax}} \left(L(b_m, t_{CPAn}, \beta_q)\right) \tag{15}$$

3. Simulation

The simulation is conducted in a basic shallow-water, range-independent waveguide environment, the sound speed in the water column is $c_w = 1500$ m/s, the water depth is 40 m, parameters of the bottom are $c_b = 1593$ m/s, $\rho_b = 1.76$ g/cc, $\alpha_b = 0.2$ dB/km/Hz. A target, positioned at a depth z of 2 m, navigates at a velocity of 5 m/s along the path depicted in Figure 1. A hydrophone, situated at a depth z_s of 7 m, remains in a fixed location. The time of the closest point of approach (t_{CPA}) between the hydrophone and the target is recorded as 500 s, with the corresponding nearest distance (r_{CPA}) being 500 m. The distance r between the hydrophone and the target is illustrated as the black line in Figure 7a. The normal mode model, Kraken, is utilized to compute the sound field between 200 s to 400 s (represented by the red line in Figure 7a), for parameter estimation, as shown in Figure 7b.

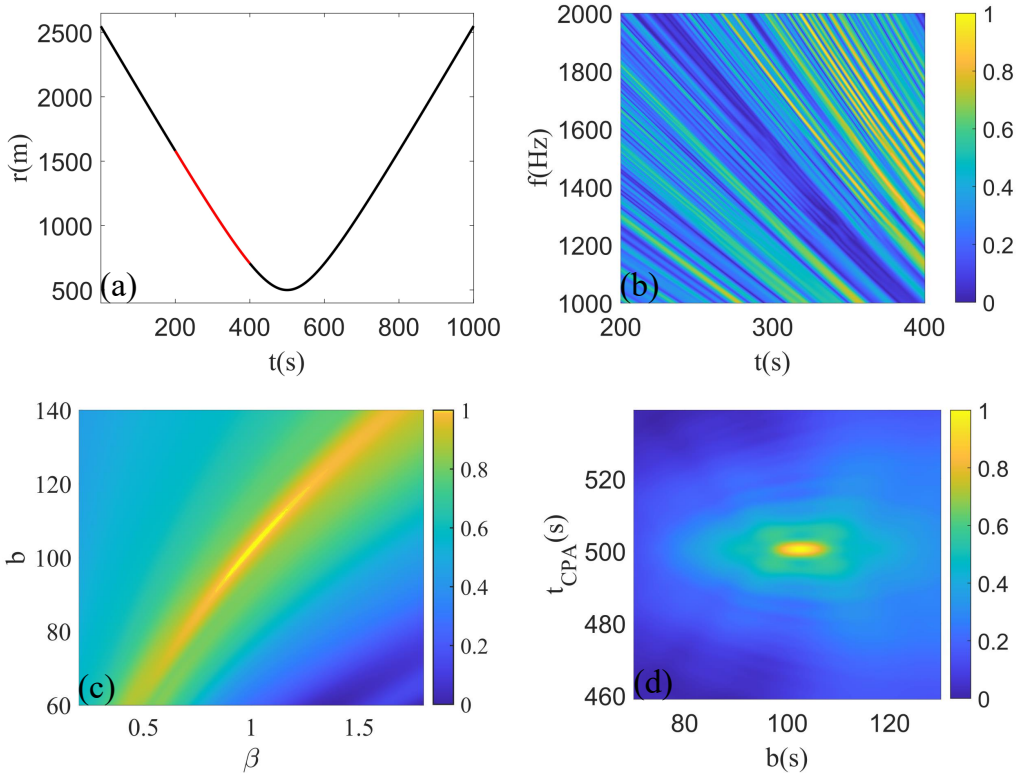


Figure 7. (**a**) The black line represents the distance between the hydrophone and the target throughout the entire time period, while the red line indicates the distance between the hydrophone and the target during the selected time segment. (**b**) Time–frequency spectrum of the signal received by the hydrophone. (**c**,**d**) Ambiguity map of motion parameter search.

According to the theoretical framework detailed in Section 2, the ambiguity map of the cost function L is presented in Figure 7c,d. The parameters corresponding to the peak value of the ambiguity map are $b = 101$ s, $t_{CPA} = 500$ s, $\beta = 0.95$, and the real parameters are $b = 100$ s, $t_{CPA} = 500$ s, $\beta = 1$. To demonstrate the algorithm's processing capability in the time–frequency domain on full-plane interference fringes, the time–frequency spectrum is transformed using the peak estimated parameters ($b = 101$ s, $t_{CPA} = 500$ s, $\beta = 0.95$), as depicted in Figure 8a. Each bright (dark) stripe in the image is a straight line. The error parameters ($b = 125$ s, $t_{CPA} = 470$ s, $\beta = 1$) are selected to transform the time–frequency spectrum, as shown in Figure 8b.

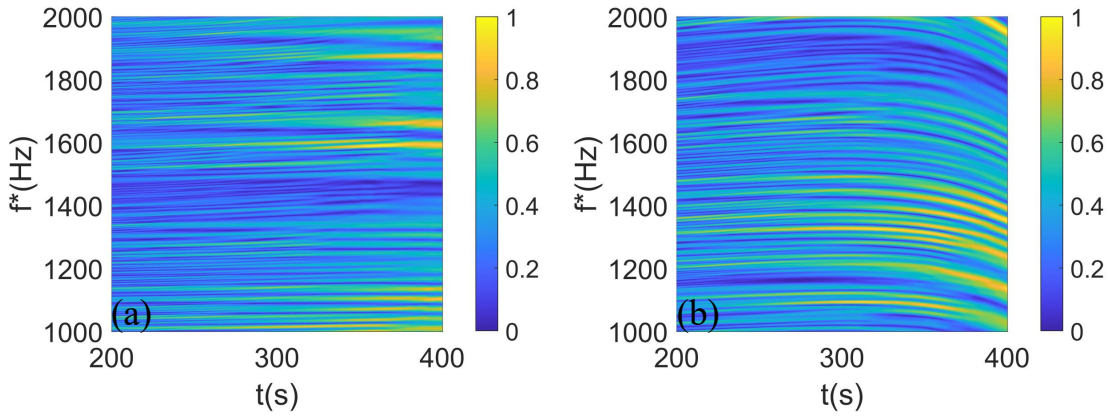


Figure 8. (**a**) The time–frequency spectrum when the parameter takes the peak value ($b = 101$ s, $t_{CPA} = 500$ s, $\beta = 0.95$) of the cost function. (**b**) The time–frequency spectrum when the parameter takes the wrong value ($b = 125$ s, $t_{CPA} = 470$ s, $\beta = 1$).

In contrast to Figure 8a,b, the algorithm presented in this paper takes into account the contributions of all interference fringes in the time–frequency domain to the analysis results.

As reported in Reference [21], the Hough transform can estimate parameters b, t_{CPA} and β, but it is designed to process individual interference fringes within the time–frequency domain.

To verify the anti-interference performance of the algorithm, we add white noise to the time–frequency spectrum diagram shown in Figure 7b according to Equation (16), so that $SNR = 10$ dB, as shown in Figure 9a. In this figure, the red dashed line represents the interference fringe selected by the Hough transform.

$$SNR = 10 * \log_{10}^{(P_s / P_n)} \tag{16}$$

where P_s represents the average power of the signal and P_n represents the average power of the noise.

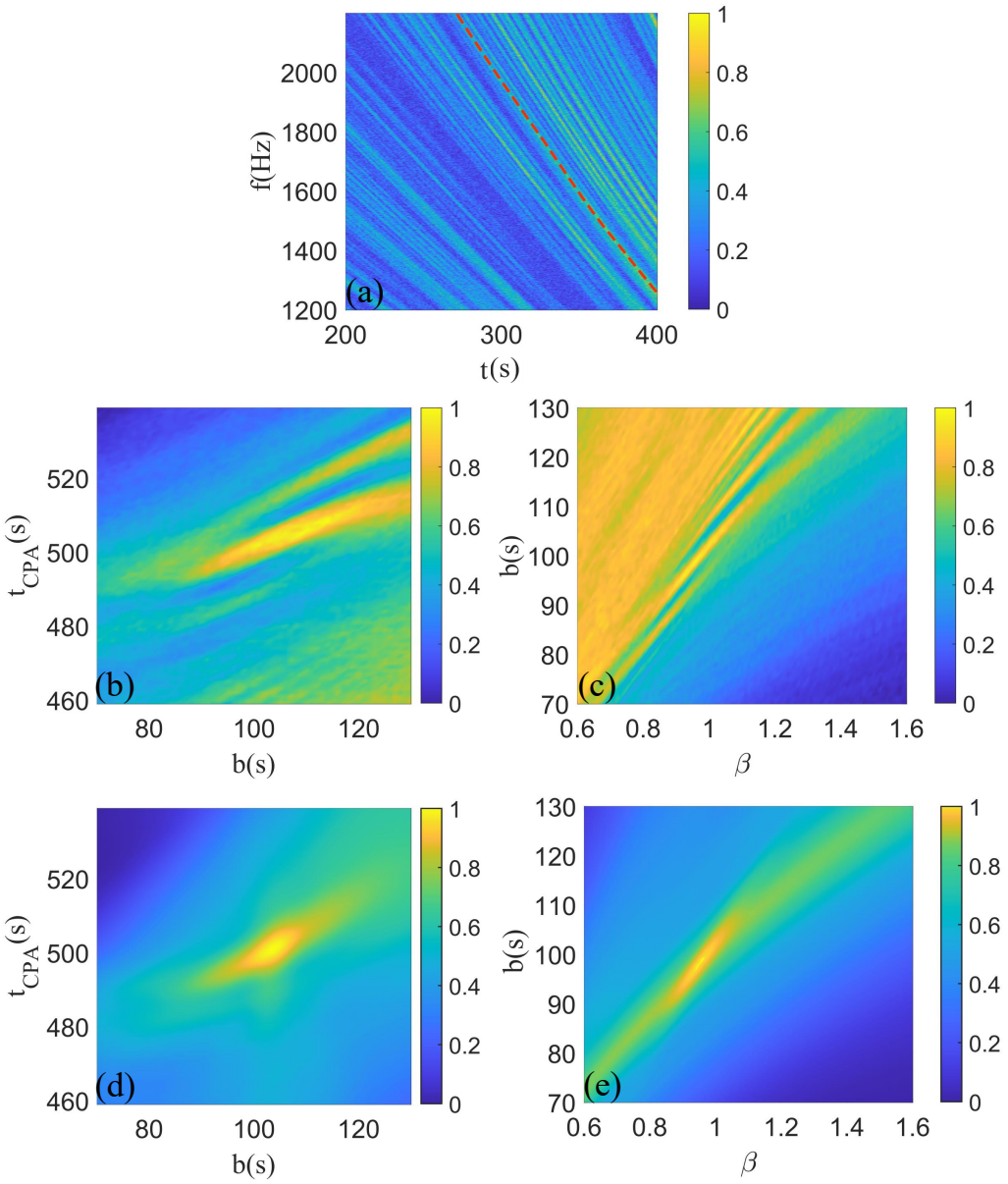


Figure 9. (**a**) Add Gaussian white noise with SNR = 10 dB in Figure 7b, where the red line represents the fringe selected by the Hough transform. (**b**,**c**) are the parameter estimation results of the Hough transform. (**d**,**e**) are the parameter estimation results of the full-plane Hyperbola-warping transform.

Analyzing the processing results of the two algorithms in Figure 9, at a signal-to-noise ratio (SNR) of 10 dB, the peak of the Hough transform is located at $b = 111$ s, $t_{CPA} = 505$ s, $\beta = 0.9$; whereas the peak of the full-plane Hyperbola-warping transform is at $b = 104$ s, $t_{CPA} = 501$ s, $\beta = 0.96$, with the actual values being $b = 100$ s, $t_{CPA} = 500$ s, $\beta = 1$. Comparing the outcomes, the conventional Hough transform can essentially predict the target parameters at this SNR, but since it processes a single stripe in the time–frequency spectrum, the parameter estimation results are subject to false peak interference, leading to relatively higher errors, as depicted in Figure 9b,c. The full-plane Hyperbola-warping transform proposed in this paper considers the contributions of all interference fringes in the time–frequency domain, therefore achieving superior parameter estimation results, as shown in Figure 9d,e.

To verify the performance of the two algorithms under different SNRs, white noise with different SNRs is added to the time–frequency spectrum of Figure 7b. The proposed algorithm and the conventional Hough transform are applied to time–frequency spectrums under varying signal-to-noise ratios, with the estimation results for three parameters presented in Figure 10a–c. It can be observed from these figures that the conventional Hough transform exhibits significant estimation errors for all three parameters when the SNR is below 6 dB. Even at SNRs of 6 dB and above, the estimation results from the Hough transform show considerable fluctuations due to noise interference. In contrast, the proposed algorithm, which takes into account the contributions of all interference fringes in the time–frequency domain, is capable of estimating parameters even at SNRs as low as −4 dB. This demonstrates approximately a 10 dB improvement in noise resistance compared to the conventional Hough transform. Moreover, at higher SNRs, the parameter estimation results from the proposed algorithm are more stable.

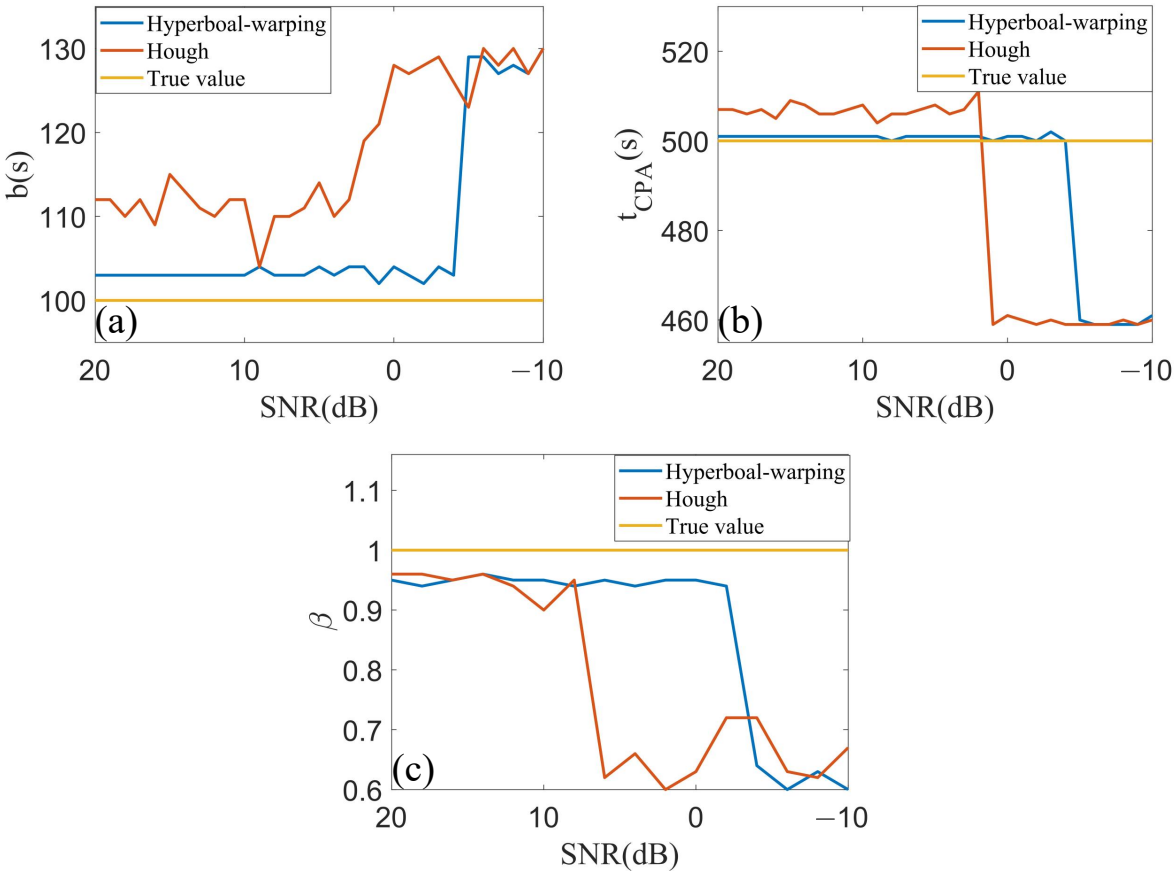


Figure 10. (**a**–**c**) present comparisons between the estimated target parameters at various SNRs and their corresponding true values.

4. Results of Sea Trial Experiments

The sea trial data come from the single hydrophone target observation experiment carried out by the Acoustic Laboratory of Ocean University of China in Qingdao offshore in June 2022. The experiment took place near Xiaogong Island in the vicinity of Qingdao, where the water depth was approximately 24 m, as shown in Figure 11a. The position of the hydrophone is marked with a red circle in Figure 11a, and the sound velocity profile is shown in Figure 11b.

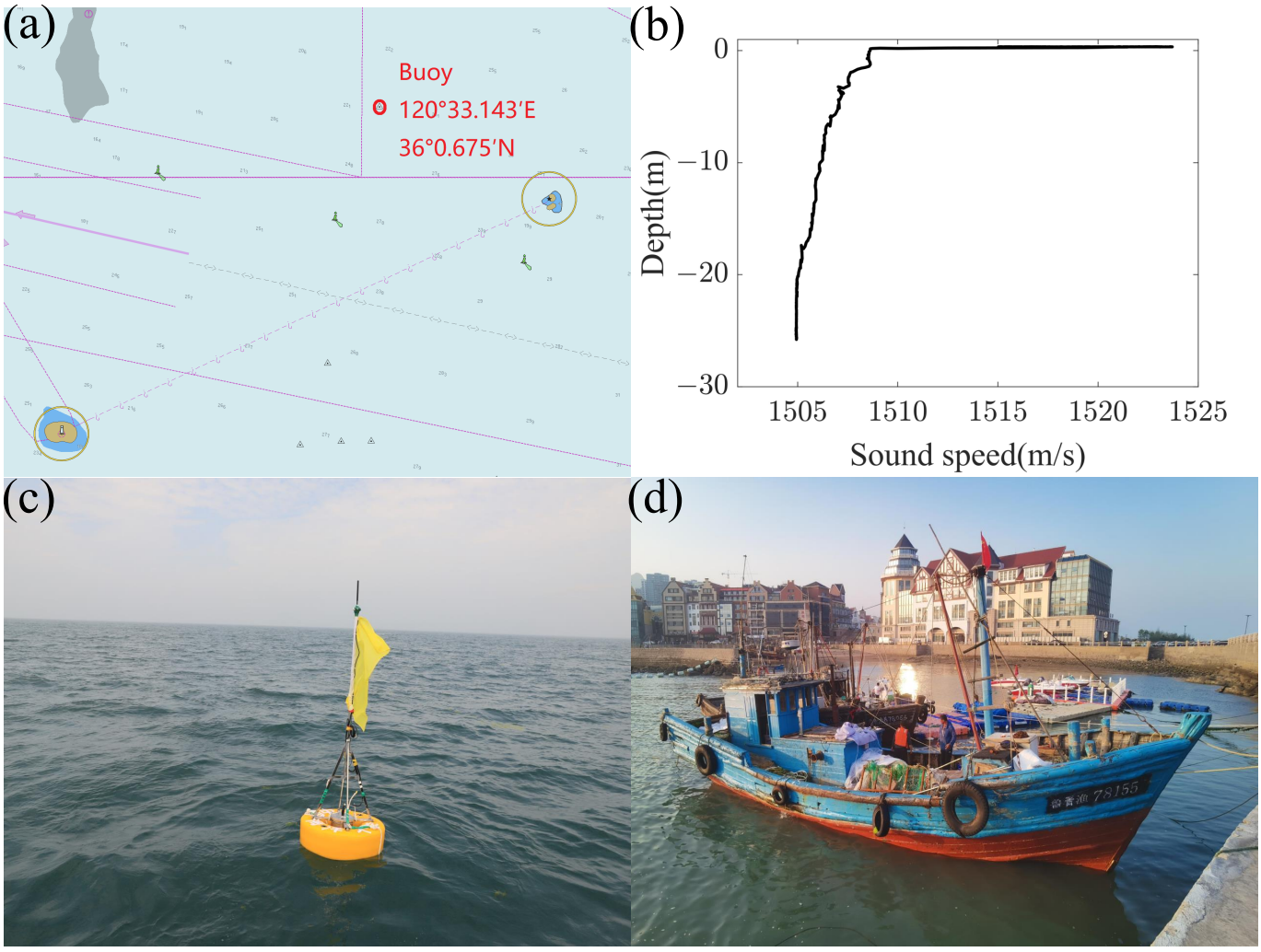


Figure 11. (**a**) Nautical chart of the experimental area, where the purple dashed line represents the shipping route. (**b**) The sound velocity profile. (**c**) Physical map of the buoy. (**d**) Physical map of the fishing boat used in the experiment (target ship).

The hydrophone used in the experiment is a self-contained hydrophone with a sensitivity of −180 dB re 1V/µPa. It is fixed on the buoy and has a depth of about 5 m, as shown in Figure 11c. The target vessel is a 150-horsepower fishing boat, depicted in Figure 11d. The vessel's trajectory and the buoy's position are detailed in Figure 12a, with the distance between them illustrated in Figure 12b. The nearest distance r_{CPA} is 233.33 m, and the nearest point time t_{CPA} is 175 s.

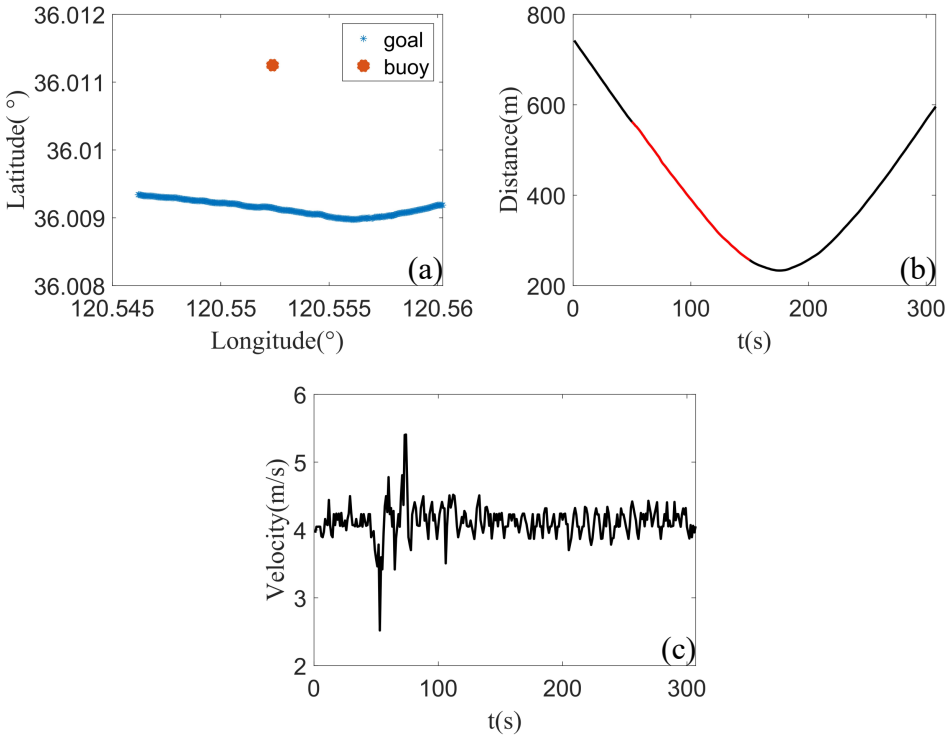


Figure 12. (**a**) The target trajectory and buoy position. (**b**) The black line indicates the distance between the target and the buoy throughout the entire time period, while the red line represents the distance during the selected time segment. (**c**) The target navigation speed.

According to the data of hand-held GPS, the moving speed of the target ship is drawn as shown in Figure 12c, and the average speed $\bar{v}$ between 90 s and 300 s is 4.1301 m/s. If we take the average speed $\bar{v}$ of the target in this period as the speed of the target, the real parameter $b = r_{CPA}/v = 233.33/4.1301 = 56.495$ (s) of the target can be inferred.

Figure 13 shows the time–frequency diagram of the noise signal received by the target during the trajectory navigation shown in Figure 12a. It can be seen from the figure that the overall shape of the interference fringes is hyperbolic, which is consistent with the simulation results, but the signal-to-noise ratio of the overall signal is relatively low. From the frequency domain axis, the more obvious interference fringes are concentrated in the frequency range of $[600, 4000]$ Hz. The frequency band less than 600 Hz is seriously disturbed by background noise. From the time axis, in the time period of $[50, 300]$ S, the interference fringes are more obvious.

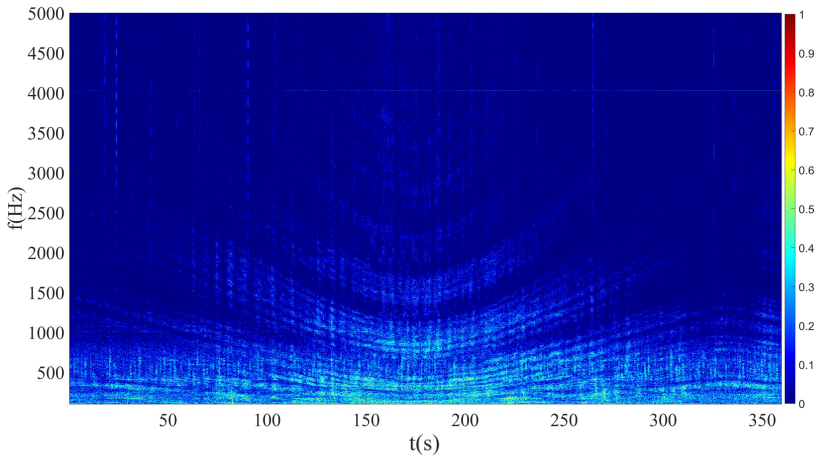


Figure 13. Time–frequency diagram of buoy receiving target ship radiated noise.

Therefore, for parameter estimation, we select the time segment from 50 s to 150 s and the frequency band from 600 Hz to 4000 Hz, as indicated in the red box of Figure 14a. Utilizing the algorithm detailed in Section 2, the cost function for parameter estimation is derived and displayed in Figure 14b,c. The parameter estimates corresponding to the peak of the ambiguity map are: $b = 58$ s, $t_{CPA} = 175$ s, $\beta = 0.95$, which are basically consistent with the real parameters($b = 56.495$ s, $t_{CPA} = 175$ s, $\beta = 1.02$).

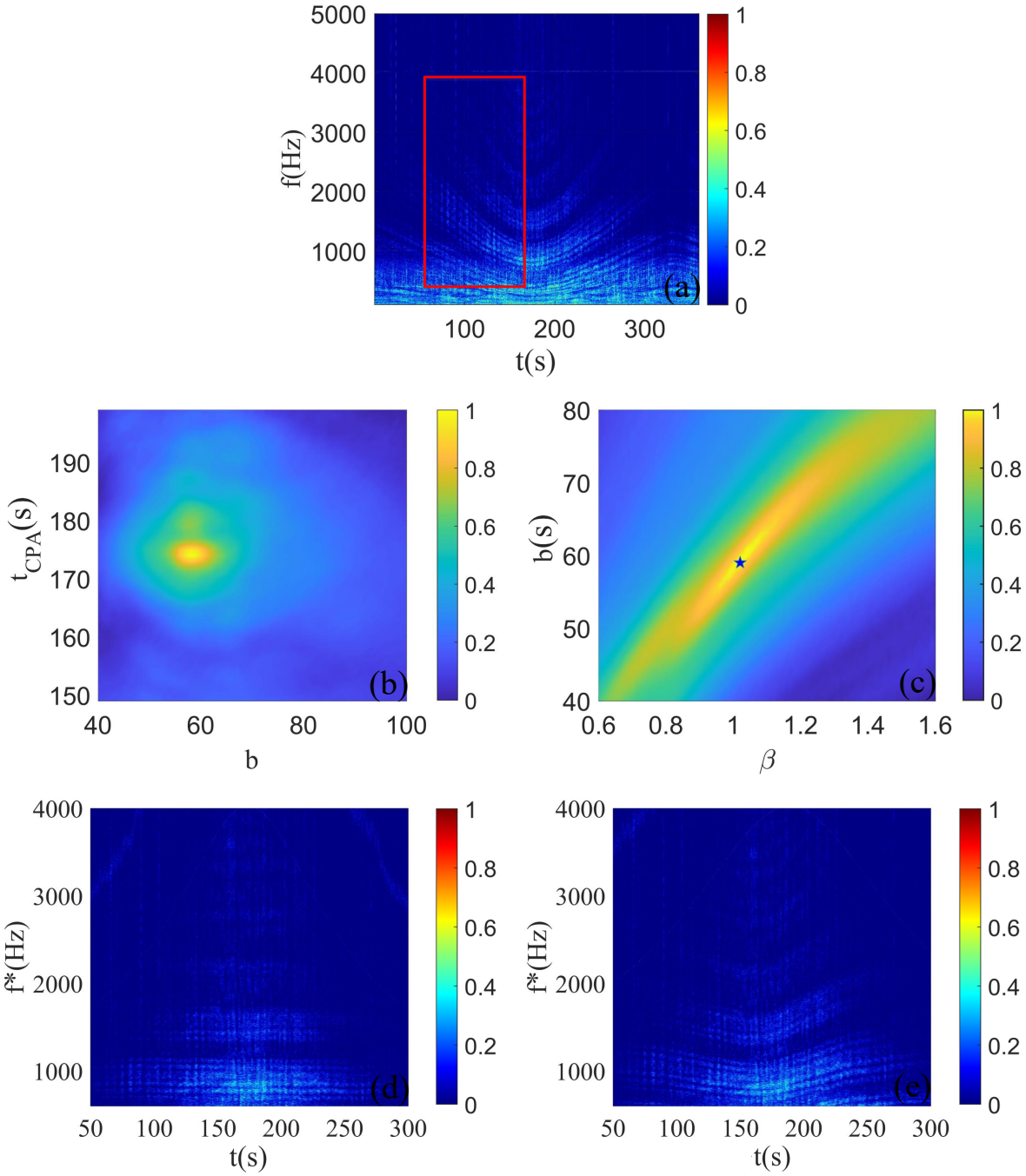


Figure 14. (**a**) Time–frequency diagram of a buoy receiving target ship radiated noise, where the red square indicates the selected time segment and frequency band. (**b**,**c**) are the parameter estimation results of the full-plane Hyperbola-warping transform, where the pentagram indicates the peak values. (**d**) The time–frequency spectrum when the parameter takes the peak value ($b = 58$ s, $t_{CPA} = 175$ s, $\beta = 0.95$) of the cost function. (**e**) The time–frequency spectrum when the parameter takes the wrong value ($b = 80$ s, $t_{CPA} = 190$ s, $\beta = 0.9$).

To show the processing ability of the algorithm for interference fringes more vividly, we set the parameter as the peak value of the cost function ($b = 58$ s, $t_{CPA} = 175$ s, $\beta = 0.95$). The converted time–frequency spectrum is shown in Figure 14d. The interference fringes in the figure become straight lines. When the parameter is set to the error value ($b = 80$ s, $t_{CPA} = 190$ s, $\beta = 0.9$), the converted time–frequency spectrum is shown in Figure 14e. The interference fringes in the figure show irregular bending.

To compare the processing capabilities of the traditional Hough transform and the full-plane Hyperbola-warping transform proposed in this paper on actual sea trial data, we selected signals from the same time segments for Hough transform processing, as shown in Figure 15a. The resulting parameter estimation is depicted in Figure 15b,c, with peaks located at $b = 58$ s, $t_{CPA} = 173$ s, $\beta = 1.06$, which are close to the true parameters ($b = 56.495$ s, $t_{CPA} = 175$ s, $\beta = 1.02$). However, the parameter estimation contains many false peaks, making the results less distinct.

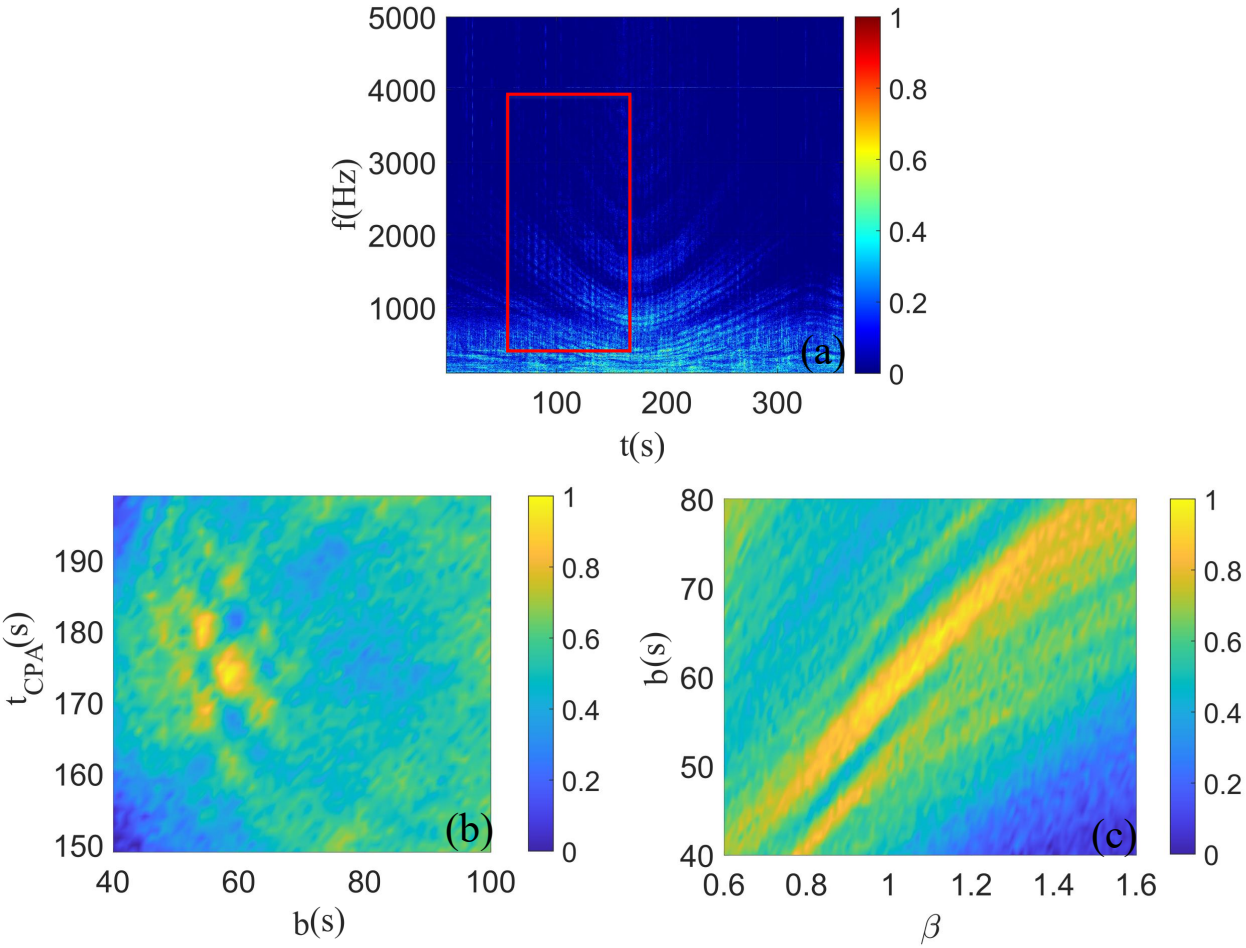


Figure 15. (**a**) The time–frequency spectrum processed by the Hough transform (within the red box). (**b**,**c**) are the parameter estimation results of the Warping transform.

To further compare the processing effects of the traditional Warping transform and the full-plane Hyperbola-warping transform on actual data, we set the frequency band to [600, 4000] Hz and selected time periods ranging from [5, 105] s to [180, 280] s, performing estimations at every 5 s interval. The comparative results for the three parameters are displayed in Figure 16a–c.

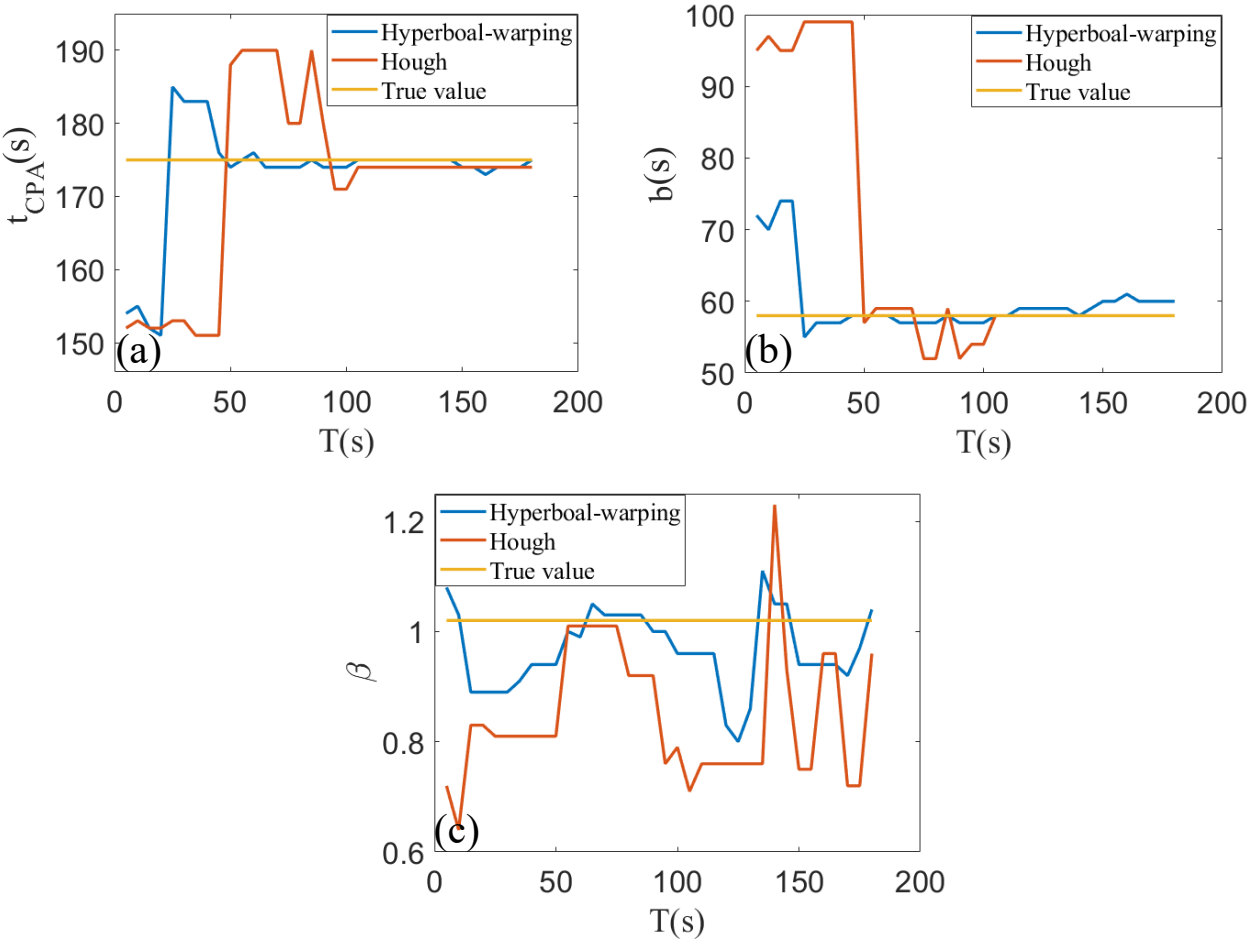


Figure 16. (**a**–**c**) present comparisons between the estimated target parameters at various SNRs and their corresponding true values.

To quantify the estimation capability of the two algorithms for the three parameters, an increase in detection distance D is introduced as a metric for their detection ability, as shown in Equation (17).

$$D = \frac{|t_{CPA} - t_W|}{|t_{CPA} - t_H|} - 1 \tag{17}$$

where t_H denotes the time when the Hough transform approaches the true value, and t_W indicates the time when the algorithm presented in this paper approaches the true value.

Analyzing Figure 16a for the estimation of the parameter t_{CPA}, the conventional Hough transform begins to approach the true value at 50 s, stabilizing around the true value around 90 s, whereas the algorithm proposed in this paper starts to approach the true value at 25 s and stabilizes around the true value by 50 s, resulting in an increase in detection distance D by 47%. Analyzing Figure 16b for the estimation of the parameter b, the traditional Hough transform approaches the true value at 50 s, while the proposed algorithm approaches the true value at 25 s, resulting in an increase in detection distance D by 20%. Analyzing Figure 16c for the estimation of the parameter β, the conventional Hough transform begins to approach the true value at 50 s with significant estimation errors, whereas the algorithm presented in this paper starts to approach the true value at 25 s and substantially reduces the estimation errors, resulting in an increase in detection distance D by 20%.

Considering the estimation results of the three parameters, the algorithm proposed in this paper demonstrates better stability, smaller estimation errors, and at least a 20% increase in detection distance compared to the conventional Hough transform.

5. Discussion and Concluding Remarks

In this paper, we introduce a novel single hydrophone target parameter estimation method that utilizes the Hyperbola-warping transform. It can estimate the motion parameters of the target: t_{CPA}, $b = r_{CPA}/v$ and β before the target reaches the nearest point when only the time–frequency spectrum of the partial target is known. Unlike conventional methods (Hough transform), our approach takes into account the contributions of all interference fringes in the time–frequency domain, leading to improved accuracy and significantly enhanced noise resistance.

In simulation experiments, the proposed method demonstrated approximately 10 dB superior anti-noise performance compared to the traditional Hough transform. Furthermore, during sea trial experiments, our method was able to predict the target parameters with a detection range improvement of 20% and with greater stability than the Hough transform. These results substantiate the proposed algorithm's enhanced anti-noise capabilities and stable parameter estimation abilities.

However, this study focuses solely on range-independent waveguide environments. The effectiveness of the algorithm in range-dependent waveguide environments warrants investigation in future research.

Author Contributions: Conceptualization, Y.L. and B.G.; methodology, Y.L.; software, Y.L.; validation, Y.L., Z.C. and Y.Y.; formal analysis, Y.L.; investigation, Y.L.; resources, D.G. and B.G.; data curation, Y.L. and Z.W.; writing—original draft preparation, Y.L.; writing—review and editing, Y.L. and B.G.; visualization, Y.L.; supervision, D.G. and B.G.; project administration, D.G.; funding acquisition, D.G. and B.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant Nos. 12274385, 12374427.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Thode, A.M.; Kuperman, W.; D'Spain, G.; Hodgkiss, W. Localization using Bartlett matched-field processor sidelobes. *J. Acoust. Soc. Am.* **2000**, *107*, 278–286. [CrossRef]
2. Thode, A.M. Source ranging with minimal environmental information using a virtual receiver and waveguide invariant theory. *J. Acoust. Soc. Am.* **2000**, *108*, 1582–1594. [CrossRef]
3. Heaney, K.D. Rapid geoacoustic characterization using a surface ship of opportunity. *IEEE J. Ocean. Eng.* **2004**, *29*, 88–99. [CrossRef]
4. Heaney, K.D. Rapid geoacoustic characterization: Applied to range-dependent environments. *IEEE J. Ocean. Eng.* **2004**, *29*, 43–50. [CrossRef]
5. Li, P.; Wu, Y.; Guo, W.; Cao, C.; Ma, Y.; Li, L.; Leng, H.; Zhou, A.; Song, J. Striation-based beamforming with two-dimensional filtering for suppressing tonal interference. *J. Mar. Sci. Eng.* **2023**, *11*, 2117. [CrossRef]
6. Song, W.; Gao, D.; Li, X.; Kang, D.; Li, Y. Waveguide invariant in a gradual range-and azimuth-varying waveguide. *JASA Express Lett.* **2022**, *2*, 056002. [CrossRef]
7. Song, H.; Byun, G. Localization of a distant ship using a guide ship and a vertical array. *J. Acoust. Soc. Am.* **2021**, *149*, 2173–2178. [CrossRef]
8. Carter, G. Time delay estimation for passive sonar signal processing. *IEEE Trans. Acoust. Speech Signal Process.* **1981**, *29*, 463–470. [CrossRef]
9. Wu, Q.; Xu, Y. A Nonlinear Data-Driven Towed Array Shape Estimation Method Using Passive Underwater Acoustic Data. *Remote Sens.* **2022**, *14*, 304. [CrossRef]
10. Zhang, W.; Jiang, P.; Lin, J.; Sun, J. Array shape calibration based on coherence of noise radiated by non-cooperative ships. *Ocean Eng.* **2024**, *303*, 117792. [CrossRef]
11. Nardone, S.; Lindgren, A.; Gong, K. Fundamental properties and performance of conventional bearings-only target motion analysis. *IEEE Trans. Autom. Control* **1984**, *29*, 775–787. [CrossRef]
12. Lowney, M.P.; Bar-Shalom, Y.; Luginbuhl, T.; Willett, P. Estimation of the Target State with Passive Measurements and Prior Range Information. In Proceedings of the OCEANS 2023-MTS/IEEE US Gulf Coast, Biloxi, MS, USA, 25–28 September 2023; IEEE: New York, NY, USA, 2023; pp. 1–7.
13. Hinich, M.J. Maximum-likelihood signal processing for a vertical array. *J. Acoust. Soc. Am.* **1973**, *54*, 499–503. [CrossRef]

14. Zhang, X.; Yang, P.; Wang, Y.; Shen, W.; Yang, J.; Wang, J.; Ye, K.; Zhou, M.; Sun, H. A novel multireceiver sas rd processor. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4203611. [CrossRef]
15. Zhang, X.; Yang, P.; Wang, Y.; Shen, W.; Yang, J.; Ye, K.; Zhou, M.; Sun, H. LBF-based CS algorithm for multireceiver SAS. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1502505. [CrossRef]
16. Wang, C.; Ma, S.; Meng, Z. Array Invariant-Based Source Localization in Deep Sea by Vertical Linear Array. In Proceedings of the 2019 IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP), Weihai, China, 28–30 September 2019; IEEE: New York, NY, USA, 2019; pp. 180–184.
17. Byun, G.; Song, H. Adaptive array invariant. *J. Acoust. Soc. Am.* **2020**, *148*, 925–933. [CrossRef]
18. Wang, Y.; Chi, C.; Li, Y.; Ju, D.; Huang, H. Passive Array-Invariant-Based Localization for a Small Horizontal Array Using Two-Dimensional Deconvolution. *Appl. Sci.* **2022**, *12*, 9356. [CrossRef]
19. Lee, S.; Makris, N.C. The array invariant. *J. Acoust. Soc. Am.* **2006**, *119*, 336–351. [CrossRef] [PubMed]
20. Tao, H.; Hickman, G.; Krolik, J.L.; Kemp, M. Single hydrophone passive localization of transiting acoustic sources. In Proceedings of the OCEANS 2007-Europe, Aberdeen, UK, 18–21 June 2007; IEEE: New York, NY, USA, 2007; pp. 1–3.
21. Turgut, A.; Orr, M.; Rouseff, D. Broadband source localization using horizontal-beam acoustic intensity striations. *J. Acoust. Soc. Am.* **2010**, *127*, 73–83. [CrossRef] [PubMed]
22. Zhao, C.; Liang, G.; Wang, J.; Qiu, L.; Liu, G.; Dong, W. Single-hydrophone-based passive localization for vertically moving targets in shallow water. *Ocean Eng.* **2024**, *306*, 118062. [CrossRef]
23. Baraniuk, R.G.; Jones, D.L. Unitary equivalence: A new twist on signal processing. *IEEE Trans. Signal Process.* **1995**, *43*, 2269–2282. [CrossRef]
24. Bonnel, J.; Thode, A.; Wright, D.; Chapman, R. Nonlinear time-warping made simple: A step-by-step tutorial on underwater acoustic modal separation with a single hydrophone. *J. Acoust. Soc. Am.* **2020**, *147*, 1897–1926. [CrossRef]
25. Qi, Y.B.; Zhou, S.H.; Zhang, R.H.; Zhang, B.; Ren, Y. Modal characteristic frequency in a range-dependent shallow-water waveguide and its application to passive source range estimation. *Acta Phys. Sin.* **2014**, *63*, 044303. [CrossRef]
26. Bonnel, J.; Dosso, S.E.; Ross Chapman, N. Bayesian geoacoustic inversion of single hydrophone light bulb data using warping dispersion analysis. *J. Acoust. Soc. Am.* **2013**, *134*, 120–130. [CrossRef] [PubMed]
27. Niu, H.Q.; Zhang, R.H.; Li, Z.L.; Guo, Y.G.; He, L. Bubble pulse cancelation in the time-frequency domain using warping operators. *Chin. Phys. Lett.* **2013**, *30*, 084301. [CrossRef]
28. Mao, J.; Peng, Z.; Zhang, B.; Wang, T.; Zhai, Z.; Hu, C.; Wang, Q. An Underwater Localization Algorithm for Airborne Moving Sound Sources Using Doppler Warping Transform. *J. Mar. Sci. Eng.* **2024**, *12*, 708. [CrossRef]
29. Li, X.; Piao, S.; Zhang, H. Modified Warping transformation for typical waveguide with semi-infinite seabed. *Harbin Gongcheng Daxue Xuebao/J. Harbin Eng.* **2018**, *39*, 1257–1263.
30. Wang, D.; Guo, L.-H.; Liu, J.-J.; Qi, Y.-B. Passive impulsive source range estimation based on warping operator in shallow water. *Acta Phys. Sin.* **2016**, *65*, 104302. [CrossRef]
31. Brekhovskikh, L.M.; Lysanov, Y.P. *Fundamentals of Ocean Acoustics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2003.
32. Chuprov, S.; Brekhovskikh, L. Interference structure of a sound field in a layered ocean. *Ocean Acoust. Curr. State* **1982**, 71–91.
33. Porter, M.B. *The KRAKEN Normal Mode Program*; Naval Research Lab: Washington, DC, USA, 1992.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

A Novel Chirp-Z Transform Algorithm for Multi-Receiver Synthetic Aperture Sonar Based on Range Frequency Division

Mingqiang Ning [1], Heping Zhong [1], Jinsong Tang [1,*], Haoran Wu [1], Jiafeng Zhang [1], Peng Zhang [2] and Mengbo Ma [3]

[1] Naval Institute of Underwater Acoustic Technology, Naval University of Engineering, Wuhan 430033, China; mingqiangning@outlook.com (M.N.); zheping525@sohu.com.cn (H.Z.); wuhaoran_wh@163.com (H.W.); 21000901@nue.edu.cn (J.Z.)
[2] School of Meteorology and Oceanography, National University of Defense Technology, Changsha 410073, China; zhangpeng23a@nudt.edu.cn
[3] Academy of Military Sciences, Beijing 100850, China; mengboma@163.com
* Correspondence: 18500901@nue.edu.com

Abstract: When a synthetic aperture sonar (SAS) system operates under low-frequency broadband conditions, the azimuth range coupling of the point target reference spectrum (PTRS) is severe, and the high-resolution imaging range is limited. To solve the above issue, we first convert multi-receivers' signal into the equivalent monostatic signal and then divide the equivalent monostatic signal into range subblocks and the range frequency subbands within each range subblock in order. The azimuth range coupling terms are converted into linear terms based on piece-wise linear approximation (PLA), and the phase error of the PTRS within each subband is less than $\pi/4$. Then, we use the chirp-z transform (CZT) to correct range cell migration (RCM) to obtain low-resolution results for different subbands. After RCM correction, the subbands' signals are coherently summed in the range frequency domain to obtain a high-resolution image. Finally, different subblocks are concatenated in the range time domain to obtain the final result of the whole swath. The processing of different subblocks and different subbands can be implemented in parallel. Computer simulation experiments and field data have verified the superiority of the proposed method over existing methods.

Keywords: synthetic aperture sonar (SAS); scaled Fourier transformation; chirp-z transform (CZT); frequency division; range cell migration correction (RCMC)

Citation: Ning, M.; Zhong, H.; Tang, J.; Wu, H.; Zhang, J.; Zhang, P.; Ma, M. A Novel Chirp-Z Transform Algorithm for Multi-Receiver Synthetic Aperture Sonar Based on Range Frequency Division. *Remote Sens.* **2024**, *16*, 3265. https://doi.org/10.3390/rs16173265

Academic Editors: Xuebo Zhang, Lei Zhang, Jingyue Lu, Haiyong Xu, Marc Pinto and Farman Ali

Received: 8 July 2024
Revised: 29 August 2024
Accepted: 30 August 2024
Published: 3 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Synthetic aperture sonar (SAS) has high resolution by coherently receiving sideboard echo data along the azimuth direction [1–5] and its azimuth resolution is independent of range [4,6–9]. Therefore, SAS is widely used to detect high-value targets such as mines [10–14], buried objects [15,16], or seabed remote sensing [17–20]. The low-speed nature of sound in seawater [21,22] results in two issues for the SAS system. The first issue is the contradiction between the maximum detection range and the minimum azimuth sampling frequency. The maximum detection range of SAS is determined by pulse repetition frequency (PRF), and the SAS system with low PRF is generally used to obtain a larger swath. At the same time, SAS samples along the azimuth direction at a fixed PRF. If the PRF is too low, it may be smaller than the Doppler bandwidth, resulting in frequency aliasing and ghost targets in the azimuth direction [23]. To solve the above contradiction, in practice, SAS systems use single-transmit/multi-receive configuration [24,25], where multiple receivers sample the echo signal synchronously. This type of SAS is also known as the multi-receiver SAS. The multi-receiver SAS's configuration improves the actual sampling frequency in the azimuth direction when PRF remains unchanged, overcoming the first issue. The second issue is that the low speed of underwater sound causes the along-track motion of the SAS platform between signal transmission and reception. SAS imaging algorithms must adopt a complex

non-stop–hop–stop (NSHS) assumption [26] instead of stop–hop–stop assumption widely used in SAR [27,28], and the platform's along-track motion between signal transmission and reception must be considered.

The phase center approximation (PCA) method [29] for multi-receiver SAS is based on NSHS assumption and uses the distance from the effective phase center (EPC) to the target as the round-trip range history. The beam center approximation (BCA) used in the PCA method leads to an increase in range history error with the rise in beamwidth [30], which cannot achieve high-resolution imaging (range history error < $\lambda/8$). Zhang et al. [31] proposed a new range history for the multi-receiver SAS based on the method of reversion (MSR) [32–36], which also adopts BCA and is inefficient due to the receiver-by-receiver processing. The LBF method [7,20,37–40] can acquire a more accurate point target reference spectrum (PTRS), also called the two-dimensional frequency spectrum, and obtain results only through one imaging, but BCA and the assumption that the transmitter and receiver equally contribute to the azimuth modulation are used. The azimuth range coupling of the PTRS is caused by range cell migration (RCM) [33,41] and has not been carefully considered yet due to the insignificance of azimuth range coupling of the PTRS in side-looking view, narrowband, and narrow beamwidth conditions.

However, high-order terms (azimuth range coupling) cannot be ignored under broadband and wide beam conditions. The most accurate method is to use omega-KA [42,43] (OKA) to directly compensate for the azimuth range coupling term without any approximation, but the Stolt interpolation is time-consuming. The range-Doppler algorithm (RDA) [7,44,45] also requires time domain interpolation processing to perform RCM Correction (RCMC) and is not efficient. The chirp scaling algorithms (CSA) [46–49] use fast Fourier transform (FFT) and phase multiplication in the RCMC to avoid interpolation and significantly improve computational efficiency. The original CSA does not consider the spatially variant range in the secondary range compression (SRC), resulting in the limited high-resolution imaging range. To increase the high-resolution imaging range, the non-linear CSA (NCSA) [50–52] solves the problem of range spatial variation in SRC by using a range-variant non-linear scaling equation to make the chirp rate linearly change with range. However, like CSA, it requires that the transmitted signal must be a chirp signal. Loffeld et al. [53] proposed the scaled Fourier transform (SCFT) to perform range migration correction, eliminating the azimuth range coupling caused by RCM in the PTRS, avoiding interpolation, and the impact of truncation interpolation [54]. However, the influence of SRC was not considered in [53]. Wu et al. [55] considered the influence of SRC and applied SCFT to the squint mode's SAR imaging, but also did not solve the problem of range variation in SRC. Although SCFT has one more FFT operation than CS, it can be applied to non-chirp signals [53] and can be achieved by the chirp-z transform (CZT) [56–58] in the digital system. Franceschetti et al. [59] divided the range into several subblocks continuously to decrease the range spatial variance of the SRC in squint mode SAR imaging. Zhang et al. [60] proposed a range-subblock processing method to decrease the PTRS error and residual quadratic coupling error for a multi-receiver SAS system with a wide-bandwidth signal. Residual coupling error will be small when the blocks are very dense, but the computational efficiency will not be high when azimuth range coupling is severe.

This paper proposes a CZT algorithm for multi-receiver synthetic aperture sonar based on range frequency division. First, the exact analytical PTRS based on the range history shifting (RHS) is obtained without any approximation in azimuth FFT. The above PTRS is expanded into a series of the range frequency and kept to the quadratic term. Then, the processed PTRS is divided into subblocks, and Taylor expansion is used in the coefficients of the range frequency at the center of each subblock. Residual RCM and SRC terms are obtained by performing high-order phase compensation and bulk phase compensation. In each range subblock, we divide the PTRS into subbands in the range frequency domain to ensure that the error of the PTRS is less than $\pi/4$. The residual SRC term can be merged into the residual RCM term by the piece-wise linear approximation (PLA), which means that the

square term of range frequency is divided into many small linear frequency segments, and then the residual RCM term can be corrected through CZT to obtain low-resolution results of different subbands. The high-resolution result is obtained by coherent superposition of low-resolution resolution results, and then all range subblocks are concatenated in the range time domain to obtain the focusing image of the whole swath.

The remaining parts of the article are as follows: In Section II, we establish multi-receiver SAS's model based on RHS and finally derive the accurate PTRS of each receiver. In Section III, the multi-receiver SAS's data is converted into the monostatic equivalent signal by monostatic conversion, and the PTRS is divided into subblocks in the range time domain and subbands in the range frequency domain. The low-resolution results of the range subbands are obtained through the CZT algorithm, and the whole image can be obtained by concatenating low-resolution results in the range time domain after coherent superposition. In Section IV, simulation and field data results are presented to demonstrate the effectiveness of the proposed method. Section V serves as a summary.

2. The Multi-Receiver SAS's Model

2.1. The Accurate Range History

The geometric schematic diagram of the multi-receiver SAS operating in side-looking mode is shown in Figure 1. The SAS platform consists of one transmitter and $M_1 + M_2 + 1$ uniformly arranged receivers, numbered $-M_1, -M_1 + 1, \ldots, 0, \ldots, M_2$. Receivers are represented in yellow, and red represents the transmitter.

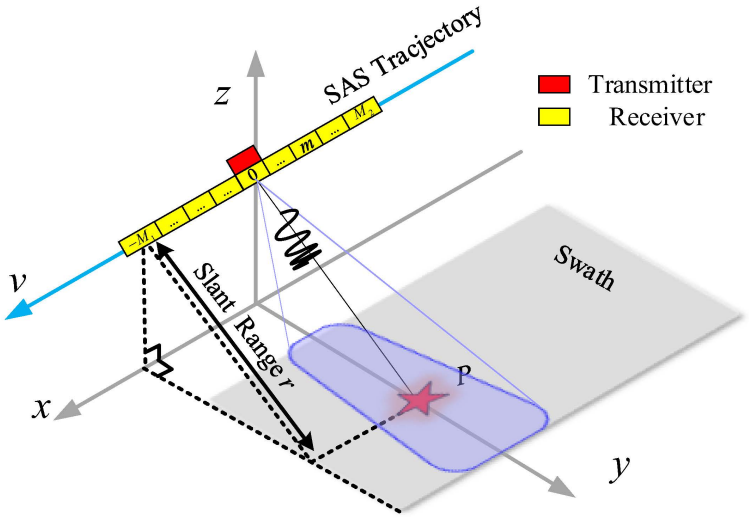


Figure 1. Geometric model of the multi-receiver SAS in the side-looking mode.

To facilitate the understanding of the imaging process, we draw an imaging geometric diagram of the slant range plane, as shown in Figure 2. We take the trajectory direction as the x-axis called the azimuth direction, also known as the along-track direction, and the direction perpendicular to the trajectory is the r-axis, which is called the range direction, also known as the cross-track direction. d_m is the distance from the m-th receiver to the reference receiver, and the position of the reference receiver coincides with that of the transmitter. The SAS platform insonifies the ideal scattering point $P(r, 0)$ represented as a red pentagram by moving in a uniform straight line at a speed v along the azimuth direction.

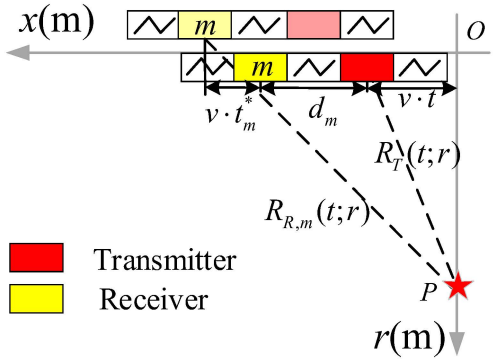


Figure 2. Schematic diagram in the slant plane.

The SAS platform emits chirp signals in a fixed *PRF* while moving along the azimuth direction. At time t, the position of the transmitter is $(vt, 0)$, represented as the solid red. Based on geometric relationships, the distance from the transmitter to the point target $P(r, 0)$ can be expressed as

$$R_T(t;r) = \sqrt{r^2 + (vt)^2}. \tag{1}$$

At time t, the position of the m-th receiver is $(vt + d_m, 0)$, represented as the solid yellow. Considering the NSHS time during signal propagation t_m^*, the azimuth movement of the SAS platform during this time is vt_m^*. It should be noted that Figure 2 is only a schematic diagram, and vt_m^* is very small in reality. The true position of the m-th receiver receiving the signal is $(vt + vt_m^* + d_m, 0)$, represented as the light yellow, so the distance from the point target $P(r, 0)$ to the m-th receiver can be expressed as

$$R_{R,m}(t;r) = \sqrt{r^2 + (vt + vt_m^*(t;r) + d_m)^2}. \tag{2}$$

According to (1) and (2), the round-trip range history of the m-th receiver considering NSHS time can be expressed as

$$R_m^*(t;r) = R_T(t;r) + R_{R,m}(t;r) = \sqrt{r^2 + (vt)^2} + \sqrt{r^2 + (vt + vt_m^*(t;r) + d_m)^2}. \tag{3}$$

Meanwhile, during the NSHS time t_m^*, the distance traveled by sound waves from emission to reception by the m-th receiver can be expressed as

$$R_m^*(t;r) = c \cdot t_m^* \tag{4}$$

where c is a constant speed of sound. The distance derived from geometric relationships and the propagation distance of sound waves is equal, so (3) and (4) can be combined, and the exact round-trip NSHS time can be obtained as

$$t_m^*(t;r) = (B + \sqrt{B^2 + AC})/A \tag{5}$$

where A, B, and C are represented as follows:

$$\begin{cases} A = c^2 - v^2 \\ B = v(vt + d_m) + c\sqrt{(vt)^2 + r^2} \\ C = {d_m}^2 + 2vtd_m \end{cases} \tag{6}$$

Substituting (5) and (6) into (4), we can obtain the accurate range history.

2.2. The Range History Shifting Method

However, the accurate range history is too complex to be conducive to subsequent processing. We use the receiver numbered 0 as the reference receiver where $d_m = 0$.

Substituting $d_m = 0$ into (4) and the accurate range history of the reference receiver is shown in the red curve in Figure 3 and can be expressed as

$$R_0^*(t;r) = \frac{2v^2ct + 2c^2\sqrt{(vt)^2 + r^2}}{c^2 - v^2}. \tag{7}$$

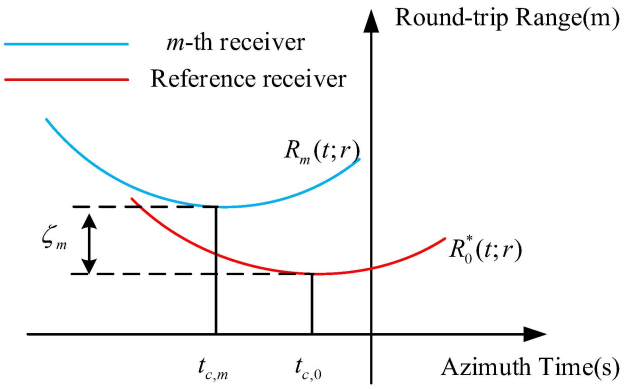


Figure 3. Range migration curves of different receivers.

Let $\gamma_0 = c^2/(c^2 - v^2)$ and $\eta_0 = v^2/(c^2 - v^2)$; then, (7) can be expressed as

$$R_0^*(t;r) \approx 2\gamma_0\sqrt{(vt)^2 + r^2} + 2\eta_0 ct. \tag{8}$$

The beam center's time is the time when the slant range history is minimum. Let $dR_0^*(t;r)/dt = 0$ we can obtain the beam center's time as follows:

$$t_{c,0} = -\frac{r}{\sqrt{c^2 - v^2}} \approx -\frac{r}{c}. \tag{9}$$

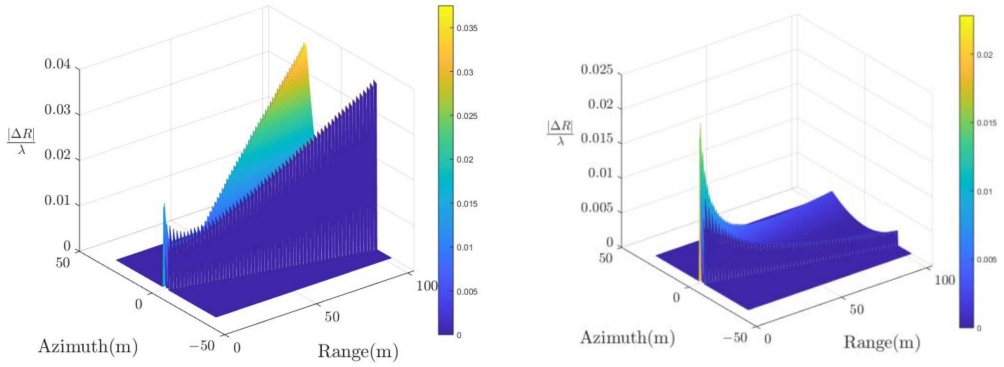
The sampling interval between the m-th receiver and the reference receiver is $d_m/2$ [61]. When the reference receiver samples at $t_{c,0}$, the equivalent sampling time of the m-th receiver is $t_{c,m} = t_{c,0} - d_m/(2v)$. If the curvature's difference of the RCM curves of different receivers is ignored, then the RCM curve of the m-th receiver is regarded as the translation of the reference receiver in the slant plane, with the translation amounts on the t-axis and r-axis being $d_m/2$ and ζ_m, respectively, as shown by the blue curve in Figure 3. Based on the geometric relationship, $\zeta_m(r)$ can be expressed as

$$\zeta_m(r) = R_m^*(t;r) - R_0^*(t + d_m/(2v);r) = R_m^*(t_{c,m};r) - R_0^*(t_{c,0};r) \tag{10}$$

According to (4), (5), (6), and (8), the analytical expressions of $\zeta_m(r)$ can be expressed as

$$\begin{aligned}\zeta_m(r) \quad = & \frac{c\left(+\sqrt{(c^2-v^2)\left(2v\left(-\frac{r}{c}-\frac{d_m}{2v}\right)d_m+{d_m}^2\right)+\left(v\left(v\left(-\frac{r}{c}-\frac{d_m}{2v}\right)+d_m\right)+c\sqrt{v^2\left(-\frac{r}{c}-\frac{d_m}{2v}\right)^2+r^2}\right)^2}\right)}{c^2-v^2} \\ & + \frac{c\left(v\left(v\left(-\frac{r}{c}-\frac{d_m}{2v}\right)+d_m\right)+c\sqrt{v^2\left(-\frac{r}{c}-\frac{d_m}{2v}\right)^2+r^2}\right)}{c^2-v^2} - 2\gamma_0\sqrt{\frac{v^2r^2}{c^2}+r^2} + 2\eta_0 r\end{aligned} \tag{11}$$

The round-trip range history can be obtained by substituting $\zeta_m(r)$ into (10), and the complex accurate range history is transformed into the SSR form with the linear term and constant term, which facilitates the subsequent derivation of PTRS. Since is obtained by shifting reference range history, we call range history shifting (RHS) method. The error compared with the accurate range history is ΔR. Figure 4 shows ΔR in whole swath using the parameters shown in Table 1.



(**a**) (**b**)

Figure 4. Range history errors. (**a**) Range history error of the PCA method. (**b**) Range history error of the RHS.

Table 1. SAS's system parameters.

Parameters	Values
Carrier Frequency	28 kHz
Chirp Bandwidth	16 kHz
Pulse duration	4 ms
PRI	0.15 s
Transmitter Size	0.102 m
Receiver Size	0.0765 m
Number of Receivers	9
SAS Platform Velocity	2.3 m/s
Swath	[5 m, 100 m]

From Figure 4, it can be seen that although the range history errors of RHS and PCA methods have not exceeded $\lambda/8$, the range history error of the RHS method does not widen with the beam, and the RHS method has a wider range of applications.

2.3. Derivation of the PTRS

The baseband form of the echo signal of target P received by the m-th receiver can be represented as

$$ss_{\mathrm{bb},m}(\tau,t;r)=A_0\cdot w_r\left(\tau-\frac{R_m(t;r)}{c}\right)\cdot w_a(t)\times\exp\left(-j2\pi f_c\frac{R_m(t;r)}{c}\right)\times\exp\left(j\pi K_r\left(\tau-\frac{R_m(t;r)}{c}\right)^2\right) \tag{12}$$

To obtain the phase of the PTRS of the echo signal, we perform range FFT on (12), and obtain the expression of the range spectrum as

$$Ss_m(f_r,t;r)=\int_{-\infty}^{+\infty}ss_{\mathrm{bb},m}(\tau,t;r)\exp(-j2\pi f_r\tau)d\tau \tag{13}$$

The phase in the integral expression of (13) is

$$\varphi_{Ss,m}(\tau,f_r,t;r)=\pi K_r\left(\tau-\frac{R_m(t;r)}{c}\right)^2-2\pi f_c\frac{R_m(t;r)}{c}-2\pi f_r\tau \tag{14}$$

According to the principle of stationary phase [62] (PSP), let $d\varphi_{Ss,m}(\tau,f_r,t;r)/d\tau=0$, the stationary phase time in the range direction can be obtained as

$$\tau_{\mathrm{PSP},m}(f_r,t;r)=\frac{R_m(t;r)}{c}+\frac{f_r}{K_r} \tag{15}$$

Bringing $\tau_{\mathrm{PSP},m}(f_r,t;r)$ into $\varphi_{Ss,m}(\tau,f_r,t;r)$, we can obtain

$$\varphi_{Ss,m}(f_r,t;r) = -\pi\frac{f_r^2}{K_r} - 2\pi\frac{(f_c+f_r)}{c}R_m(t;r) \tag{16}$$

Based on (16), the range spectrum can be rewritten as

$$Ss_m(f_r,t;r) = W_r(f_r)w_{a,m}(t)\exp\left(-j\pi\frac{f_r^2}{K_r}\right)\times\exp\left(-2\pi\frac{(f_c+f_r)}{c}R_m(t;r)\right). \tag{17}$$

where $W_r(f_r) = w_r(f_r/K_r)$.

The PTRS can be obtained by performing azimuth FFT, and the PTRS is expressed as

$$SS_m(f_r,f_a;r) = \int_{-\infty}^{+\infty} Ss_m(f_r,t;r)\exp(-j2\pi f_a t)dt \tag{18}$$

Substituting (17) into (18), the phase in the integral is

$$\theta_m(f_r,f_a;r,t) = -\pi\frac{f_r^2}{K_r} - 2\pi\frac{f_r+f_c}{c}\left(\zeta_m + 2\gamma_0\sqrt{\left(v\left(t+\frac{d_m}{2v}\right)\right)^2 + r^2} + 2\eta_0 c\left(t+\frac{d_m}{2v}\right)\right). \tag{19}$$

Using PSP, let $\partial\theta_m(f_r,f_a;r,t)/\partial t = 0$, and the azimuth phase stationary time can be expressed as

$$t_{\mathrm{PSP},m} = -\frac{cr((f_c+f_r)\eta_0 + f_a/2)}{v^2\gamma_0(f_c+f_r)\cdot D_{2\mathrm{df}}} - \frac{d_m}{2v} \tag{20}$$

where $D_{2\mathrm{df}}(f_r,f_a)$ is the range migration factor, expressed as

$$D_{2\mathrm{df}}(f_r,f_a) = \sqrt{1 - \frac{((2\eta_0(f_c+f_r)+f_a)c)^2}{4v^2\gamma_0^2(f_c+f_r)^2}}. \tag{21}$$

Bringing $t_{\mathrm{PSP},m}$ into (18), we can obtain the analytical expression of the PTRS, expressed as

$$SS_m(f_r,f_a;r) = W_r(f_r)W_{a,m}(f_a)\times\exp\left(+j\left(-\frac{4\pi r(f_c+f_r)D_{2\mathrm{df}}}{c} + \frac{\pi d_m f_a}{v} - \frac{\pi f_r^2}{K_r} - \frac{2\pi(f_c+f_r)\zeta_m}{c}\right)\right) \tag{22}$$

where $W_{a,m}(f_a)$ is the envelope of Doppler frequency. D_T is the transmitter size, and D_R is the receiver size. Perform Taylor expansion on (22) at $f_r = 0$ and preserve it to the quadratic term, we can obtain

$$\begin{aligned} SS_m(f_r,f_a;r) \quad &= W_r(f_r)W_{a,m}(f_a)\times\exp\left(-j2\pi\tfrac{f_c+f_r}{c}\zeta_m\right)\times\exp\left(+j\tfrac{\pi f_a d_m}{v}\right)\times\exp\left(j\varphi_{\mathrm{RC}}f_r^2\right) \\ &\times\exp\left(j\left(\varphi_{\mathrm{AC},m}(r,f_a) + \varphi_{\mathrm{RCM},m}(r,f_a)f_r + \varphi_{\mathrm{SRC},m}(r,f_a)f_r^2\right)\right)\times\exp(j\varphi_{\mathrm{HOP},m}(f_r,f_a;r)) \end{aligned} \tag{23}$$

where φ_{RC} represents the coefficient of range compression, $\varphi_{\mathrm{AC},m}(r,f_a)$ represents the azimuth compression term, $\varphi_{\mathrm{RCM},m}(r,f_a)$ represents the coefficient of the RCM term, and $\varphi_{\mathrm{SRC},m}(r,f_a)$ represents the coefficient of the SRC term, $\varphi_{\mathrm{HOP},m}(f_r,f_a;r)$ represents the higher order phase (HOP), which cannot be ignored in severe azimuth range coupling. The above coefficients can be expressed as follows:

$$\begin{cases} \varphi_{\mathrm{RC}} = -\pi / K_r \\ \varphi_{\mathrm{AC},m}(r, f_a) = -\frac{4\pi r \gamma_0 f_c D_a}{c} \\ \varphi_{\mathrm{RCM},m}(r, f_a) = -\frac{4\pi r \gamma_0 \mu_1}{c D_a} \\ \varphi_{\mathrm{SRC},m}(r, f_a) = -\frac{2\pi r \gamma_0 \left(\mu_2 D_a^2 - \mu_1^2\right)}{c f_c D_a^3} \\ \varphi_{\mathrm{HOP},m}(f_r, f_a; r) = \frac{4\pi r \gamma_0 f_c}{c}\left(D_a + \frac{\mu_1}{D_a}\frac{f_r}{f_c} + \left(\frac{\mu_2 D_a^2 - \mu_1^2}{2{D_a}^3}\right)\frac{f_r^2}{f_c^2}\right) - \frac{4\pi r \gamma_0 (f_c + f_r) D_{2\mathrm{df}}}{c}. \end{cases} \tag{24}$$

where μ_1, μ_2 and $D_a(f_a)$ are expressed as follows:

$$\begin{cases} \mu_1 = 1 - \frac{(2 f_c \eta_0 + f_a)\eta_0 c^2}{2 v^2 {\eta_0}^2 f_c} \\ \mu_2 = 1 - \frac{\eta_0 c^2}{v^2 {\gamma_0}^2} \\ D_a(f_a) = \sqrt{1 - \frac{((2\eta_0 f_c + f_a)c)^2}{4 v^2 \gamma_0^2 {f_c}^2}} \end{cases} \tag{25}$$

3. The Subblock–Subband CZT Algorithm

3.1. Monostatic Conversion

According to the first exponential term in (23), it can be seen that ζ_m will cause envelope offset and phase error, and the phase compensation function is

$$H_{\mathrm{phase},m}(r) = \exp\left(+j2\pi \frac{f_c}{c}\zeta_m(r)\right) \tag{26}$$

Assuming the transmitter is located in the middle of the array, the curves of $\zeta_m(r)$ changing with r are shown in Figure 5 based on the parameters in Table 1. From Figure 5, it can be seen that the variation of ζ_m exceeds one range resolution unit, so range subblocks processing or interpolation [60] is needed to ensure that the variation of ζ_m in each range subblock is less than one range resolution unit. Therefore, for the n-th subblock, $r_{c,n}$ is the center of the subblock, and $\zeta_m(r_{c,n})$ can be used to perform envelope offset (delay) compensation. The corresponding compensation function is

$$H_{\mathrm{delay},n,m}(f_r) = \exp\left(+j2\pi \frac{f_r}{c}\zeta_m(r_{c,n})\right) \tag{27}$$

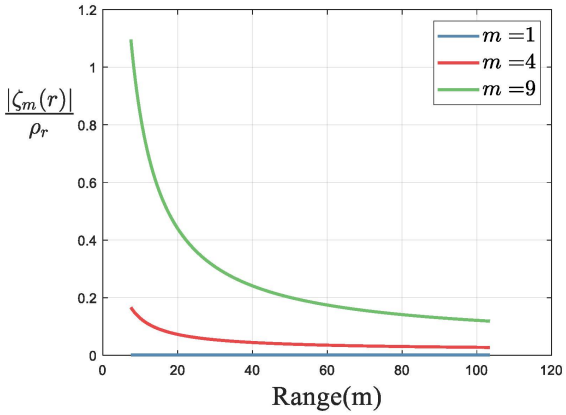


Figure 5. Curves of $\zeta_m(r)$ changing with r.

The second exponential term in (23) is the bistatic deformation term, which is due to the different positions of the transmit/receive arrays in multi-receiver SAS. According to the time shifting property of FFT, this term will cause the data of the m-th receiver to shift in the azimuth axis by an amount of $d_m/2$. Therefore, we can offset the impact of this term by rearranging the data in the time domain at intervals of $d_m/2$, which is equivalent to compensating for the bistatic deformation term.

After compensating for the receiver-dependent terms, the monostatic conversion has been completed, and the converted signal is represented as

$$\begin{aligned} SS_m(f_r, f_a; r) &= W_r(f_r)W_{a,m}(f_a) \times \exp\left(j\varphi_{\mathrm{RC}} f_r^2\right) \times \exp(j\varphi_{\mathrm{HOP},m}(f_r, f_a; r)) \\ &\times \exp\left(j\left(\varphi_{\mathrm{AC},m}(r, f_a) + \varphi_{\mathrm{RCM},m}(r, f_a) f_r + \varphi_{\mathrm{SRC},m}(r, f_a) {f_r}^2\right)\right) \end{aligned} \tag{28}$$

3.2. Range Time Domain Subblocks

Firstly, perform range compression with a compensation filter of $H_{\mathrm{rc}}(f_r) = -\varphi_{\mathrm{RC}} f_r^2 = \pi f_r^2 / K_r$, and the range compressed signal is

$$\begin{aligned} SS_m(f_r, f_a; r) &= W_r(f_r)W_{a,m}(f_a) \times \exp(j\varphi_{\mathrm{HOP},m}(f_r, f_a; r)) \exp(j\varphi_{\mathrm{AC},m}(r, f_a)) \\ &\times \exp\left(j\left(\varphi_{\mathrm{RCM},m}(r, f_a) f_r + \varphi_{\mathrm{SRC},m}(r, f_a) {f_r}^2\right)\right) \end{aligned} \tag{29}$$

If $\varphi_{\mathrm{HOP},m}(f_r, f_a; r)$ is discarded, it is equivalent to ignoring azimuth range coupling terms. The high-order phase error increases with the increase of range. When the HOP term is equal to $\pi/4$, the corresponding range is the maximum allowable imaging range, also known as the high-resolution imaging range (relative to the starting point of the swath). The larger the relative bandwidth, the smaller the high-resolution imaging range. In low-frequency broadband signals, the azimuth range coupling is very severe, severely limiting the high-resolution imaging range. If the HOP term at the center of the swath is compensated, then the high-order phase at the center of the swath is zero, which can be fully focused. The maximum allowable imaging range is transformed into the maximum permissible imaging range relative to the center of the swath, which is half the swath. However, the HOP terms at both edges of the swath may still exceed $\pi/4$. To solve the above problem, we divide the range into many subblocks in the time domain, and the center of each subblock is used as a reference range to compensate for the HOP, ensuring that the HOP terms at both edges of each range subblock are less than $\pi/4$.

The total number of sampling points in the time domain is N_r. If the range is uniformly divided into P subblocks in the time domain, and the number of points in each block is $N_P = N_r/P$, then the total number of sampling points in the time domain is expressed as

$$l = pN_P + n \tag{30}$$

where p is the range time domain subblock subscript, and p can be expressed as

$$p = -P/2, -P/2+1, \ldots, P/2-1 \tag{31}$$

Assuming the center reference range of the p-th subblock is r_p, the range within the p-th subblock can be discretized as

$$r = r_p + (n - N_P/2) \cdot c/(2f_s) \tag{32}$$

where f_s is range sampling frequency and n represents the range time domain sampling point in the p-th subblock and can be expressed as

$$n = 0, 1, \ldots, N_P - 1 \tag{33}$$

The HOP compensation is applied to the p-th subblock's center, and the expression for the HOP compensation is

$$SS_{\mathrm{HOP},m}(f_r, f_a; r) = SS_m(f_r, f_a; r) \times \exp(-j\varphi_{\mathrm{HOP},m}(f_r, f_a; r_p)) \tag{34}$$

The PTRS after the HOP compensation is

$$\begin{aligned} SS_{\mathrm{HOP},m}(f_r, f_a; r) &= W_r(f_r)W_a(f_a) \times \exp(j\Delta\varphi_{\mathrm{HOP},m}(f_r, f_a; r)) \times \exp(j\varphi_{\mathrm{AC},m}(r, f_a)) \\ &\times \exp\left(j\left(\varphi_{\mathrm{RCM},m}(r, f_a)f_r + \varphi_{\mathrm{SRC},m}(r, f_a){f_r}^2\right)\right) \end{aligned} \tag{35}$$

where $\Delta\varphi_{\mathrm{HOP},m}(f_r, f_a; r)$ represents the residual HOP error after HOP compensation, expressed as

$$\Delta\varphi_{\mathrm{HOP},m}(f_r, f_a; r) = \varphi_{\mathrm{HOP},m}(f_r, f_a; r) - \varphi_{\mathrm{HOP},m}(f_r, f_a; r_p) \tag{36}$$

As the higher order Taylor expands, $\Delta\varphi_{\mathrm{HOP},m}(f_r, f_a; r)$ becomes smaller but the expression becomes more complex. It can be seen that $\Delta\varphi_{\mathrm{HOP},m}(f_r, f_a; r_p) = 0$, which means at the reference range r_p, the residual HOP is zero, and the image at the reference range can be completely focused. The $\varphi_{\mathrm{RCM},m}(r, f_a)$, $\varphi_{\mathrm{RCM},m}(r, f_a)$ and $\varphi_{\mathrm{Cubic},m}(r, f_a)$ are expanded at $r = r_p$ and kept as first-order terms for subsequent CZT processing. This is equivalent to the azimuth-frequency range decoupling (ARD), and the phase expression is obtained as follows:

$$\begin{aligned} SS_{\mathrm{ARD},m}(f_r, f_a; r) &= W_r(f_r)W_a(f_a)\exp(+j\varphi_{\mathrm{AC},m}(r, f_a)) \\ &\times \exp\left(+j\left(\varphi_{\mathrm{RCM},m}(r_p, f_a) + \left.\frac{d\varphi_{\mathrm{RCM},m}(r,f_a)}{dr}\right|_{r=r_p}(r - r_p)\right)f_r\right) \\ &\times \exp\left(+j\left(\varphi_{\mathrm{SRC},m}(r_p, f_a) + \left.\frac{d\varphi_{\mathrm{SRC},m}(r,f_a)}{dr}\right|_{r=r_p}(r - r_p)\right){f_r}^2\right) \\ &\times \exp(+j\Delta\varphi_{\mathrm{HOP},m}(f_r, f_a; r)) \times \exp(+j\Delta\varphi_{\mathrm{ARD},m}(f_r, f_a; r)) \end{aligned} \tag{37}$$

where $\Delta\varphi_{\mathrm{ARD},m}(f_r, f_a; r)$ is the high-order phase term of the Taylor expansion at $r = r_p$ for $\varphi_{\mathrm{RCM},m}(r, f_a)$, $\varphi_{\mathrm{RCM},m}(r, f_a)$, and $\varphi_{\mathrm{Cubic},m}(r, f_a)$, and can be expressed as

$$\begin{aligned} \Delta\varphi_{\mathrm{ARD},m}(f_r, f_a; r) &= \varphi_{\mathrm{RCM},m}(r, f_a)f_r - \left(\varphi_{\mathrm{RCM},m}(r_p, f_a) + \left.\frac{d\varphi_{\mathrm{RCM},m}(r,f_a)}{dr}\right|_{r=r_p}(r - r_p)\right)f_r \\ &+\varphi_{\mathrm{SRC},m}(r, f_a){f_r}^2 - \left(\varphi_{\mathrm{SRC},m}(r_p, f_a) + \left.\frac{d\varphi_{\mathrm{SRC},m}(r,f_a)}{dr}\right|_{r=r_p}(r - r_p)\right)f_r^2 \end{aligned} \tag{38}$$

Perform bulk compensation on (37) by removing range-independent terms from the first and second order coefficients of f_r, so that the RCM and SRC terms at the center r_p of each subblock are zero, and the remaining terms are the residual range variant RCM and SRC terms. The phase multiplication factor is

$$\varphi_{\mathrm{Bulk},m}(f_r, f_a; r) = -\varphi_{\mathrm{RCM},m}(r_p, f_a)f_r - \varphi_{\mathrm{SRC},m}(r_p, f_a)f_r^2 \tag{39}$$

After the above bulk compensation, the compensated PTRS is expressed as

$$\begin{aligned} SS_{\mathrm{Bulk},m}(f_r, f_a; r) &= W_r(f_r)W_a(f_a) \times \exp(+j\varphi_{\mathrm{AC},m}(r, f_a)) \\ &\times \exp\left(+j\left.\frac{d\varphi_{\mathrm{RCM},m}(r,f_a)}{dr}\right|_{r=r_p}(r - r_p)f_r\right) \\ &\times \exp\left(+j\left.\frac{d\varphi_{\mathrm{SRC},m}(r,f_a)}{dr}\right|_{r=r_p}(r - r_p){f_r}^2\right) \\ &\times \exp(+j\Delta\varphi_{\mathrm{HOP},m}(f_r, f_a; r)) \times \exp(+j\Delta\varphi_{\mathrm{ARD},m}(f_r, f_a; r)) \end{aligned} \tag{40}$$

At the center r_p of each subblock, the RCM term and SRC term are zero, and $\Delta\varphi_{\mathrm{HOP},m}(f_r, f_a; r)$ and $\Delta\varphi_{\mathrm{ARD},m}(f_r, f_a; r)$ increase with range relative to r_p. It is only necessary to ensure that there is always $|\Delta\varphi_{\mathrm{HOP},m}(f_r, f_a; r) + \Delta\varphi_{\mathrm{ARD},m}(f_r, f_a; r)| \leq \pi/4$ in each subblock for high-resolution imaging. However, it is difficult to meet the requirements of high-resolution imaging solely through range time domain subblocks, as the azimuth range coupling is more pronounced in the range frequency domain, and $|\Delta\varphi_{\mathrm{HOP},m}(f_r, f_a; r) + \Delta\varphi_{\mathrm{ARD},m}(f_r, f_a; r)|$ changes rapidly with f_r. Therefore, we need to divide the range frequency into subbands in each range time domain subblock.

3.3. Range Frequency Domain Subbands

If the number of sampling points is N_r, then the range frequency in the p-th subblock can be discretized as

$$f_r = k \cdot f_s / N_r \tag{41}$$

where the range frequency domain full sampling points can be expressed as

$$k = -N_r/2, -N_r/2 + 1, \ldots, N_r/2 - 1. \tag{42}$$

Based on (32) and (41), $PP_{\mathrm{Bulk},m}(f_r, f_a; r)$ can be discretized as

$$\begin{aligned} SS_{\mathrm{Bulk},m}(k, f_a; \quad & pN_P + n) = W_r\left(kf_s/N_Q\right) W_a(f_a) \times \exp(+j\varphi_{\mathrm{AC},m}(pN_P + n, f_a)) \\ & \times \exp\left(-u_m(f_a)\left(n - \tfrac{N_P}{2}\right)k - v_m(f_a)\left(n - \tfrac{N_P}{2}\right)k^2\right) \\ & \times \exp(+j\Delta\varphi_{\mathrm{HOP},m}(k, f_a; pN_P + n)) \times \exp(+j\Delta\varphi_{\mathrm{ARD},m}(k, f_a; pN_P + n)) \end{aligned} \tag{43}$$

where $u_m(f_a)$ and $v_m(f_a)$ are expressed as

$$\begin{cases} u_m(f_a) = -\frac{c}{2N_r} \cdot \left.\frac{d\varphi_{\mathrm{RCM},m}(r, f_a)}{dr}\right|_{r=r_p} \\ v_m(f_a) = -\frac{cf_s}{2N_r^2} \cdot \left.\frac{d\varphi_{\mathrm{SRC},m}(r, f_a)}{dr}\right|_{r=r_p} \end{cases}. \tag{44}$$

Next, the range frequency domain fully sampling point k is divided into subbands in the p-th subblock. The number of fully frequency sampled points is N_r and the range frequency is uniformly divided into Q subbands. The size of each subband is $N_Q = N_r/Q$, and k can be expressed as

$$k = qN_Q + i \tag{45}$$

where the subband subscript is $q = -\frac{Q}{2}, -\frac{Q}{2} + 1, \ldots, \frac{Q}{2} - 1$, and the frequency point in the q-th subband is $i = 0, 1, \ldots, N_Q - 1$. After dividing k into subbands, k^2 in (43) can be linearly approximated in each subband, which we refer to as piece-wise linear approximation (PLA). For each segment, the tangent at the center of the frequency subband is taken as the approximate value of the subband, so that the center can be fully compensated and there is no phase jump at the edge, as shown in Figure 6. The approximate expressions for k^2 is expressed as follows:

$$k^2 \approx \left(q + \frac{1}{2}\right)^2 {N_Q}^2 + 2\left(q + \frac{1}{2}\right) N_Q \left(i - \frac{N_Q}{2}\right). \tag{46}$$

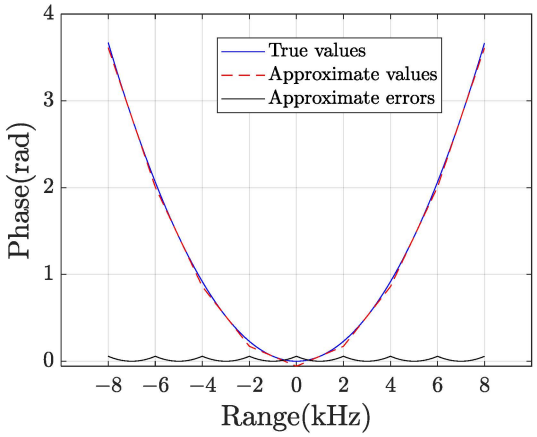


Figure 6. Schematic diagram of PLA.

After PLA, the PTRS corresponding to the q-th frequency domain subband in the p-th time domain subblock can be expressed as

$$\begin{aligned} PP_{\mathrm{PLA},m}(qN_Q+i,f_a;pN_P+n) &= W_r\left(kf_s/N_Q\right)W_a(f_a)\times\exp(+j\varphi_{\mathrm{AC},m}(pN_P+n,f_a)) \\ &\times\exp\left(-ju_m(f_a)(n-N_P/2)\left(qN_Q+i\right)\right) \\ &\times\exp\left(-jv_m(f_a)(n-N_P/2)\left(\begin{array}{l}(q+1/2)^2N_Q^2\\ +2(q+1/2)N_Q(i-N_Q/2)\end{array}\right)\right) \\ &\times\exp\left(j\Delta\varphi_{\mathrm{HOP},m}(qN_Q+i,f_a;pN_P+n)\right) \\ &\times\exp\left(j\Delta\varphi_{\mathrm{ARD},m}(qN_Q+i,f_a;pN_P+n)\right) \\ &\times\exp\left(j\Delta\varphi_{\mathrm{PLA},m}(qN_Q+i,f_a;pN_P+n)\right) \end{aligned} \quad (47)$$

where $\Delta\varphi_{\mathrm{PLA},m}$ is the phase error caused by PLA, and it can be expressed as

$$\Delta\varphi_{\mathrm{PLA},m}(qN_Q+i,f_a;pN_P+n)=v_m(f_a)\left(n-\frac{N_P}{2}\right)\left[\left(\left(q+\frac{1}{2}\right)^2N_Q^2+2\left(q+\frac{1}{2}\right)N_Q\left(i-\frac{N_Q}{2}\right)\right)-\left(qN_Q+i\right)^2\right] \quad (48)$$

Combining $\Delta\varphi_{\mathrm{PLA},m}$ with $\Delta\varphi_{\mathrm{HOP},m}$ and $\Delta\varphi_{\mathrm{ARD},m}$, the cumulative error is

$$\begin{aligned} \Delta\varphi_m(qN_Q+i,f_a;pN_P+n) &= \Delta\varphi_{\mathrm{HOP},m}(qN_Q+i,f_a;pN_P+n)+\Delta\varphi_{\mathrm{ARD},m}(qN_Q+i,f_a;pN_P+n) \\ &+\Delta\varphi_{\mathrm{PLA},m}(qN_Q+i,f_a;pN_P+n) \end{aligned} \quad (49)$$

Now, it is necessary to ensure that there is $\Delta\varphi_m(qN_Q+i,f_a;pN_P+n)\le\pi/4$ within each frequency domain subband. $\Delta\varphi_m(qN_Q+i,f_a;pN_P+n)$ is zero at r_p of the p-th subblock (corresponding sampling point $n_c=N_P/2$), and the maximum value at the edge $n_{\max}=N_P-1$ of the subblock is $\Delta\varphi_m(qN_Q+i,f_a;pN_P+N_P-1)$. For the PTRS at $n_{\max}$, PTRS is divided into Q subbands. If $q<0$, $\Delta\varphi_m(qN_Q+i,f_a;pN_P+N_P-1)$ reaches its maximum value $\Delta\varphi_m(qN_Q,B_a\left(qN_Q\right)/2;pN_P+N_P-1)$ in the q-th subband when the minimum range frequency $i=i_{\min}=0$ and the maximum azimuth frequency $f_a=f_{a,\max}=B_a\left(qN_Q\right)/2$ are taken. If $q\ge0$, $\Delta\varphi_m(qN_Q+i,f_a;pN_P+N_P-1)$ reaches its maximum value $\Delta\varphi_m(qN_Q+N_Q-1,B_a(qN_Q+N_Q-1)/2;pN_P+N_p-1)$ in the q-th subband when the range frequency and azimuth frequency are both maximized. It can be seen that the errors in different subbands within the p-th subblock are only related to N_P and N_Q. According to (45), Q subbands are traversed in the p-th subblock to obtain the maximum phase error, which can be expressed as

$$\Delta\varphi_m(k_m, B_a(k_m)/2; pN_P + N_P - 1) = \max\left\{ \begin{array}{l} \max\limits_{q=-\frac{Q}{2},-\frac{Q}{2}+1,\ldots,-1} \{\Delta\varphi_m(qN_Q, B_a(qN_Q)/2; pN_P + N_P - 1)\}, \\ \max\limits_{q=0,1,\ldots,\frac{Q}{2}-1} \{\Delta\varphi_m(qN_Q + N_Q - 1, B_a(qN_Q + N_Q - 1)/2; pN_P + N_p - 1)\} \end{array} \right\} \quad (50)$$

where k_m is the range frequency sampling point corresponding to the maximum value of $\Delta\varphi_m(k, f_a; pN_P + N_P - 1)$.

Next, according to (30), P subblocks are traversed to obtain the maximum phase error over the entire swath, which can be expressed as

$$\Delta\varphi_{\max,m}(k_m, B_a(k_m)/2; l_m) = \max_{p=-\frac{P}{2},-\frac{P}{2}+1,\ldots,\frac{P}{2}-1} \{\Delta\varphi_m(k_m, B_a(k_m)/2; pN_P + N_P - 1)\} \quad (51)$$

where l_m represents the range sampling point when $\Delta\varphi_m(k_m, B_a(k_m)/2; pN_P + N_P - 1)$ reaches its maximum value. The maximum error value $\Delta\varphi_m(k_m, B_a(k_m)/2; l_m)$ is only related to l_m and k_m, while l_m is only associated with and k_m is only related to N_Q. Based on $N_P = N_r/P$ and $N_Q = N_r/Q$, the curves of $\Delta\varphi_{\max,m}$ changing with P and Q can be obtained. Based on the parameters in Table 1, we obtain $\Delta\varphi_{\max,m}$ for different P and Q, as shown in Figure 7. It can be seen that the higher the PTRS is expanded, the smaller the number of time domain subblocks required. However, if the expanding order is too high, it will increase the algorithm complexity in program execution. If range frequency division is not performed (Q = 1), P needs to be large enough to meet $\Delta\varphi_{\max,m} \leq \pi/4$, which will cause a sharp increase in computational complexity. If P is fixed, $\Delta\varphi_{\max,m}$ will decrease with the growth of Q, but it cannot be infinitely reduced. Considering the computational complexity, we expand the PTRS to the quadratic phase term and take P = 16 and Q = 6 as the required division parameters for the SAS shown in Table 1.

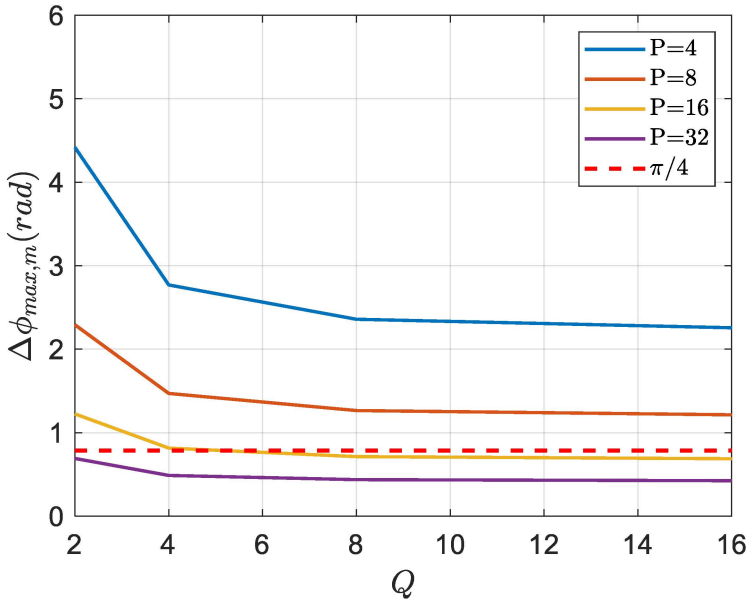


Figure 7. Curves of $\Delta\varphi_{\max,m}$ changing with P and Q.

3.4. The CZT Algorithm

After the above subblock–subband processing and PLA, if $\Delta\varphi_{\max,m} \leq \pi/4$ in each subband, the error shown in can be ignored, and the focusing process in each subband meets the requirements of high-resolution imaging. From here, it can be seen that there are no longer quadratic terms of range frequency, and the first-order term for range frequency can be corrected through RCMC. Therefore, the PLA process is equivalent to converting the residual SRC term to RCM terms. SCFT is used to avoid time-consuming interpolation

operations in RCMC and efficiently implemented in digital systems through CZT, which can effectively handle first-order spatially range cell migration [63]. After CZT processing, the RD domain data of the q-th frequency domain subband under the p-th time domain subblock is expressed as

$$pP_{q,m}(pN_P+n,f_a)=\sum_{i=0}^{N_Q-1}\left\{\begin{array}{l} PP_{\mathrm{PLA},m}(qN_Q+i,f_a;pN_P+n) \\ \times exp\left[j\left(\begin{array}{l} u_m(f_a)\cdot\left(n-\frac{N_P}{2}\right)(qN_Q+i) \\ +v_m(f_a)\left(n-\frac{N_P}{2}\right)\left(\left(q+\frac{1}{2}\right)^2N_Q^2+2\left(q+\frac{1}{2}\right)N_Q\left(i-\frac{N_Q}{2}\right)\right)\end{array}\right)\right]\end{array}\right\} \tag{52}$$

According to Bluestein's equation [64,65]

$$2ni=n^2+i^2-(n-i)^2 \tag{53}$$

we can rewrite (52) as

$$\begin{aligned} pP_m(pN_P+n,f_a) &= \exp\left(j\cdot g(qN_Q,f_a)\left(n-\tfrac{N_P}{2}\right)\right) \\ &\times\exp\left(j\cdot h(f_a)\cdot n^2\right)\sum_{i=0}^{N_Q-1}\left[\begin{array}{l} PP_{\mathrm{SLA},m}(qN_Q+i,f_a;pN_P+n) \\ \cdot\exp\left(-j\cdot h(qN_Q,f_a)N_Pi\right) \\ \cdot\exp\left(j\cdot h(qN_Q,f_a)i^2\right) \\ \cdot\exp\left(j\cdot h(qN_Q,f_a)\left(-(i-n)^2\right)\right)\end{array}\right]\end{aligned} \tag{54}$$

where $g(qN_Q,f_a)$ and $h(qN_Q,f_a)$ are, respectively, represented as

$$\left\{\begin{array}{l} g(qN_Q,f_a)=u_m(f_a)N_Qq+v_m(f_a)N_Q^2\left(q^2-\frac{1}{4}\right) \\ h(qN_Q,f_a)=\frac{u_m(f_a)}{2}+v_m(f_a)N_Q\left(q+\frac{1}{2}\right)^2\end{array}\right. \tag{55}$$

Equation (54) is equivalent to $PP_{\mathrm{PLA},m}(qN_Q+i,f_a;pN_P+n)$ multiplied by $\exp(-jh(qN_Q,f_a)N_Pi)$, then multiplied by $\exp\left(jh(qN_Q,f_a)i^2\right)$, and finally convolved with $\exp(-jh(qN_Q,f_a)i^2)$. To achieve higher computational efficiency, frequency domain convolution can generally be achieved through time domain multiplication, and finally multiplied by $\exp\left(j\cdot g(qN_Q,f_a)\left(n-\frac{N_P}{2}\right)\right)\times\exp\left(j\cdot h(f_a)\cdot n^2\right)$ to obtain the result after RCMC and SRC. The quick implementation of CZT is shown in Figure 8.

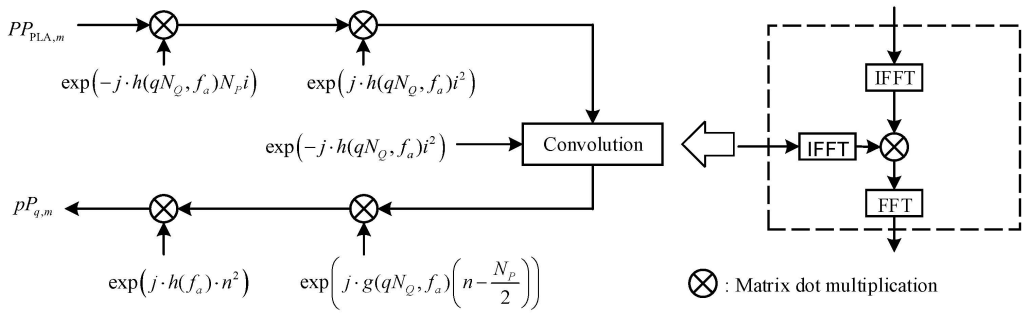


Figure 8. Quick implementation of CZT.

The bandwidth of each subband is one-Nth of the total range chirp bandwidth, so the range resolution is only one-Nth of the original signal. Therefore, the result obtained in this step is called a low-resolution result. Coherently summing Q low-resolution results can obtain a high-resolution result.

$$pP_m(pN_P+n,f_a)=\sum_{q=-Q/2}^{Q/2-1}pP_{q,m}(pN_P+n,f_a)=w_r\left(\tau-\frac{(pN_P+n)}{f_s}\right)\cdot W_a(f_a)\cdot\exp(+j\varphi_{\mathrm{AC},m}(pN_P+n,f_a)) \tag{56}$$

where $w_r(\cdot)$ represents the envelope in the range time domain, a sinc function. $(pN_P + n)/f_s$ represents the time corresponding to the true range coordinate of the target and is independent on f_a, indicating that the RCM has been corrected.

Concatenating data along the range axis in the RD domain to obtain data of the whole swath, the expression of the concatenated signal is

$$pP_m(l, f_a) = \underset{p=-\frac{P}{2},-\frac{P}{2}+1,\frac{P}{2}-1}{CAT}[pP_m(pN_P + n, f_a)] = w_r\left(\tau - \frac{l}{f_s}\right) \cdot W_a(f_a) \cdot \exp(j\varphi_{\mathrm{AC},m}(l, f_a)) \quad (57)$$

where $CAT[\cdot]$ represents the range time domain concatenation function. Azimuth compression is performed with a phase multiplication factor of $-\varphi_{\mathrm{AC},m}(l, f_a)$. The compressed signal in the azimuth direction is subjected to azimuth IFFT to obtain the 2-D time domain focusing result of the m-th receiver. Finally, $(M_1 + M_2 + 1)$ receivers' azimuth-aliased results are coherently summed to form the final high-resolution result, and the detailed flowchart of the proposed method is shown in Figure 9.

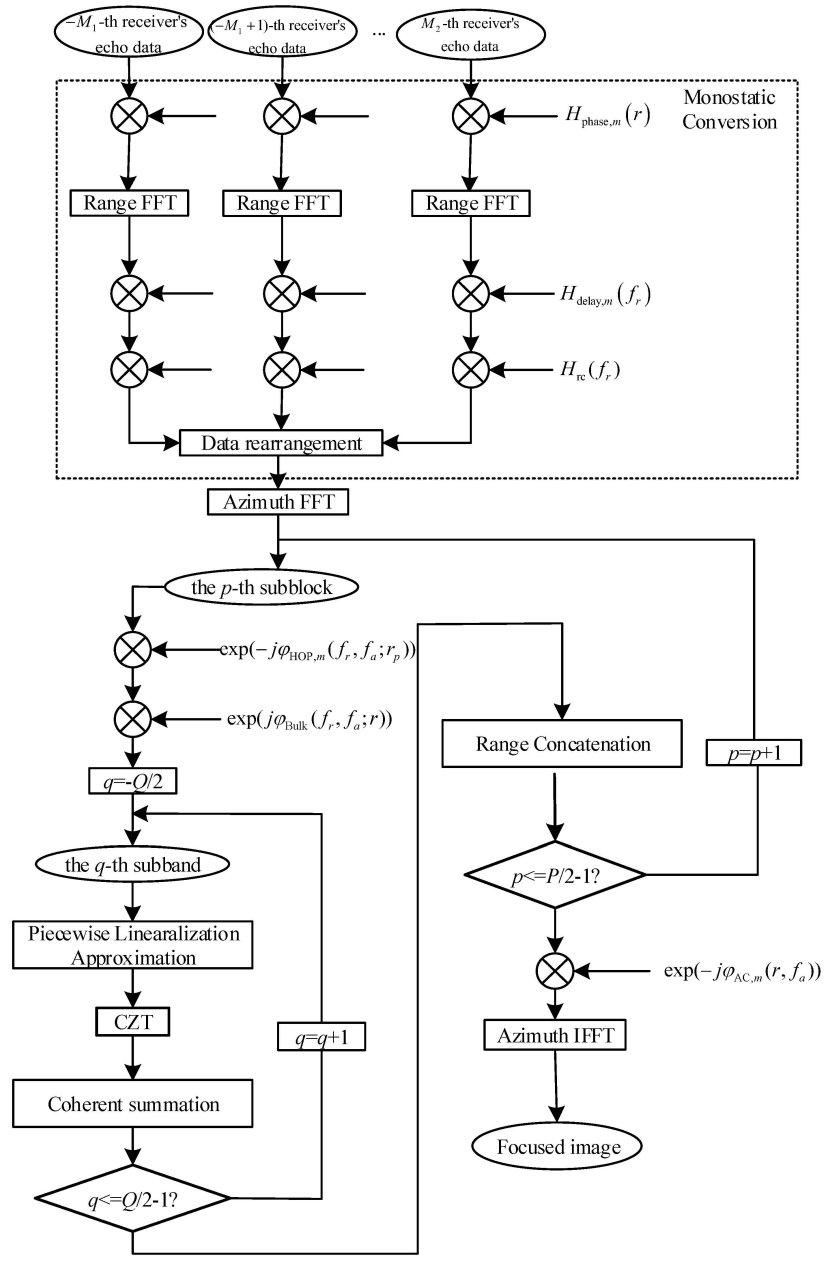


Figure 9. The flowchart of the proposed method.

4. Simulation and Experimental Results

4.1. Computer Simulation

To verify the proposed method's superiority, we conducted computer simulation. The imaging geometry is shown in Figures 1 and 2, and the SAS system parameters are shown in Table 1. Point targets represented as the red dots are located as shown in Figure 10. The computer CPU used in the simulation experiment is Intel i7-10700@2.90 GHz, RAM is 32 GB, and MATLAB version is R2020A. We used the BPA [66,67] as the evaluation criterion and compared the range time domain subblocks CZT algorithm based on the RHS, called the subblock CZT method [60], and the range time domain subblocks/frequency domain subbands algorithm based on the RHS, called the subblock–subband CZT method. The imaging results of different methods are shown in Figure 11. To display the differences of different methods more clearly, range and azimuth cross-sections of P1 (0 m, 20 m), P2 (0 m, 50 m), and P3 (0 m, 80 m) are shown in Figures 12–14 without any windowing processing [68]. From range and azimuth cross-sections, the point target imaging result of the proposed method is similar to that of BPA, both of which are superior to the subblock CZT method that only uses subblock processing. The quantitative analysis can be found in the following text.

For quantitative analysis, PSLRs, ISLRs, and IRWs of P1, P2 and P3 are measured, and the results are shown in Table 2. For the measurement results in the range direction, since the range envelope after pulse compression is a standard sinc function, the theoretical values of PSLR and ISLR in the range direction are −13.26 dB and −9.73 dB, respectively, and the IRW's theoretical value (−4 dB) is 4.69 cm. Due to the main-lobe's widening and the fusion of the first side-lobe into the main lobe [41], PSLRs may exceed the theoretical value, causing that the PSLR and ISLR of the second side lobe were actually measured. In addition, unaliased frequency recovery is achieved by the Nyquist sampling frequency, but to attain undistorted amplitude recovery, a higher frequency than the Nyquist sampling frequency is required [69]. Due to the limited sampling frequency, there may be distortion in the envelope amplitude, resulting in a gap between PSLR and theoretical values. However, IRW does not experience distortion. Therefore, we mainly use IRW as the evaluation indicator for the focusing effect. For the azimuth measurement results, since the azimuth envelope after pulse compression is the product of two sinc functions and then a weighted average of multiple receivers' direction maps. The theoretical values of PSLR and ISLR are challenging to acquire, and we directly use the PSLR and ISLR of BPA's results as the evaluation criterion.

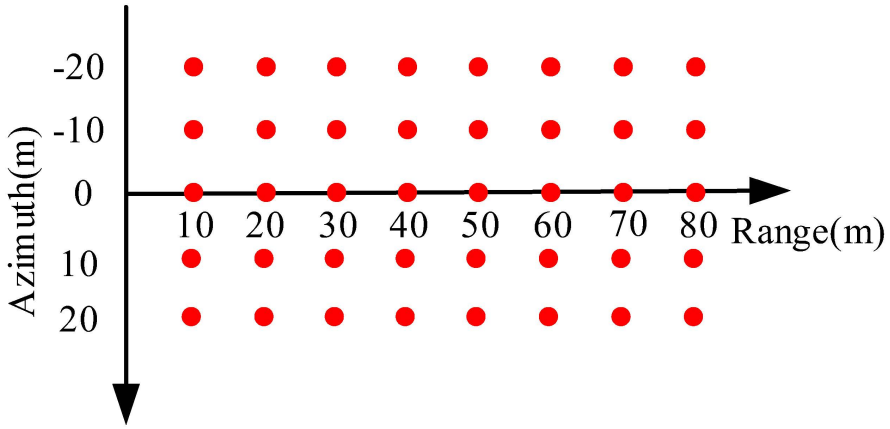


Figure 10. Location of simulation point targets.

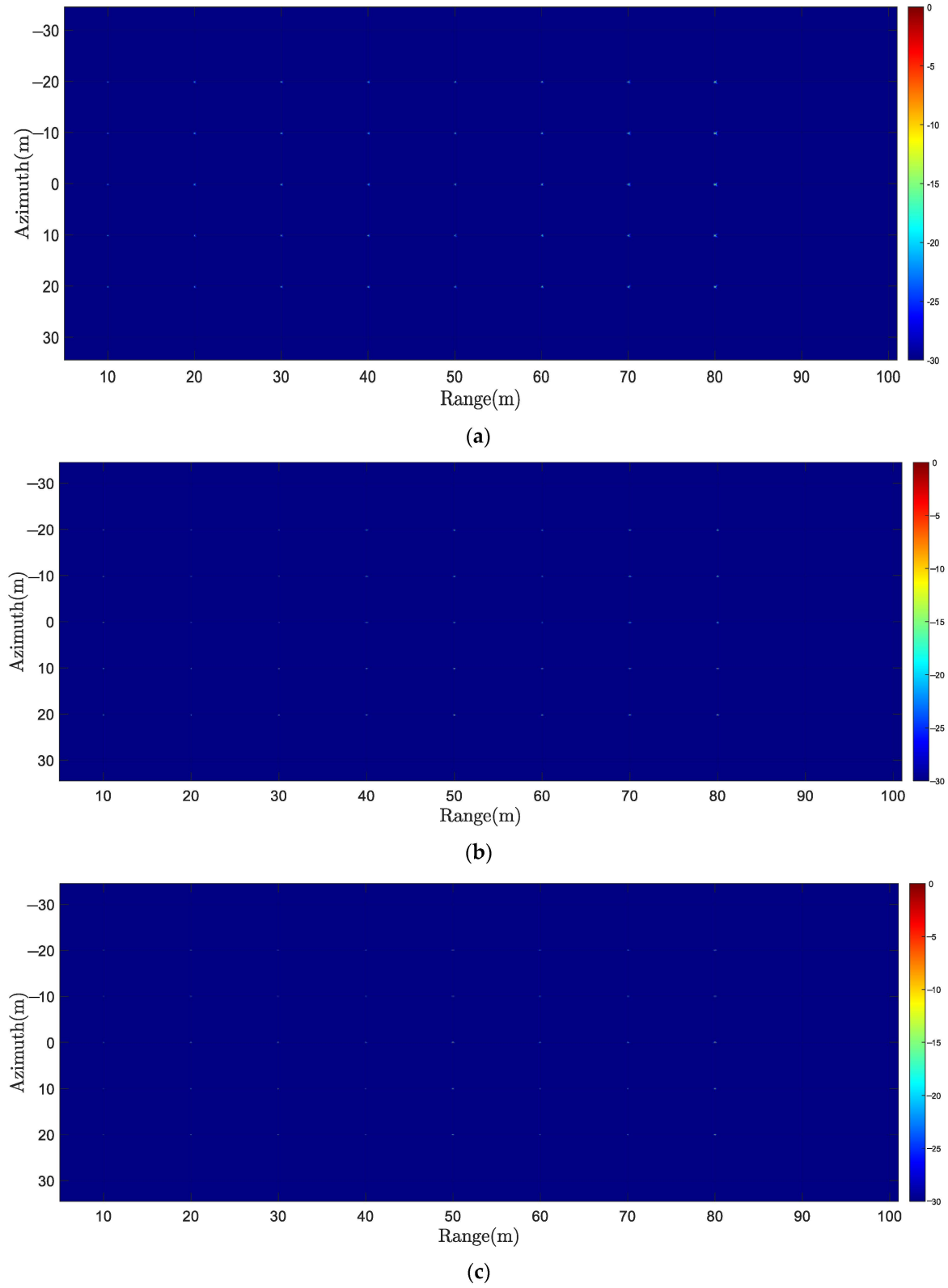


Figure 11. The imaging results of different methods. (**a**) The subblock CZT method. (**b**) The subblock–subband CZT method. (**c**) BPA.

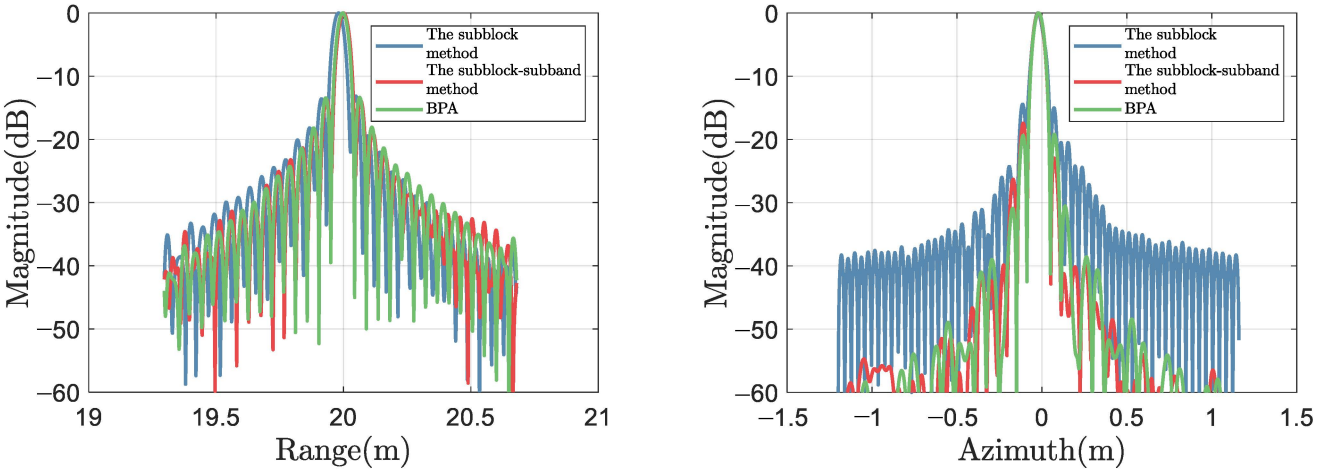


Figure 12. Cross-sections of the point target P1.

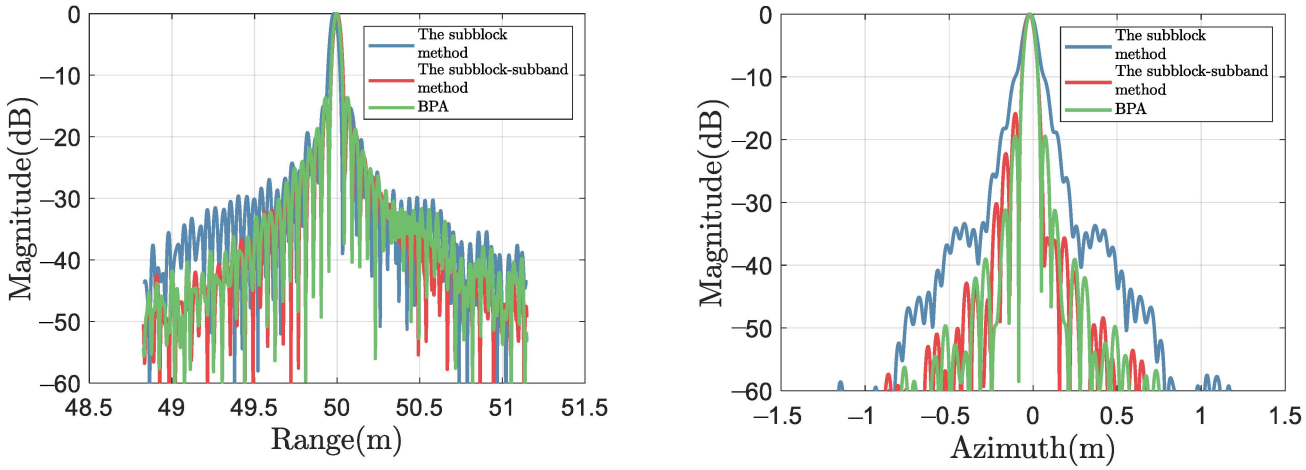


Figure 13. Cross-sections of the point target P2.

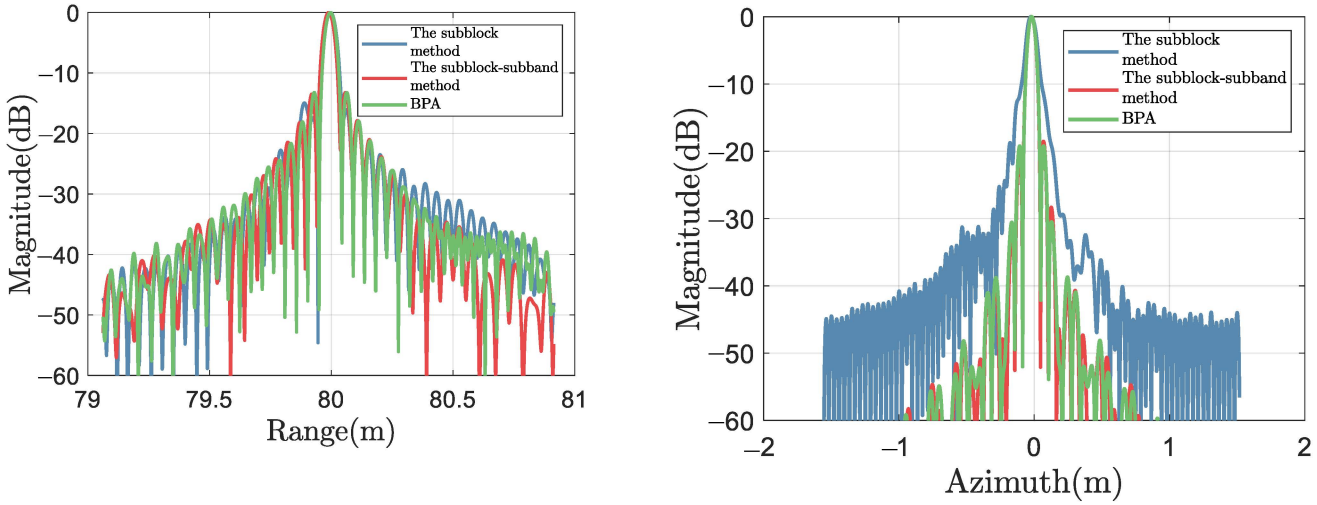


Figure 14. Cross-sections of the point target P3.

Table 2. Measured values of different point targets.

Point Targets	Methods	Range Cross-Section			Azimuth Cross-Section		
		PSLR (dB)	**ISLR (dB)**	**IRW (cm)**	**PSLR (dB)**	**ISLR (dB)**	**IRW (cm)**
P1 (0 m, 20 m)	The subblock CZT method	−13.14	−10.71	4.80	−14.44	−13.10	5.50
	The subblock–subband CZT method	−13.17	−10.42	4.76	−17.42	−18.94	5.74
	BPA	−13.27	−10.49	4.75	−19.2	−18.88	5.77
P2 (0 m, 50 m)	The subblock CZT method	−15.07	−8.58	5.09	−18.21	−18.33	7.63
	The subblock–subband CZT method	−13.17	−9.7	4.89	−17.87	−17.91	5.61
	BPA	−13.16	−9.6	4.71	−19.43	−19.14	5.53
P3 (0 m, 80 m)	The subblock CZT method	−16.06	−8.62	5.03	−18.53	−18.03	7.41
	The subblock–subband CZT method	−13.23	−9.41	4.84	−18.74	−19.03	5.52
	BPA	−13.23	−9.43	4.78	−19.21	−18.87	5.54

The theoretical value of azimuth IRW is 5.10 cm, related to the SAS's transmitter size. Indicators that cannot truly reflect image quality due to main-lobe broadening or amplitude distortion are represented by underlined numbers, while black numbers represent superior results except for BPA's results. From Table 2, it can be seen that for range cross-sections, the subblock CZT method has poor IRWs on all targets, while the subblock–subband CZT method has improved IRWs, which is related to theoretical values. This is due to the subband processing. For azimuth cross-sections, the ISLR of the subblock–subband CZT method is 5.84 dB lower than that of the subblock method for the P1 target, resulting in a reduction in the possibility of azimuth ghosting targets. For P2 and P3, due to azimuthal IRW broadening, the measured values of PSLR and ISLR of the subblock CZT method are distorted. However, the measured values of the subblock–subband CZT method are close to theoretical values. Therefore, simulation results can demonstrate the superiority of the proposed subband processing method.

4.2. Real Data Processing

To further validate the effectiveness of this method, we presented the measured data collected in 2019 by Wuhan Huanda Electronics Co., Ltd. at Zhanghe Reservoir in Jingmen City, Hubei Province, China. The sonar parameters used in the experiment are the same as the parameters used in simulation. The subblock CZT method [60], the proposed subblock–subband CZT method, and BPA [66] were used to focus the measured data, and the results are shown in Figure 15. From the red dashed rectangular box, it can be seen that without subband processing, the subblock CZT method yields the worst results, making it difficult to distinguish the focusing result from the background. The use of subband processing can improve the quality of imaging. The imaging quality of the proposed subblock–subband CZT method is comparable to that of BPA, and contour information can be clearly seen.

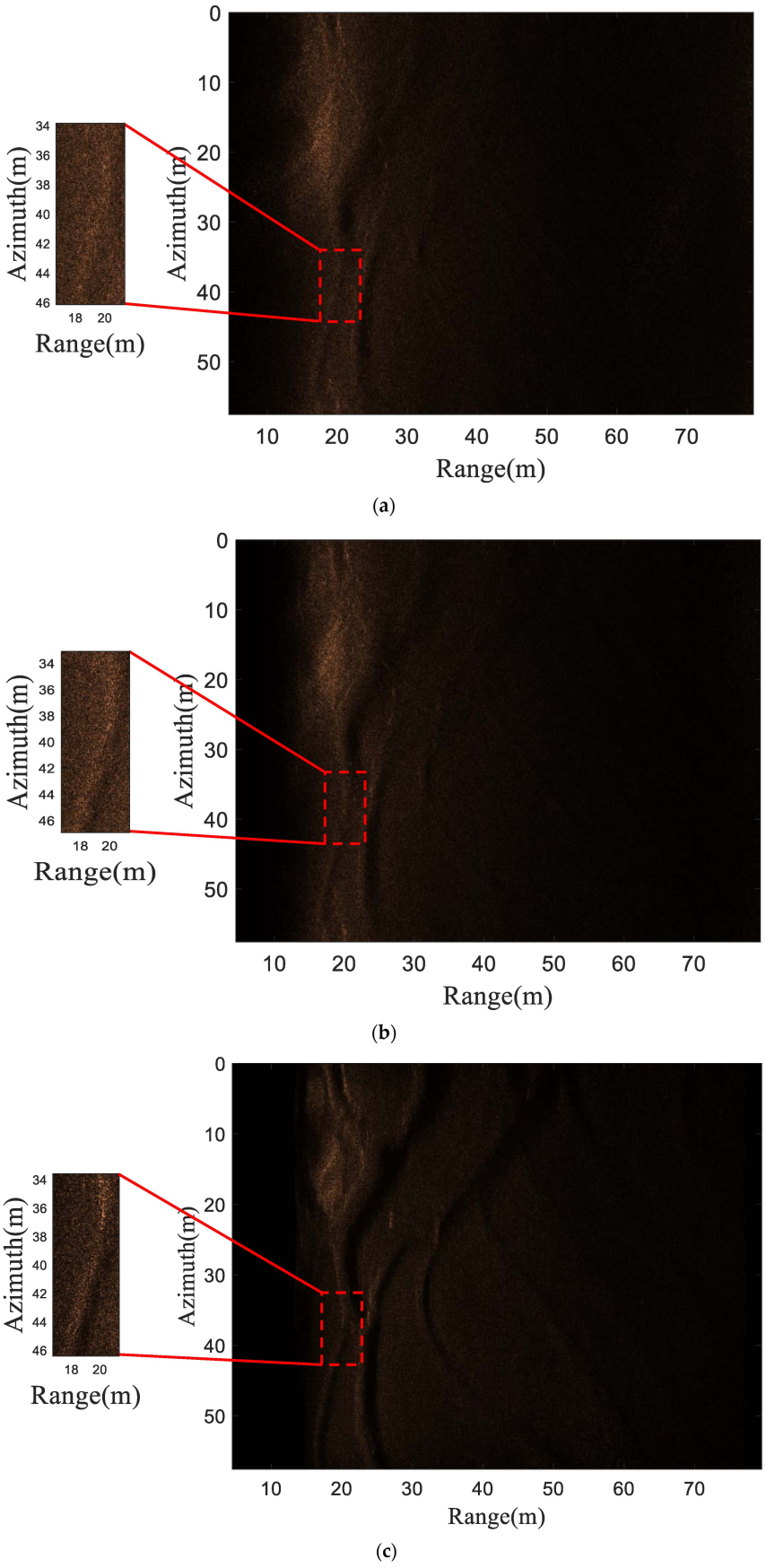


Figure 15. Image results of the field data. (**a**) The subblock CZT method. (**b**) The subblock–subband-CZT method. (**c**) BPA.

5. Conclusions

This paper proposes a range subblock–subband CZT algorithm, and we divide the echo data into the range frequency subbands within each range subblock compared to the traditional range subblocks method. The azimuth range coupling terms are converted into linear terms based on PLA. Then, we use the CZT to correct the RCM to obtain low-resolution results for different subbands. After RCM correction, the subbands' signals are coherently summed in the range frequency domain to obtain a high-resolution image. Finally, different subblocks are concatenated in the range time domain to obtain the final result of the whole swath. The processing of different subblocks and different subbands can be implemented in parallel and can efficiently deploy on hardware with parallel processors.

Author Contributions: Conceptualization, M.N. and J.T.; methodology, M.N.; software, H.Z.; validation, M.M. and P.Z.; formal analysis, H.W.; investigation, H.W. and J.Z.; resources, H.Z.; data curation, H.Z.; writing—original draft preparation, M.N.; writing—review and editing, M.N. and J.T.; visualization, J.Z.; supervision, J.T.; project administration, H.Z.; funding acquisition, H.Z., M.M. and P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China under Grant 42176187 and 62301592 and the Science Foundation of the National University of Defense Technology under Grant ZK21-30.

Data Availability Statement: The real data presented in this paper is not readily available due to technical limitations and copyright issues of Huanda Electronics Co., Ltd.

Acknowledgments: The authors would like to thank all reviewers and editors for their comments on this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Putney, A.; Chang, E.; Chatham, R.; Marx, D.; Nelson, M.; Warman, L.K. Synthetic aperture sonar-the modern method of underwater remote sensing. In Proceedings of the IEEE Aerospace Conference Proceedings, Big Sky, MT, USA, 10–17 March 2001.
2. Marx, D.; Nelson, M.; Chang, E.; Gillespie, W.; Putney, A.; Warman, K. An introduction to synthetic aperture sonar. In Proceedings of the Proceedings of the Tenth IEEE Workshop on Statistical Signal and Array Processing, Manor, PA, USA, 14–16 August 2000.
3. Xu, J.; Tang, J.-S.; Zhang, C.-H. Multi-aperture Synthetic Aperture Sonar lmaging Algorithm. *Signal Process.* **2003**, *19*, 157–160. [CrossRef]
4. Bruce, M.P. A processing requirement and resolution capability comparison of side-scan and synthetic-aperture sonars. *IEEE J. Ocean. Eng.* **1992**, *17*, 106–117. [CrossRef]
5. Huang, P.; Yang, P.-X. Synthetic aperture imagery for high-resolution imaging sonar. *Front. Mar. Sci.* **2022**, *9*, 1049761. [CrossRef]
6. Zhang, J.-F.; Cheng, G.-L.; Tang, J.-S.; Xie, Z.-M.; Wu, H.-R. A Novel Imaging Algorithm for Wide-Beam Multiple-Receiver Synthetic Aperture Sonar Systems. *Remote Sens.* **2023**, *15*, 3745. [CrossRef]
7. Zhang, X.-B.; Yang, P.-X.; Wang, Y.-M.; Shen, W.-Y.; Yang, J.-C.; Wang, J.-F.; Ye, K.; Zhou, M.-Z.; Sun, H.-X. A Novel Multireceiver SAS RD Processor. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4203611. [CrossRef]
8. Wu, H.-R.; Zhou, F.-Y.; Xie, Z.-M.; Tang, J.-S.; Zhong, H.-P.; Zhang, J.-F. Two-Dimensional Space-Variant Motion Compensation Algorithm for Multi-Hydrophone Synthetic Aperture Sonar Based on Sub-Beam Compensation. *Remote Sens.* **2024**, *16*, 2144. [CrossRef]
9. Zhang, X.-B. An efficient method for the simulation of multireceiver SAS raw signal. *Multimed. Tools Appl.* **2024**, *83*, 37351–37368. [CrossRef]
10. Sternlicht, D.D.; Femandez, J.E.; Holtzapple, R.; Kucik, D.P.; Montgomery, T.C.; Loeffler, C.M. Advanced sonar technologies for autonomous mine countermeasures. In Proceedings of the OCEANS'11 MTS/IEEE KONA, Waikoloa, HI, USA, 19–22 September 2011.
11. Fossum, T.G.; Sæbø, T.O.; Langli, B.; Callow, H.; Hansen, R.E. HISAS 1030—High resolution interferometric synthetic aperture sonar. In Proceedings of the Canadian Hydrographic Conference and National Surveyors Conference, Victoria, BC, Canada, 6–7 May 2008.
12. Zhang, P.; Tang, J.-S.; Zhong, H.-P.; Ning, M.-Q.; Liu, D.-D.; Wu, K. Self-Trained Target Detection of Radar and Sonar Images Using Automatic Deep Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4701914. [CrossRef]
13. Hagen, P.E.; Hansen, R.; Fossum, T.; Langli, B. Development of High-Resolution Synthetic Aperture Sonar for Demanding AUV Applications. In Proceedings of the 9th Unmanned Underwater Vehicle Showcase, Southampton, UK, 26–27 September 2007.
14. Zhang, X.; Dai, X.; Yang, B. Fast imaging algorithm for the multiple receiver synthetic aperture sonars. *IET Radar Sonar Navig.* **2018**, *12*, 1276–1284. [CrossRef]

15. Piper, J.E.; Commander, K.W.; Thorsos, E.I.; Williams, K.L. Detection of buried targets using a synthetic aperture sonar. *IEEE J. Ocean. Eng.* **2002**, *27*, 495–504. [CrossRef]
16. Larsen, L.J.; Wilby, A.; Stewart, C. Deep ocean survey and search using synthetic aperture sonar. In Proceedings of the OCEANS 2010 MTS/IEEE SEATTLE, Washington, DC, USA, 20–23 September 2010.
17. Chatillon, J.; Adams, A.E.; Lawlor, M.A.; Zakharia, M.E. SAMI: A low-frequency prototype for mapping and imaging of the seabed by means of synthetic aperture. *IEEE J. Ocean. Eng.* **1999**, *24*, 4–15. [CrossRef]
18. Neto, A.A.; Rodrigues, G.P.; Alvarenga, I.D. Seabed Mapping with HISAS Sonar For Decommissioning Projects High-Resolution Surveying for Decom Planning. Available online: https://api.semanticscholar.org/CorpusID:49237309 (accessed on 25 June 2024).
19. Charlot, D.; Couade, M.; Marty, B.; Fabre, M.N.; Alain, P.; Delbecke, J.; Laquet, T.; Chemisky, B.; Mosca, F.; Bouhier, M.E.; et al. The Synthetic Aperture Mapping Sonar SAMS150 onboard UlyX AUV 6000m: An advanced solution for simultaneous detection and identification of deep-sea features. In Proceedings of the OCEANS 2023, Limerick, Ireland, 5–8 June 2023.
20. Zhang, X.-B.; Yang, P.-X.; Sun, M. Experiment results of a novel sub-bottom profiler using synthetic aperture technique. *Curr. Sci.* **2022**, *122*, 461–464. [CrossRef]
21. Li, Z.-L.; Yu, Y.-X. Overview of deep water acoustics. *Chin. Sci. Bull.* **2022**, *67*, 125–134. [CrossRef]
22. Dai, W. Research on the Characteristics of Underwater Acoustic Channels and the Propagation Characteristics of Signals in Them. *Ship Electron. Eng.* **2023**, *43*, 200–204. [CrossRef]
23. Shannon, C.E. Communication in the Presence of Noise. *Proc. Proc. IRE* **1949**, *37*, 10–21. [CrossRef]
24. Moreira, A.; Prats-Iraola, P.; Nannini, M.; Martín-del-Campo-Becerra, G.D.; Pardini, M.; Papathanassiou, K.; Reigber, A. Spaceborne Multi-Baseline Synthetic Aperture Radar (SAR) Imaging. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023.
25. Zhang, X.-B.; Tan, C.; Ying, W.-W. Imaging Algorithm for Multireceiver Synthetic Aperture Sonar. *J. Electr. Eng. Technol.* **2019**, *14*, 471–478. [CrossRef]
26. Zhang, X.-B.; Chen, X.-H.; Wu, Q. Influence of the stop-and-hop assumption on synthetic aperture sonar imagery. In Proceedings of the 2017 IEEE 17th International Conference on Communication Technology (ICCT), Chengdu, China, 27–30 October 2017.
27. Liang, Z.; Fu, X.; Lv, X. New Channel Errors Estimation Method for Multichannel SAR Based on Virtual Calibration Source. *Remote Sens.* **2021**, *13*, 3625. [CrossRef]
28. Lv, Y.-N.; Shang, M.-Y.; Zhong, L.-H.; Qiu, X.-L.; Ding, C.-B. A Novel Imaging Scheme of Squint Multichannel SAR: First Result of GF-3 Satellite. *Remote Sens.* **2022**, *14*, 3962. [CrossRef]
29. Yang, H.-L.; Zhang, S.; Tang, J.-S. Study on Simulation of Multiple-receiver Synthetic Aperture Sonar Imagery Based on Wide Swath. *J. Syst. Simul.* **2011**, *23*, 1424–1428. [CrossRef]
30. Zhang, X.-B.; Liu, Y.-Q.; Deng, X.-Y. Influence of Phase Centre Approximation Error on SAS Imagery. In Proceedings of the 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS), Chengdu, China, 23–26 April 2021.
31. Zhang, X.-B.; Tang, J.-S.; Zhang, S.; Bai, S.-X.; Zhong, H.-P. Four-order Polynomial Based Range-Doppler Algorithm for Multi-receiver Synhetic Aperture Sonar. *J. Electron. Inf. Technol.* **2014**, *36*, 1592–1598. [CrossRef]
32. Neo, Y.L.; Wong, F.; Cumming, I.G. A Two-Dimensional Spectrum for Bistatic SAR Processing Using Series Reversion. *IEEE Geosci. Remote Sens. Lett.* **2007**, *4*, 93–96. [CrossRef]
33. Wu, H.-R.; Jin-Song, T.; Zhong, H.-P.; Yi-Shuo, T. Multi-aperture range-Doppler imaging algorithm based on spectrumofseries reversion. *J. Huazhong Univ. Sci. Technol. (Nat. Sci. Ed.)* **2018**, *46*, 6. [CrossRef]
34. Zhang, X.-B.; Ying, W.-W.; Liu, Y.-Q.; Deng, X.-Y. Processing Multireceiver SAS Data Based on the PTRS Linearization. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021.
35. Zhang, X.-B.; Yang, P.-X.; Sun, H.-X. Frequency-domain multireceiver synthetic aperture sonar imagery with Chebyshev polynomials. *Electron. Lett.* **2022**, *58*, 995–998. [CrossRef]
36. Zhang, X.-B.; Yang, P.-X.; Dai, X.-T. Focusing Multireceiver SAS Data Based on the Fourth-Order Legendre Expansion. *Circuits Syst. Signal Process.* **2019**, *38*, 2607–2629. [CrossRef]
37. Wang, R.; Loffeld, O.; Ul-Ann, Q.; Nies, H.; Ortiz, A.M.; Samarah, A. A Bistatic Point Target Reference Spectrum for General Bistatic SAR Processing. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 517–521. [CrossRef]
38. Zhang, X.-B.; Wu, H.-R.; Sun, H.-X.; Ying, W.-W. Multireceiver SAS Imagery Based on Monostatic Conversion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10835–10853. [CrossRef]
39. Wu, J.-J.; Li, Z.-Y.; Huang, Y.-L.; Yang, J.-Y.; Liu, Q.-H. An Omega-K Algorithm for Translational Invariant Bistatic SAR Based on Generalized Loffeld's Bistatic Formula. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 6699–6714.
40. Zhang, X.-B.; Yang, P.-X.; Wang, Y.-M.; Shen, W.-Y.; Yang, J.-C.; Ye, K.; Zhou, M.-Z.; Sun, H.-X. LBF-Based CS Algorithm for Multireceiver SAS. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1502505. [CrossRef]
41. Cumming, I.G.; Wong, F.H. *Digital Processing of Synthetic ApertureRadar Data: Algorithms and Implementation*; Artech House Inc.: Norwood, OH, USA, 2004.
42. Zhang, X.-B.; Yang, P.-X.; Sun, H.-X. An omega-k algorithm for multireceiver synthetic aperture sonar. *Electron. Lett.* **2023**, *13*, e12859. [CrossRef]
43. Tan, C.; Zhang, X.-B.; Yang, P.-X.; Sun, M. A Novel Sub-Bottom Profiler and Signal Processor. *Sensors* **2019**, *19*, 5052. [CrossRef]

44. Zhang, X.-B.; Wang, Y.-M.; Yang, J.-C.; Shen, W.-Y.; Sun, H.-X. Range-Doppler Imaging Algorithm for Multireceiver Synthetic Aperture Sonar. *J. Electron. Inf. Technol.* **2022**, *44*, 2104–2110. [CrossRef]
45. Zhu, J.-H.; Xie, Z.; Jiang, N.; Song, Y.-P.; Han, S.-D.; Liu, W.-J.; Huang, X.-T. Delay-Doppler Map Shaping through Oversampled Complementary Sets for High-Speed Target Detection. *Remote Sens.* **2024**, *16*, 2898. [CrossRef]
46. Runge, H.; Bamler, R. A Novel High Precision SAR Focussing Algorithm Based On Chirp Scaling. In Proceedings of the 12th Annual International Geoscience and Remote Sensing Symposium, Houston, TX, USA, 26–29 May 1992.
47. Raney, R.K.; Runge, H.; Bamler, R.; Cumming, I.G.; Wong, F.H. Precision SAR processing using chirp scaling. *IEEE Trans. Geosci. Remote Sens.* **1994**, *32*, 786–799. [CrossRef]
48. Ma, M.-B.; Tang, J.-S.; Zhong, H.-P.; Wu, H.-R. Multireceiver Synthetic Aperture Sonar Chirp Scaling Algorithm Considering Intrapulse Doppler Shift. *IEEE J. Ocean. Eng.* **2022**, *47*, 433–444. [CrossRef]
49. Yang, P.-X. An imaging algorithm for high-resolution imaging sonar system. *Multimed. Tools Appl.* **2024**, *83*, 31957–31973. [CrossRef]
50. Davidson, G.W.; Cumming, I.G.; Ito, M.R. A chirp scaling approach for processing squint mode SAR data. *IEEE Trans. Aerosp. Electron. Syst.* **1996**, *32*, 121–133. [CrossRef]
51. Zhang, X.-B.; Yang, P.-X.; Feng, X.; Sun, H.-X. Efficient imaging method for multireceiver SAS. *IET Radar Sonar Navig.* **2022**, *16*, 1470–1483. [CrossRef]
52. Li, C.; Zhang, H.; Deng, Y.-K. Focus Improvement of Airborne High-Squint Bistatic SAR Data Using Modified Azimuth NLCS Algorithm Based on Lagrange Inversion Theorem. *Remote Sens.* **2021**, *13*, 1916. [CrossRef]
53. Loffeld, O.; Hein, A.; Schneider, F. SAR focusing: Scaled inverse Fourier transformation and chirp scaling. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Seattle, WA, USA, 6–10 July 1998.
54. Wei, W.; Zhu, D.-Y.; Wu, D. Wavenumber Domain Algorithm Based on the Principle of Chirp Scaling for SAR Imaging. *J. Radars* **2020**, *9*, 9. [CrossRef]
55. Wu, Y.-J.; Huang, Y. Inverse Scaled Fourier Transformation Algorithmfor Squint Mode SAR lmaging. *Signal Process.* **2010**, *26*, 4.
56. Ma, M.-B.; Tang, J.-S.; Wu, H.-R.; Zhang, P.; Ning, M.-Q. CZT Algorithm for the Doppler Scale Signal Model of Multireceiver SAS Based on Shear Theorem. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5201412. [CrossRef]
57. Ma, M.-B.; Jin-Song, T.; Zhong, H.-P. CZT Algorithm for Multiple-Receiver Synthetic Aperture Sonar. *IEEE Access* **2020**, *8*, 1902–1909. [CrossRef]
58. Ma, M.-B.; Jin-Song, T.; Zhong, H.-P.; Tian, Z. CZT Algorithm for Multiple-receiver Synthetic Aperture Sonar. *J. Huazhong Univ. Sci. Technol. (Nat. Sci. Ed.)* **2019**, *47*, 6. [CrossRef]
59. Franceschetti, G.; Lanari, R.; Marzouk, E.S. A new two-dimensional squint mode SAR processor. *IEEE Trans. Aerosp. Electron. Syst.* **1996**, *32*, 854–863. [CrossRef]
60. Zhang, X.-B.; Yang, P.-X. An Improved Imaging Algorithm for Multi-Receiver SAS System with Wide-Bandwidth Signal. *Remote Sens.* **2021**, *13*, 5008. [CrossRef]
61. Zhang, S.; Jin-Song, T.; Ming, C.; Sheng-Xiang, B. Development and sea trial of interferometric synthetic aperture sonar. *Tech. Acoust.* **2012**, *31*, 167–173. [CrossRef]
62. Raney, R.K. A New And Fundamental Fourier Transform Pair. In Proceedings of the IGARSS '92 International Geoscience and Remote Sensing Symposium, Houston, TX, USA, 26–29 May 1992.
63. Chen, J.-L. Study on Signal Modeling and Imaging Algorithm for Airborne/Spaceborne SAR with Nonlinear Trajectory. Ph.D. Thesis, Xidian University, Xi'an, China, 2018.
64. Rabiner, L.R.; Schafer, R.W.; Rader, C.M. The chirp z-transform algorithm and its application. *Bell Syst. Tech. J.* **1969**, *48*, 1249–1292. [CrossRef]
65. Bluestein, L. A linear filtering approach to the computation of discrete Fourier transform. *IEEE Trans. Audio Electroacoust.* **1970**, *18*, 451–455. [CrossRef]
66. Zhang, X.-B.; Tang, J.-S.; Wang, F.; Bai, S.-X.; Liu, D.-D. Accurate back projection imaging algorithm for multi-receiverSAS in engineering application. *J. Nav. Univ. Eng.* **2014**, *26*, 5. [CrossRef]
67. Zhang, X.-B.; Yang, P.-X. Back Projection Algorithm for Multi-Receiver Synthetic Aperture Sonar Based on Two Interpolators. *J. Mar. Sci. Eng.* **2022**, *10*, 718. [CrossRef]
68. Zhu, J.-H.; Song, Y.-p.; Jiang, N.; Xie, Z.; Fan, C.-Y.; Huang, X.-T. Enhanced Doppler Resolution and Sidelobe Suppression Performance for Golay Complementary Waveforms. *Remote Sens.* **2023**, *15*, 2452. [CrossRef]
69. Shi, J.-C.; Yu, S.-X. The Analysis of the Error in Sampling. *Pet. Instrum.* **1995**, *9*, 154–159.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

Joint Luminance-Saliency Prior and Attention for Underwater Image Quality Assessment

Zhiqiang Lin [1], Zhouyan He [1,*], Chongchong Jin [1], Ting Luo [1] and Yeyao Chen [2]

[1] College of Science and Technology, Ningbo University, Ningbo 315212, China; 2211170008@nbu.edu.cn (Z.L.); jinchongchong@nbu.edu.cn (C.J.); luoting@nbu.edu.cn (T.L.)
[2] Faculty of Information Science and Engineering, Ningbo University, Ningbo 315212, China; chenyeyao@nbu.edu.cn
* Correspondence: hezhouyan@nbu.edu.cn

Abstract: Underwater images, as a crucial medium for storing ocean information in underwater sensors, play a vital role in various underwater tasks. However, they are prone to distortion due to the imaging environment, which leads to a decline in visual quality, which is an urgent issue for various marine vision systems to address. Therefore, it is necessary to develop underwater image enhancement (UIE) and corresponding quality assessment methods. At present, most underwater image quality assessment (UIQA) methods primarily rely on extracting handcrafted features that characterize degradation attributes, which struggle to measure complex mixed distortions and often exhibit discrepancies with human visual perception in practical applications. Furthermore, current UIQA methods lack the consideration of the perception perspective of enhanced effects. To this end, this paper employs luminance and saliency priors as critical visual information for the first time to measure the enhancement effect of global and local quality achieved by the UIE algorithms, named JLSAU. The proposed JLSAU is built upon an overall pyramid-structured backbone, supplemented by the Luminance Feature Extraction Module (LFEM) and Saliency Weight Learning Module (SWLM), which aim at obtaining perception features with luminance and saliency priors at multiple scales. The supplement of luminance priors aims to perceive visually sensitive global distortion of luminance, including histogram statistical features and grayscale features with positional information. The supplement of saliency priors aims to perceive visual information that reflects local quality variation both in spatial and channel domains. Finally, to effectively model the relationship among different levels of visual information contained in the multi-scale features, the Attention Feature Fusion Module (AFFM) is proposed. Experimental results on the public UIQE and UWIQA datasets demonstrate that the proposed JLSAU outperforms existing state-of-the-art UIQA methods.

Keywords: underwater image quality assessment (UIQA); luminance prior; saliency prior; multi-scale

Citation: Lin, Z.; He, Z.; Jin, C.; Luo, T.; Chen, Y. Joint Luminance-Saliency Prior and Attention for Underwater Image Quality Assessment. *Remote Sens.* **2024**, *16*, 3021. https://doi.org/10.3390/rs16163021

Academic Editors: Xuebo Zhang, Lei Zhang, Jingyue Lu, Haiyong Xu, Marc Pinto and Farman Ali

Received: 8 July 2024
Revised: 10 August 2024
Accepted: 15 August 2024
Published: 17 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Underwater images serve as an important medium for transmitting ocean information, supporting autonomous underwater vehicles, remotely operated vehicles, and other marine instruments to help them efficiently complete tasks such as ocean resource exploration and deep-sea facility monitoring [1]. However, due to the unique properties of the underwater environment, the imaging process often suffers from light absorption and scattering, resulting in low contrast, color casts, haze effects, and other visual impediments that affect subsequent image processing systems in underwater instruments [2]. Consequently, extensive research on Underwater Image Enhancement (UIE) algorithms [3] has received significant attention, aiming to improve the visual quality of underwater images and aiding downstream tasks (i.e., saliency detection, object detection, image segmentation, etc.). However, due to the lack of high-quality underwater reference images, most existing UIE algorithms rely on synthetic datasets, and none are universally effective,

limiting their applicability in real-world scenarios. Enhanced underwater images may still retain unresolved issues or introduce additional distortions, such as artifacts related to over-enhancement or detail loss related to under-enhancement. Therefore, Image Quality Assessment (IQA) [4–7] methods are often necessary for objective assessment, promoting the development of image systems for various optical sensors.

Thus far, many traditional in-air IQA methods have been proposed [8–11]. Among them, Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are classic Full Reference (FR) methods that often serve as benchmarks for optimizing UIE algorithms. However, MSE and PSNR, defined based on pixel-wise differences, fail to adequately capture features affecting human perception, such as texture details [12]. Given the unavailability of reference images in many cases, the application value of No-Reference (NR) IQA methods becomes more prominent. Existing NR-IQA methods primarily gauge image distortion by extracting Natural Scene Statistics (NSS) features in spatial or transform domains. These encompass classic NSS features, like Mean Subtracted Contrast Normalized (MSCN) coefficients [8], wavelet transformation coefficients [9], and discrete cosine transformation coefficients [10], among others, which have shown success in in-air images. However, due to the unique imaging environment, statistical NSS features change in underwater images, making it difficult to capture the increased noise and non-Gaussian distribution variations present in these images. Additionally, some progress has been made in deep learning (DL)-based IQA methods for in-air images [13–20]. For example, some methods utilize transfer learning [13] and meta-learning [16] to learn distortion features of in-air images, thereby achieving distortion classification and discrimination. However, the imaging characteristics and primary degradations of underwater images differ from those of in-air images. More importantly, the UIE algorithms may lead to more complex and variable artificial distortions of underwater images. Therefore, directly applying IQA methods designed for air images to evaluate underwater image quality may yield suboptimal results.

To this end, many traditional underwater IQA (UIQA) methods [21–30] have been proposed. Most of the existing traditional UIQA methods rely on extracting handcrafted features to reflect the distortion situations. For instance, methods like underwater color image quality evaluation (UCIQE) [21] gauge underwater image quality by extracting statistical features like chroma, saturation, and contrast from the CIELab color space, but do not consider any human perception factors. The underwater image quality measure (UIQM) [22], inspired by characteristics of the Human Vision System (HVS), incorporates measurements of colorfulness, sharpness, and contrast to evaluate underwater image degradation. However, the method of assigning manual weights to the extracted features to represent image quality has poor generalization and fails to adequately represent diverse degradation scenarios. Although these methods have a certain ability to assess common distortions such as low contrast and blurriness, they struggle to precisely quantify compound distortions caused by complex and dynamic underwater environments. Particularly, they inadequately perceive the intricate variations in color and structure introduced by UIE algorithms, like redshifts and artifacts. Due to the excellent feature learning capabilities of DL, some deep models for UIQA have been proposed [31–33]. Most of the DL-based UIQA methods use additional priors or conditions beyond the input to better learn the quality features. For example, some methods integrate color histogram priors as quality tokens to supplement global degradation information [31], but they fail to reflect the impact of local spatial structure on human perception. Others utilize depth maps as the weight to differentiate between foreground and background [32], but lack consideration of phenomena such as low contrast caused by luminance distortions. Previous works [34,35] used saliency maps as weights to measure perceptual quality. Typically, they directly apply or resize the saliency map to fit the required input size without delving into their inherent perceptual information. Although these embedded conditions or priors help the methods to understand the image quality to some extent, there is still a lack of sufficient consideration for the visual information introduced by UIE algorithms. Therefore, it is necessary to

improve the UIQA methods to evaluate the quality of the enhanced underwater images more accurately.

In this paper, we propose a novel joint saliency–luminance prior UIQA model called JLSAU. Specifically, considering distortions at different scales, such as texture details and haze effects, a pyramid-structured backbone network is constructed to capture multi-scale distortion features. Additionally, to measure the impact of the UIE algorithms on visual information, luminance and saliency priors are incorporated into multi-scale features, where luminance prior and saliency prior reflect the enhancement effect of global and local quality information, respectively. To this end, the Luminance Feature Extraction Module (LFEM) is proposed to comprehensively measure the global distortion of luminance, which is highly sensitive to the HVS. It supplements multi-scale features to enhance the perception of global distortions, such as low contrast. Furthermore, the Saliency Weight Learning Module (SWLM) is designed to learn saliency features that reflect variations in local quality, serving as supplementary priors to enhance the perception of textures and details within multi-scale features. Finally, multi-scale features containing rich perceptual quality information are obtained and the AFFM based on attention mechanism is proposed to model the visual information at different levels. The experimental results show that, compared to existing UIQA methods, the proposed JLSAU has better performances for two different underwater image datasets. The main contributions of our work are summarized as follows:

- To enhance the global quality perception of multi-scale features, the LFEM is proposed to learn quality-aware representations of luminance distribution as prior knowledge. This includes learning quantitative statistical features through histogram statistics and designing a convolutional network to encode and supplement the positional information missing from the statistical features.
- To improve the local quality perception of multi-scale features, the SWLM is designed to extract saliency features in the channel and spatial domains as prior knowledge. These features reflect the enhancement effect of UIE algorithms on local quality, thereby enhancing the perception of structure and texture.
- To model the relationship of perceptual information among multi-scale features augmented by luminance and saliency prior, AFFM is introduced. It effectively integrates perception quality information from different levels of features using attention mechanisms, aiming to comprehensively perceive distortions in enhanced images.

2. Related Work

2.1. Underwater Image Enhancement Algorithms

Jaffe McGlamery [36] defines the underwater Image Forming Model (IFM) as a linear combination of direct reflection, background scattering, and forward scattering of light. The forward scattering is almost negligible. When only the direct scattering and the backscattering are considered, the underwater image $\boldsymbol{I}_\alpha$ can be defined as follows:

$$\boldsymbol{I}_\alpha = \boldsymbol{S}_\alpha \cdot \boldsymbol{\tau}_\alpha + \boldsymbol{B}_\alpha(1 - \boldsymbol{\tau}_\alpha) \tag{1}$$

where $\alpha \in \{R, G, B\}$ represents the color channel, $\boldsymbol{S}_\alpha$ represents the original scene, $\boldsymbol{\tau}_\alpha$ represents the transmission map, and $\boldsymbol{B}_\alpha$ represents the background light.

Based on the involvement of IFM, Underwater Image Enhancement (UIE) algorithms can be categorized as model-based or model-free. Model-based algorithms typically rely on estimating parameters to reverse IFM. For instance, Peng et al. [37] estimated the background light, scene depth, and transmission map through blurriness and light absorption of underwater images. Zhao et al. [38] derived the relevant attenuation and scattering coefficients of RGB channels from the background color of underwater images for dehazing and color correction. However, parameter estimation often lacks sufficient constraints, requiring appropriate prior information for optimization. To address this issue, Drew et al. [39] introduced the underwater dark channel prior based on the characteristic that red light is more

easily absorbed in water, which provides a good constraint for estimating the transmission map. The utilization of reasonable prior information and constraints has indeed improved the accuracy of parameter estimation to some extent. However, due to the complexity and variability of the underwater environment, significant estimation deviations still persist. For example, inaccurate depth estimation may result in structure distortion. In addition, to mitigate the impact of environmental factors on image degradation, some studies begun to explore extended IFM. For example, Li et al. [40] proposed an effective variational framework based on an extended IFM.

Model-free algorithms adopt direct pixel value adjustments to improve image quality. For example, Fu et al. [41] proposed a two-step UIE algorithm, incorporating color correction and contrast enhancement. Li et al. [42] decomposed underwater images into high- and low-frequency components, suppressing noise in the high-frequency component and adaptively enhancing color and contrast in the low-frequency component. Lu et al. [43] utilized a diffusion model for UIE, developing a new accelerated sampling method to optimize the probability distribution in the diffusion stage. Additionally, reinforcement learning-based UIE algorithms have made significant progress. For instance, Wang et al. [44] first introduced reinforcement learning into UIE algorithms by using UIQM as a feedback function and a positive learning direction. Sun et al. [45] modeled UIE as a Markov decision process, optimizing it based on reinforcement learning strategies. Wang et al. [46] proposed a reinforcement learning paradigm that controls actions to enhance visual effects and improve object detection performance, but it did not consider human perception factors. To address this, they later introduced a human visual perception-driven paradigm [47] and a method [48] that incorporates color priors into the scoring process. Moreover, some reinforcement learning methods are optimized based on visual preferences, such as the Meta underwater camera [49], which uses reinforcement learning strategies to globally optimize the configuration of parameter values in a comprehensive cascade of UIE methods. There is also a framework [50] that fine-tunes and improves hierarchical probability networks using reinforcement learning. Additionally, the Metalantis framework [51] supports IFM through reinforcement learning with virtually generated data. However, these reinforcement learning-based UIE algorithms, which use image quality as a reward, are limited by the performance of UIQA. Although these algorithms have improved underwater image quality to some extent, the complexity of the underwater imaging environment makes it challenging to completely restore image details. Additionally, these algorithms may introduce additional distortions, such as red offsets, artifacts, and structure loss, resulting in no substantial improvement in visual quality. Therefore, appropriate visual quality assessment methods are necessary to evaluate and enhance the performance of UIE algorithms.

2.2. Underwater Image Quality Assessment

In recent years, Image Quality Assessment (IQA) has made significant strides in in-air images [9–20]. Among them, Peak Signal-to-Noise Ratio (PSNR), Mean Square Error (MSE), and Structural Similarity Index Measure (SSIM) [52] are commonly used as Full Reference (FR) metrics for optimizing and evaluating UIE algorithms. Since FR-IQA requires reference images that are often unavailable, studying No Reference (NR) IQA methods becomes more practically significant. For in-air images, numerous NR-IQA methods based on Natural Scene Statistics (NSS) have been developed. For example, Mittal et al. [8] proposed the use of Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), which quantifies naturalness using NSS features of local normalized brightness coefficients without frequency domain decomposition. Moorthy et al. [9] proposed an NR-IQA method called Distortion Identification-based Image INtegrity and Verity Evaluation (DIIVINE), which extracts NSS features in the wavelet domain to identify image distortion types. Since deep learning (DL) can automatically learn image distortion features, attention has shifted to DL-based NR-IQA methods [13–20]. For example, Kang et al. [14] designed a Convolutional Neural Network (CNN)-based method that represents overall quality by predicting the quality of image patches. Su et al. [16] introduced a meta-learning

based method, called Meta-IQA, which adapts to real unknown distortions by learning shared meta-knowledge. Meanwhile, Golestaneh et al. [17] developed an evaluation model using a hybrid structure of CNN and Transformer, incorporating relative ranking loss and self-consistency loss as constraints. However, due to the distinct imaging environments and principles of underwater versus in-air images, there are significant differences in the types and levels of distortion. Therefore, IQA methods designed for in-air images may not perform optimally for evaluating underwater images.

To address the unique distortions in underwater images and gain insights into UIE algorithm development, Yang et al. [21] proposed an NR method, called Underwater Color Image Quality Evaluation Metric (UCIQE), which is a linear combination of chrominance, saturation, and contrast. It quantitatively measures distortions such as color shift, blur, and low contrast, but lacks sufficient consideration of human visual perception-related factors. Inspired by the properties of HVS, Panetta et al. [22] introduced the underwater image quality measure (UIQM), which includes three components, i.e., color measurement, sharpness measurement, and contrast measurement. However, linear weighting of handcrafted features still struggles to accurately capture the diversity of degradation types and degrees in underwater images. Then, Wang et al. [23] quantified color loss, blur, and haze effects by analyzing the influence of light absorption and scattering. Jiang et al. [24] derived quality-aware features from chrominance and luminance components for assessing color shift, inadequate luminosity, and degradation of sharpness or detail. Yang et al. [25] combined with human visual perception properties, proposed a method that considers the degradation effects in underwater images, albeit overlooking naturalness and structural information. Zheng et al. [26] introduced the Underwater Image Fidelity (UIF) method, which extracts the statistical features from the CIElab space to construct naturalness, sharpness, and structure indicators, yet disregards color information. Li et al. [27] extracted statistical features based on underwater dark channel and underwater bright channel priors in their proposed method. In contrast to prior approaches focusing on a limited set of image attributes, Liu et al. [28] developed six quality perception features aimed at providing a more comprehensive evaluation of luminance, sharpness, color balance, contrast, haze density, and noise. Zhang et al. [29] introduced a series of quality awareness features, including naturalness, color, contrast, clarity, and structure. Hou et al. [30] introduced new measures of contrast, clarity, and naturalness for underwater images. Although traditional UIQA methods based on handcrafted features have achieved some success, DL-based methods have gradually gained public attention due to their performance advantages.

In recent years, some DL-based methods have been proposed, using additional priors or conditions. For example, Guo et al. [31] utilized the RGB color histogram prior as an embedding feature to supplement global degradation information, combined with a dynamic cross-scale correspondence module for multi-scale feature fusion. Nevertheless, the histogram prior derived solely from pixel statistics fails to reflect the impact of different color information within specific spaces and structures on human perception. Li et al. [32] considered the impact of spatial information on local quality by using the depth map prior as a weight to guide the model in distinguishing foreground–background differences. Fu et al. [33] generated samples of varying qualities by linear mixing and trained a Siamese network using a self-supervised mechanism to learn quality ranking. However, they did not consider cases of over-enhancement introduced by UIE algorithms, leading to artifacts and other issues. Despite this, most UIQA methods are still primarily based on basic visual features, such as saturation and sharpness, lacking consideration for HVS characteristics and their relevance to downstream tasks. Therefore, this paper focuses on the changes in saliency detection tasks brought about by the enhancement effects of UIE and combines human visual characteristics to propose the JLSAU method.

3. The Proposed Method

To capture the distortion present in enhanced underwater images, we propose a novel multi-scale UIQA model named JLSAU, with a joint luminance–saliency prior. To focus on

the variations in image saliency caused by local distortions, saliency features are extracted as visual weights to measure local perceptual quality. Additionally, the perceptual quality representation of luminance information is learned by extracting grayscale histogram statistical features, and positional information is supplemented. Finally, considering the multi-scale perceptual characteristics of HVS, a fusion module based on the attention mechanism is constructed to model the distortion information between different scale features. The overall framework of the proposed JLSAU is shown in Figure 1. In this section, the key parts of JLSAU will be explained in detail, including its pyramid-structured backbone (hereinafter referred to as backbone), Luminance Feature Extraction Module (LFEM), Saliency Weight Learning Module (SWLM), Attention Feature Fusion Module (AFFM), and loss function setting.

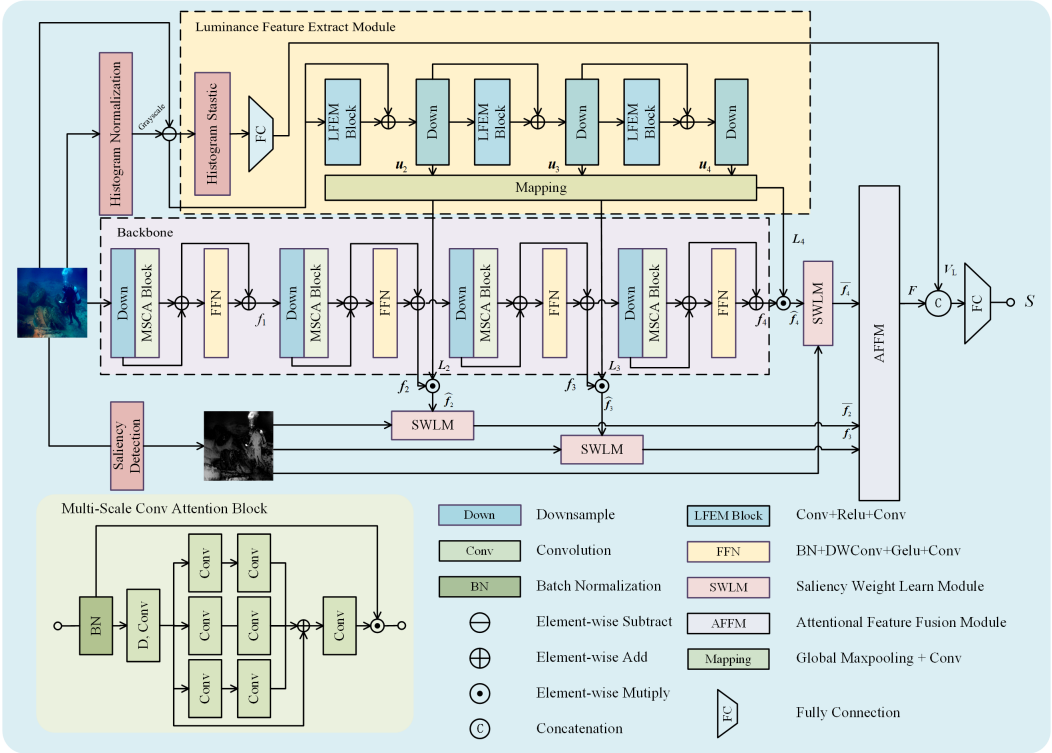


Figure 1. Framework of the proposed method.

3.1. Backbone

Enhanced underwater images not only include distortions caused by the inherent properties of the water medium, but also artificial distortions introduced by UIE algorithms. The distortions can be categorized at different scales into local distortions, such as structure loss, artifacts, and false edges, as well as global distortions, such as low contrast and color deviations. To capture the different scales of distortion information, we design a pyramid structure module as the backbone for JLSAU. The backbone extracts multi-scale features from the RGB space containing abundant raw color information. The extracted multi-scale features contain rich perceptual quality information such as color, texture, etc. The backbone consists of four stages. Each stage is composed of a downsampling layer, a Multi-Scale Conv Attention (MSCA), and a Feed-Forward Network (FFN). To reduce information loss, overlapping convolution is used instead of pooling for downsampling, which also facilitates modeling the local continuity of structural information. The MSCA consists of a batch normalization layer, a depth-wise convolution, multi-branch strip convolutions, and a 1×1 convolution. As shown in Figure 1, MSCA can be represented as follows:

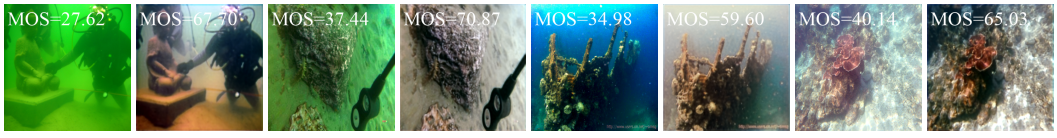
$$MSCA(x) = Conv_{1\times1}\left(\sum_{n=1}^{4} Scale_n(Conv_{\mathrm{dw}}(BN(x)))\right) \odot BN(x) \tag{2}$$

where x represents the input feature, $\odot$ represents the element-wise matrix multiplication, $BN(\cdot)$ represents batch normalization, $Conv_{1\times1}(\cdot)$, and $Conv_{\mathrm{dw}}(\cdot)$ represent 1×1 convolution and depth-wise convolution, respectively, and $Scale_n(\cdot)$ represents the n-th branch of multi-branch strip convolutions, where $n \in \{1, 2, 3, 4\}$.

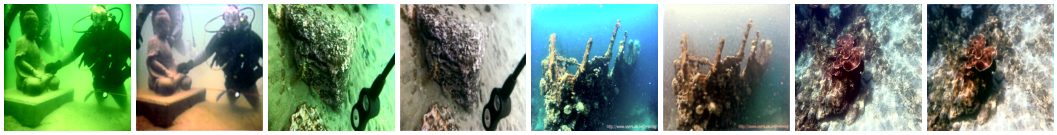
Specifically, depth-wise convolution aggregates local information, multi-branch strip convolutions capture multi-scale features, and 1×1 convolution aggregates the information between channels. FFN is composed of BN-Conv-Gelu-Conv, serving to efficiently integrate and transform channel-wise features, enhancing the representational power of the network for each stage. Finally, the output features are obtained during every stage, denoted as $\boldsymbol{f}_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, where $i \in \{1, 2, 3, 4\}$ represents stages C_i and H_i and W_i represent the number of channels, height, and width of the feature maps at the i-th stage, respectively.

3.2. Luminance Feature Extraction Module

Underwater images typically exhibit global distortions such as low contrast and low visibility in terms of luminance, which significantly affects their quality. To analyze the performance of enhanced underwater images of varying quality in terms of luminance attributes, adaptive histogram equalization is applied. Histogram equalization can distribute pixel values uniformly across the entire luminance range, enhance image contrast, and improve detail. As shown in Figure 2, four groups of enhanced underwater images and their corresponding histogram equalization results are selected from the UIQE dataset. Each group includes enhanced images obtained by different UIE algorithms that vary from low quality to high quality. Obviously, compared to low-quality images, high-quality images exhibit smaller differences before and after processing, resulting in higher visual similarity. Conversely, the global distortions such as low contrast in low-quality images have been significantly improved. Viewing histogram equalization as a mapping process, the mapping distance between the before and after results can reflect image similarity and thereby indicate image quality. Thus, given an enhanced underwater image $\boldsymbol{I}$ and its histogram equalization result $\overline{\boldsymbol{I}}$, the difference between them can reflect the quality of $\boldsymbol{I}$. To quantify the difference, LFEM is designed to extract high-dimensional luminance features containing differential information.



(a) Enhanced underwater images.



(b) Histogram equalization images.

Figure 2. Four groups of enhanced underwater images with different qualities obtained by different UIE algorithms and the corresponding histogram equalization images.

Specifically, $\boldsymbol{I}$ and $\overline{\boldsymbol{I}}$ are transferred to grayscale and the results are denoted as $\boldsymbol{I}_{\mathrm{g}}$ and $\overline{\boldsymbol{I}}_{\mathrm{g}}$. Let $\overline{\boldsymbol{I}}_{\mathrm{diff}}$ be the difference map obtained by subtracting $\boldsymbol{I}_{\mathrm{g}}$ from $\overline{\boldsymbol{I}}_{\mathrm{g}}$, which serves as the input of the LFEM. Then, the statistical features of $\overline{\boldsymbol{I}}_{\mathrm{diff}}$ are obtained through histogram statistics, which are sent into a fully connected layer to obtain the final luminance statistical vector, denoted as $\boldsymbol{V}_{\mathrm{L}}$. It can be expressed as follows:

$$\boldsymbol{V}_{\mathrm{L}} = FC\left(histogram\left(\overline{\boldsymbol{I}}_{\mathrm{diff}}\right)\right) \tag{3}$$

where $histogram(\cdot)$ represents the histogram statistical operation in the grayscale space, and $FC(\cdot)$ represents the fully connection.

Histogram statistics only capture the quantitative statistical relationship of grayscale levels without providing the specific spatial information of pixels. Inspired by the encoding ability of the convolution layer with the relative position [53], a convolutional branch is designed for implicit position encoding. Specifically, it comprises three stages, each comprising an LFE block (i.e., Conv-ReLU-Conv layer), a skip connection, and a downsampling layer. The LFE block not only captures local features but also preserves the spatial context, allowing the network to infer relative positions implicitly, which can be expressed as follows:

$$\boldsymbol{u}_{i+1} = down(Conv(ReLU(Conv(\boldsymbol{u}_i))) + \boldsymbol{u}_i) \tag{4}$$

where $i \in \{1, 2, 3\}$ represents the number of stages, $\boldsymbol{u}_i$ represents the output of the convolutional branch at stage i, $\boldsymbol{u}_1$ is equivalent to $\overline{\boldsymbol{I}}_{\text{diff}}$, $down(\cdot)$ represents downsampling, and $Relu(\cdot)$ represents the ReLU activation function.

Subsequently, due to the subsequent processes not involving the handling of the variable whose subscript i is 1, to avoid ambiguity and describe concisely, the subscript i will be replaced with λ below, where $\lambda \in \{2, 3, 4\}$. The obtained features $\boldsymbol{u}_\lambda$ are subjected to a mapping operation that includes global max pooling and convolutional layers to obtain multi-scale luminance features $\boldsymbol{L}_\lambda$, which can be expressed as follows:

$$\boldsymbol{L}_\lambda = Conv(GMP(\boldsymbol{u}_\lambda)) \tag{5}$$

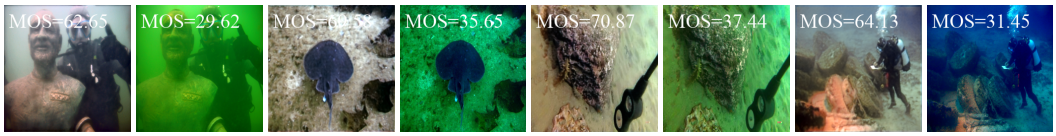
where $GMP(\cdot)$ represents global max pooling.

Finally, by pixel-wise multiplication with $\boldsymbol{L}_\lambda$ containing relative position information, multi-scale features $\boldsymbol{f}_\lambda$ are supplemented with luminance information, which can be expressed as follows:

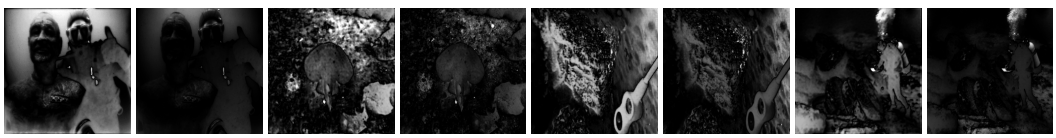
$$\widehat{\boldsymbol{f}}_\lambda = \boldsymbol{f}_\lambda \odot \boldsymbol{L}_\lambda \tag{6}$$

3.3. Saliency Weight Learning Module

Influenced by attributes such as color, contrast, and texture, human perception tends to focus on certain parts of a scene, namely regions of saliency. Saliency detection and UIQA are essentially related, as they both depend on how HVS perceives images. If visual saliency is affected, it implies a change in visual quality. Moreover, visual salient regions are also crucial aspects that UIQA needs to consider. Figure 3 shows different enhanced underwater images and their corresponding saliency maps, including two control groups of high and low quality. It can be observed that different enhancement effects lead to varying degrees of improvement in the results of saliency detection, which can reflect the quality of the input enhanced images. Therefore, to better perceive local quality, SWLM is designed to extract saliency features. This module indirectly reflects image-quality changes introduced by UIE algorithms and assigns visual weights to different regions. The specific framework is shown in Figure 4.



(a) Enhanced underwater images.



(b) Saliency maps.

Figure 3. Illustration of the enhanced underwater images with different qualities and the corresponding saliency maps.

Firstly, the method proposed by Montalone [54] is utilized to calculate the saliency map. For an enhanced underwater image $\boldsymbol{I}$, its saliency map is denoted as $\boldsymbol{I}_s$, which is input into two convolutional-activation layers to expand channel dimensions and obtain low-level saliency features $\overline{\boldsymbol{I}}_s$, which can be represented as follows:

$$\overline{\boldsymbol{I}}_s = Relu(Conv(Relu(Conv(\boldsymbol{I}_s)))) \tag{7}$$

Then, a dual-branch network is designed to obtain the perceptual features of $\overline{\boldsymbol{I}}_s$ in both spatial and channel. The spatial branch aims to capture complex spatial relationships, while the channel branch aims to capture uneven channel degeneration. For the channel branch, it includes a global average pooling layer and a global max pooling layer to refine the feature, followed by a fully connected layer and sigmoid activation function to generate channel weights $\boldsymbol{V}$, which can be expressed as follows:

$$\boldsymbol{V} = Sig\left(FC\left(GAP\left(\overline{\boldsymbol{I}}_s\right) © \, GMP\left(\overline{\boldsymbol{I}}_s\right)\right)\right) \tag{8}$$

where $Sig(\cdot)$ represents the sigmoid activation function, $GAP(\cdot)$ represents global average pooling, and © represents splicing on a specific dimension.

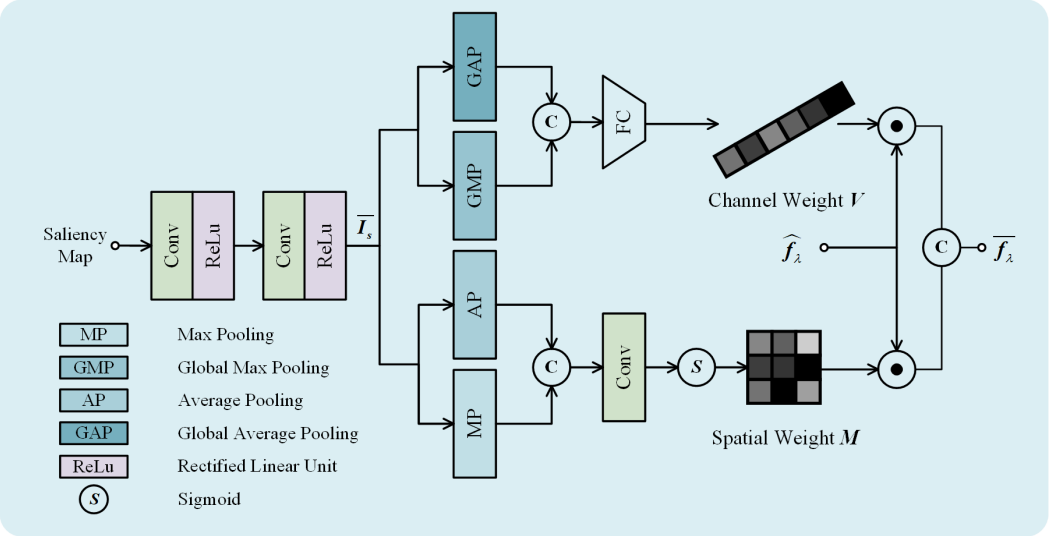


Figure 4. Framework of the proposed SWLM.

For the spatial branch, it contains an average pooling layer and a max pooling layer, followed by a convolution layer and sigmoid activation function to obtain the spatial weight $\boldsymbol{M}$, which can be expressed as follows:

$$\boldsymbol{M} = Sig\left(Conv\left(AP\left(\overline{\boldsymbol{I}}_s\right) © \, MP\left(\overline{\boldsymbol{I}}_s\right)\right)\right) \tag{9}$$

where $AP(\cdot)$ and $MP(\cdot)$ represent average pooling and max pooling, respectively.

To maintain correspondence with the size of the multi-scale feature $\widehat{f}_\lambda$, $\boldsymbol{M}_\lambda$ and $\boldsymbol{V}_\lambda$ are obtained by adjusting the output dimension of $FC(\cdot)$ and the convolution kernel size of $Conv(\cdot)$. Finally, the feature $\widehat{f}_\lambda$ is weighted with the corresponding $\boldsymbol{M}_\lambda$ and $\boldsymbol{V}_\lambda$ obtained from the saliency prior, and concatenated on the channel dimension to obtain weighted features $\overline{f}_\lambda$, which can be expressed as follows:

$$\overline{f}_\lambda = \left(\widehat{f}_\lambda \odot \boldsymbol{V}_\lambda + \widehat{f}_\lambda\right) © \left(\widehat{f}_\lambda \odot \boldsymbol{M}_\lambda + \widehat{f}_\lambda\right) \tag{10}$$

3.4. Attention Feature Fusion Module

Feature fusion, as a common component of DL models, typically integrates low-level and high-level features through long skip connections or summation to obtain features

with high resolution and strong semantic information [55]. Finding the appropriate fusion method is crucial for achieving the desired goals for different tasks. For UIQA, the HVS exhibits multi-scale perception characteristics, where features at different scales have complementary advantages. Low-level features possess a small receptive field, demonstrating strong representations of geometric details that can reveal texture structure distortion. Conversely, high-level features have large receptive fields and a robust ability to represent semantic information, capturing global distortions such as the haze effect. To this end, considering that employing an attention module can enhance the understanding of contextual information within complex features and emphasize more important parts [56–58], we propose an attention-based fusion module that is used to model the distortion information contained between multi-scale features. The main framework is illustrated in Figure 5.

Firstly, to achieve dimensional transformation and feature extraction, the input feature is divided into fixed-size patches and linearly mapped, thus enabling the capture of essential representations for subsequent relationship modeling. Multi-scale feature $\overline{f}_\lambda$ has different receptive fields and resolutions. To ensure the mapped patches maintain consistency in image content, different convolutions are used for patch embedding, followed by a flattened layer to reduce the dimension. The size of the convolution kernel is consistent with the patch size, and the number of output channels is consistent with the embedding size. After feature mapping, the spatial and channel attention are calculated independently. Spatial attention is used to model multi-scale contextual information as local features, enhancing expressive ability. Different scales of contextual information help capture the relationship between structure and texture in enhanced images. Then, the spatial attention feature $\boldsymbol{F}_s$ is obtained as follows:

$$\boldsymbol{F}_s = \mathrm{Softmax}\left(\frac{l\left(\overline{f}_2\right) l\left(\overline{f}_3\right)^T}{\sqrt{d}}\right) l\left(\overline{f}_4\right) \tag{11}$$

where $l(\cdot)$ represents linear mapping, $Softmax(\cdot)$ represents the softmax function, $\sqrt{d}$ represents the square root of the feature dimension, and T represents transpose.

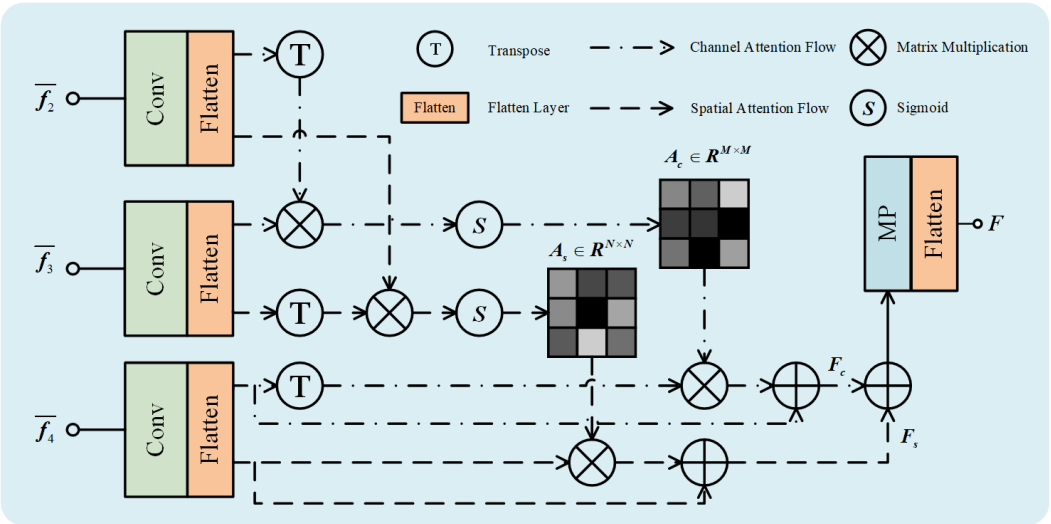


Figure 5. Framework of the proposed AFFM.

In addition, to model the interdependence between different channels and improve the perception of key channel information, we utilize channel attention feature $\boldsymbol{F}_c$, which is obtained as follows:

$$\boldsymbol{F}_c = \mathrm{Softmax}\left(\frac{l\left(\overline{f}_2\right)^T l\left(\overline{f}_3\right)}{\sqrt{d}}\right) l\left(\overline{f}_4\right) \tag{12}$$

Finally, the perceptual feature $\boldsymbol{F}$ is obtained by element-wise addition $\boldsymbol{F}_c$ and $\boldsymbol{F}_s$, which can be expressed as follows:

$$\boldsymbol{F} = \boldsymbol{F}_c + \boldsymbol{F}_s \tag{13}$$

3.5. Regression

By combining the luminance statistical vector $\boldsymbol{V}_{\mathrm{L}}$ and perceptual feature $\boldsymbol{F}$, the final quality feature is obtained, which is then input into the fully connected layer to calculate the final quality score S, which can be expressed as follows:

$$S = FC(\boldsymbol{F} \odot \boldsymbol{V}_{\mathrm{L}}) \tag{14}$$

The optimization of perceived quality prediction results employs MSE as the loss function L_{MSE}, which can be expressed as follows:

$$L_{\mathrm{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (Q - S)^2 \tag{15}$$

where S represents the quality predicted by the proposed JLSAU, and Q represents the Mean Opinion Score (MOS).

4. Experimental Results and Analysis

In this section, we first introduce the experimental settings in detail, including datasets, specific implementation details, and criteria for performance evaluation. Through comparison with a series of state-of-the-art IQA methods, the superior performance of JLSAU has been demonstrated, and is accompanied by extensive ablation studies to validate the contributions of key modules within JLSAU.

4.1. Experimental Settings

4.1.1. Datasets

UIQE [27] dataset, UWIQA [25] and UID2021 [59] datasets are used to train and test the proposed JLSAU. These datasets are composed of real underwater images and enhancement results of various UIE algorithms, with each enhanced image having a Mean Opinion Score (MOS) serving as the ground truth. Specifically, the UIQE dataset contains 405 images derived from 45 real underwater images processed by nine representative UIE algorithms. Each real underwater image and its enhanced results are regarded as a scene, and their MOS values are obtained using a five-level classification scale in a laboratory environment, with the scale administered by 25 volunteers employing a dual stimulus strategy. The UWIQA dataset contains 890 real underwater images with varying degrees of distortion, covering a variety of underwater scenes and common distortion types of underwater images. The MOS value for each image is derived by averaging the ratings from three assessments conducted in a laboratory environment by 21 observers with relevant expertise, following outlier removal. The UID2021 dataset consists of 960 images, including 60 scenes and enhanced results from 15 UIE algorithms, covering six common underwater scenes. The annotations for the UID2021 dataset are obtained through pairwise comparison and sorting by 52 observers in a laboratory environment.

4.1.2. Implementation Details

The dataset is randomly split into a training set and a testing set in a ratio of 8:2 based on scenes. During the training phase, data augmentation techniques, such as random horizontal flipping and random rotation, are applied to enhance the dataset. The enhanced images are segmented by clipping strategy, resulting in image patches of size $3 \times 192 \times 192$, which serve as the input. The proposed JLSAU is implemented based on the PyTorch and trained on an NVIDIA RTX 3080 GPU. The Adam optimizer is used for optimization. The initial learning rate is set to 8×10^{-4} and is updated by the cosine annealing strategy. MSE

is employed as the loss function, and the batch size is set to 16, with a total of 300 epochs during training.

4.1.3. Evaluation Criteria

The performance of the proposed JLSAU is evaluated using three widely recognized indicators, including Spearman Rank Order Correlation Coefficient (SROCC), Kendall Rank Order Correlation Coefficient (KROCC), and Pearson Linear Correlation Coefficient (PLCC). These indicators measure the consistency between the score predicted by the objective method and the subjective MOS. Specifically, SROCC and KROCC are utilized to evaluate the accuracy of image quality prediction ranking, while PLCC measures the precision of objective score prediction. When SROCC, PLCC, and KROCC values approach 1, it indicates a higher consistency between the predicted quality scores and human subjective perception of quality.

4.2. Performance Comparisons

To establish the superiority of the proposed JLSAU, a variety of mainstream NR-IQA methods are selected for comparison. These includes two categories of methods, i.e., in-air IQA and UIQA. The in-air methods include DIIVINE [9], BRISQUE [8], BMPRI [11], CNN-IQA [14], TReS [17], and HyperIQA [19]. UIQA methods consist of UIQM [22], UCIQE [21], CCF [23], FDUM [25], UIQEI [27], NUIQ [24], Twice-Mixing [33], Uranker [31], UIQI [28], and CSN [30]. Both categories include methods based on traditional regression and deep learning. To ensure a fair comparison, all experiments are conducted using the source code published by the respective authors. For traditional regression methods, their source code is used to extract handcrafted features and then retrain support vector machines (SVMs) under the same conditions. For DL-based methods, the testing principles align with those of the proposed JLSAU, and network parameters are optimized to achieve the best performance.

Table 1 shows the performance comparison results for the UIQE dataset and the UWIQA dataset. For each column of comparative indicators, the best results are highlighted in bold. For the UIQE dataset, the proposed JLSAU demonstrates the best performance, with SROCC, KROCC, and PLCC reaching 0.9102, 0.7433, and 0.9233, respectively. For traditional in-air IQA methods, most of them, like DIIVINE, have poor performance, showing weak correlations. Even the best-performing traditional method, BRISQUE, still has a large gap with the proposed JLSAU, with SROCC as 0.7278, KROCC as 0.5000, and PLCC as 0.7507. The reasons can be analyzed as follows: traditional in-air IQA methods are mostly based on the NSS features of in-air images, which have good applicability and reliability in the atmosphere environment. However, there are significant differences in the forms and degrees of degradation between underwater and in-air images. The differing distortion statistical distributions make traditional NSS features unable to accurately reflect the perceived quality of underwater images. In addition, the performance of the proposed JLSAU is better than all the comparing UIQA methods. Compared with the best-performing UIQEI method in UIQA with SROCC, KROCC, and PLCC are 0.8568, 0.6456, and 0.8705, respectively, JLSAU achieves gains of nearly 6.25%, 15.14%, and 6.06% in the former three indicators, respectively. Most of the UIQA methods have limited generalization on the UIQE dataset. In particular, although the traditional UIQA methods (i.e., UIQM, UCIQE, CCF) were proven to have good performance in most underwater scenes at that time, they had a poor correlation with human perception in predicting the quality of enhanced underwater images. Although FDUM designed for raw underwater images is a recently proposed method, it lacks generalization performance when applied to enhanced underwater images. This may be due to the rapid development of UIE algorithms in recent years, as they have introduced complex color and structural changes in the enhanced underwater images that traditional UIQA methods cannot perceive. Additionally, CSN and UIQI attain the fourth and fifth position in the UIQA methods, demonstrating a certain degree of predictive accuracy. This may be attributed to the introduction of visual saliency by CSN, which

is more consistent with human perception, while UIQI is attributed to consideration of a broader spectrum of attributes in enhanced underwater images. Finally, apart from Twice-Mixing, most DL-based methods (i.e., CNN-IQA, TReS, HyperIQA, Uranker) can achieve better performance and generalization ability than traditional methods. This is because Twice-Mixing performs the mixing in a way that cannot fit the uniform of the HVS and cannot cover the over-enhanced cases, whereas DL-based methods automatically learn high-level features and representations from images, enabling them to better capture high-quality information. Despite potential challenges such as model complexity and hyperparameter settings, DL-based methods generally outperform traditional methods in UIQA tasks. Fortunately, the proposed JLSAU achieves the best performance by considering the distortion characteristics of enhanced underwater images and incorporating knowledge about the priors of HVS.

Table 1. Performance comparison of different IQA methods on the UIQE and UWIQA datasets.

		Method	**UIQE**			**UWIQA**		
			SROCC ↑	**KROCC ↑**	**PLCC ↑**	**SROCC ↑**	**KROCC ↑**	**PLCC ↑**
In-air IQA	Traditional	DIIVINE [9]	0.1278	0.1084	0.0997	0.4399	0.2915	0.3724
		BRISQUE [8]	0.7278	0.5000	0.7507	0.3456	0.2562	0.3669
		BMPRI [11]	0.6152	0.4167	0.6020	0.6815	0.4964	0.6875
	DL-based	CNN-IQA [14]	0.7840	0.5849	0.7765	0.6158	0.4711	0.5327
		TReS [17]	0.8721	0.6924	0.8845	0.6720	0.5365	0.6817
		HyperIQA [19]	0.7114	0.5841	0.7288	0.6501	0.5040	0.6799
Underwater IQA	Traditional	UIQM [22]	0.1556	0.1984	0.3112	0.6180	0.4730	0.6080
		UCIQE [21]	0.3254	0.2248	0.4551	0.6220	0.4740	0.5950
		CCF [23]	0.2556	0.1517	0.3010	0.4790	0.3510	0.4090
		FDUM [25]	0.2343	0.1685	0.3030	0.6830	0.5300	0.6380
		UIQEI [27]	0.8568	0.6456	0.8705	N/A	N/A	N/A
		NUIQ [24]	0.4433	0.3067	0.4023	0.4651	0.3766	0.4702
		UIQI [28]	0.7131	0.5157	0.7270	0.7423	0.5912	0.7412
		CSN [30]	0.7265	0.5258	0.7422	0.7423	0.6014	0.7648
	DL-based	Twice-Mixing [33]	0.5690	0.4142	0.5506	0.4727	0.3501	0.4422
		Uranker [31]	0.8188	0.6504	0.8172	0.5289	0.3992	0.5135
		Proposed	**0.9102**	**0.7433**	**0.9233**	**0.7669**	**0.6193**	**0.7836**

To further validate the superiority and generalization of the proposed JLSAU, a comparative test is carried out on the UWIQA dataset. Unlike the UIQE dataset, the UWIQA dataset is composed of various real underwater images that have not been enhanced, featuring a richer variety of underwater scenes and diverse image sizes. As shown in Table 1, currently, various IQA methods exhibit mediocre performance on the UWIQA dataset, indicating that UIQA remains a challenge in the evaluation of real underwater images. The proposed JLSAU achieves competitive prediction performance, with the SROCC and PLCC values reaching 0.7669 and 0.7836—about 3.31% and 2.46% higher than the second-best CSN method, with SROCC reaching 0.7423 and PLCC reaching 0.7648, respectively. The results of other comparison methods are basically consistent with those on the UIQE dataset. The traditional in-air IQA methods perform poorly when predicting underwater images, and most of the DL-based methods are better than those traditional methods, whether for underwater or in-air images. It is worth mentioning that compared with the UIQE dataset, the traditional UIQA methods (i.e., UIQM, UCIQE, CCF, FDUM) show better performance. This is because the UWIQA dataset is composed of original underwater images without the uncertainty introduced by the UIE algorithms. In a word, compared to existing comparative methods, the proposed JLSAU has advantages in evaluating the perceptual quality of underwater images. The good performance of these two datasets proves the superiority of the proposed JLSAU.

In addition, we also conduct comparison experiments on the large-scale underwater dataset UID2021. Note that UIQEI is not open source and cannot be compared; the remaining comparison methods are consistent with those included in Table 1. As shown in Table 2, the proposed JLSAU achieves the best performance among various UIQA methods. This indicates that even in datasets with richer scenes and more comprehensive distortion coverage, the proposed JLSAU still has certain advantages. However, among all the methods compared, TReS achieves the best results. This may be due to the fact that TReS pay more attention to the relative distance information between each batch of images during design and have a stronger ability to sort relative quality. This ability is amplified when the training data are expanded. This insight is valuable for future considerations in designing UIQA methods.

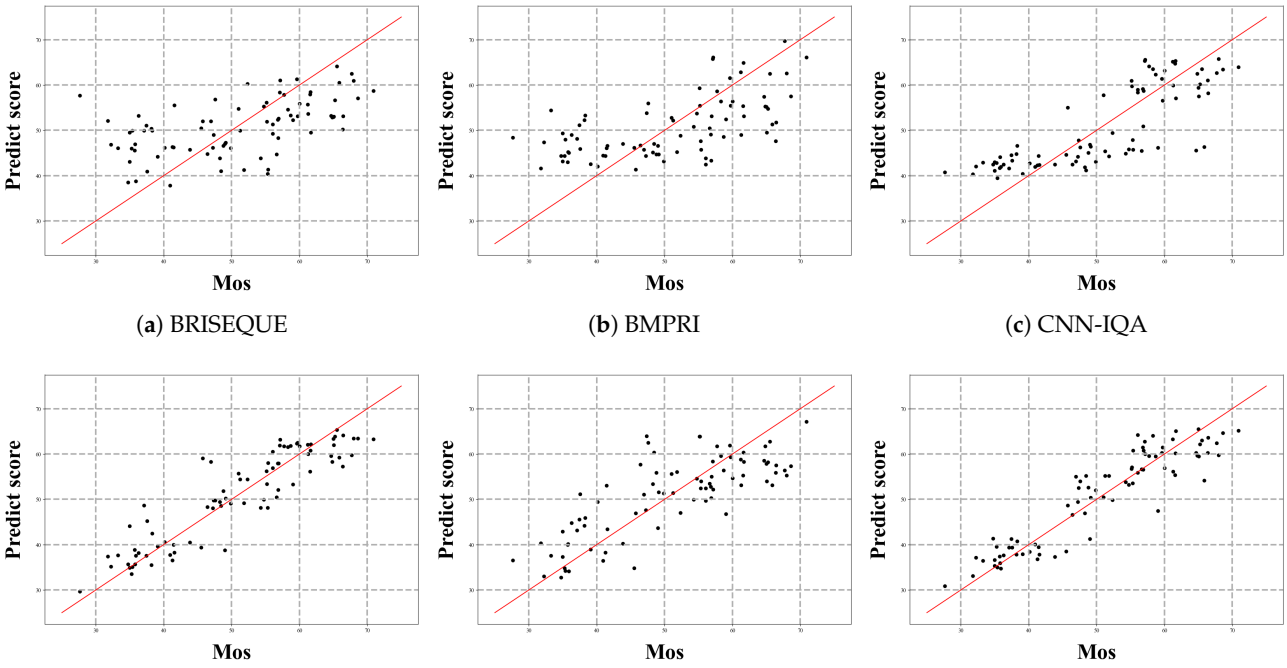
At the same time, to visually demonstrate the performance comparison results, Figure 6 shows the scatter plot of subjective MOS values versus the predicted results of the proposed JLSAU on the UIQE testing set, along with those of other comparison methods. To provide an intuitive comparison within a reasonable length, we chose the top two traditional in-air IQA methods (i.e., BRISQUE, BMPRI), the top two DL-based in-air IQA methods (i.e., CNN-IQA, TReS), and top two DL-based UIQA methods (i.e., UIQEI, JLSAU) to plot the scatter plot. In the scatter plot, the y-axis represents the predicted scores and the x-axis represents the subjective MOS values. Clearly, a scatter plot demonstrating better prediction results should have points that are closer to the fitting curve. It can be easily observed from Figure 6 that the proposed JLSAU produces the best-fitting results on the UIQE dataset, with predicted objective scores closer to the subjective MOS values.

Table 2. Performance comparison of different IQA methods on the UID2021 dataset.

		Method	**UID2021**		
			SROCC ↑	**KROCC ↑**	**PLCC ↑**
In-air IQA	Traditional	DIIVINE [9]	0.6112	0.4363	0.6264
		BRISQUE [8]	0.4689	0.3192	0.4794
		BMPRI [11]	0.5455	0.3823	0.5524
	DL-based	CNN-IQA [14]	0.6257	0.4766	0.6039
		TReS [17]	**0.8335**	**0.6491**	**0.8304**
		HyperIQA [19]	0.8022	0.6073	0.7864
Underwater IQA	Traditional	UIQM [22]	0.5349	0.3785	0.5689
		UCIQE [21]	0.5892	0.4340	0.6335
		CCF [23]	0.4577	0.3314	0.5371
		FDUM [25]	0.6406	0.4589	0.6464
		NUIQ [24]	0.7168	0.5293	0.7266
		UIQI [28]	0.6921	0.5093	0.6794
		CSN [30]	0.7210	0.5157	0.7157
	DL-based	Twice-Mixing [33]	0.6952	0.5113	0.7060
		Uranker [31]	0.7279	0.5448	0.7261
		Proposed	0.7467	0.5509	0.7353

To further demonstrate the ability of the proposed JLSAU to perceive distortion in the enhanced underwater images, a series of images are chosen from the UIQE testing set. The predicted scores from different methods along with the corresponding MOS values are provided, (i.e., from left to right: JLSAU, BRISQUE, BMPRI, CNN-IQA, TReS, and UIQEI). Note that the number in parentheses represents the predicted ranking of the corresponding method for this scene, and methods with correct rankings are highlighted in red. As shown in Figure 7, the images are arranged from left to right based on MOS values, with perceptual quality ranging from low to high. It is evident that the enhanced images closer to the left exhibit lower overall visual perceptual quality, with most of them showing inadequate enhancement and severe global distortions, such as color distortion and low contrast. As enhancement improves, the perceptual quality of the underwater images

gradually increases, addressing issues like color shifts and low contrast, while local details and texture structures become clearer. In Figure 7, for the enhanced images of the same scene, the proposed JLSAU provides an accurate objective ranking. Although the TReS also provides the correct ranking, the scores obtained by JALSU are generally closer to the MOS values, which enables accurate differentiation of the enhancement quality from good to poor. However, most other methods exhibit a wide range of predicted quality for enhanced underwater images at different levels within the same scene, failing to perceive subtle changes and trends, particularly BMPRI and BRISQUE, while TReS and UIQEI perform relatively better in their predictions.



(**a**) BRISEQUE (**b**) BMPRI (**c**) CNN-IQA

(**d**) TReS (**e**) UIQEI (**f**) Proposed

Figure 6. Illustration of scatter plots of prediction results using different methods on the UIQE dataset. Each point represents an enhanced underwater image used for testing.

4.3. Ablation Experiment

The proposed JLSAU consists of several key components: SWLM, LFEM, and AFFM. Table 3 shows the results of a series of ablation studies conducted to analyze the contribution of each module in JLSAU.

Firstly, to evaluate the influence of SWLM on the overall method, we present the performance results after removing this module, as shown in part (c) of Table 3. In comparison to methods incorporating the SWLM (i.e., (a), (b), and (d)), (c) displays inferior performance, with lower values across all three correlation coefficients used for performance evaluation. It is evident that SWLM positively impacts the overall method, improving the ability to predict the quality of enhanced underwater images. This can be attributed to the strong correlation between saliency priors and the HVS, highlighting crucial regions in enhanced underwater images. Saliency features concentrate on visual information in these critical regions, enhancing the perception of local quality.

Subsequently, to demonstrate the effectiveness of the LFEM in JLSAU, we conduct an ablation experiment by removing this module, as shown in (b) of Table 3. The results clearly show that removing the LFEM leads to a general decrease in the correlation coefficients for method (b) by 0.01 to 0.03 on both the UIQE and UWIQA datasets, with a particularly significant performance degradation for the UWIQA dataset. Hence, the LFEM plays an important role in quality assessment. This is because enhanced underwater

images typically manifest distortion in luminance, and human vision is highly sensitive to luminance information. Effectively perceiving changes in luminance information is key to objective assessment.

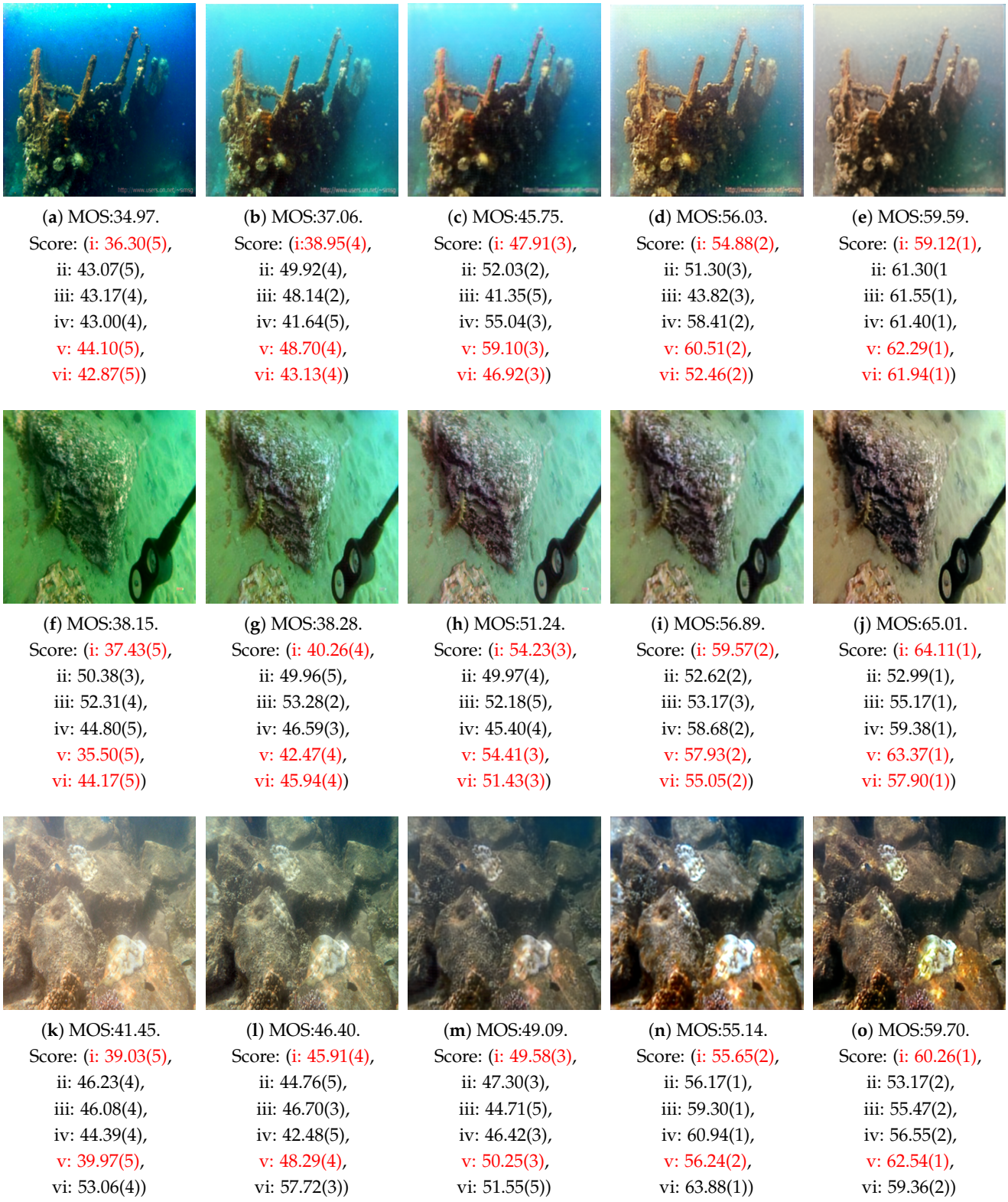


Figure 7. Illustration of enhanced underwater images with increasing quality from left to right and the predicted scores of the proposed JLSAU and the compared methods. i: JLSAU. ii: BRISQUE. iii: BMPRI. iv: CNN-IQA. v: TReS. vi: UIQEI.

Table 3. Ablation experiments on the effects of different components.

Method	SWLM	LFEM	AFFM	UIQE			UWIQA		
				SROCC	KROCC	PLCC	SROCC	KROCC	PLCC
(a)	✓	✓		0.9049	0.7412	0.9213	0.7542	0.6074	0.7698
(b)	✓		✓	0.9017	0.7356	0.9200	0.7309	0.5929	0.7591
(c)		✓	✓	0.8989	0.7382	0.9163	0.7375	0.5935	0.7708
(d)	✓	✓	✓	**0.9102**	**0.7433**	**0.9233**	**0.7669**	**0.6193**	**0.7836**

Finally, the ablation experiment of AFFM is conducted, and the performance without AFFM is shown in part (a) of Table 3. Compared to the proposed JLSAU, method (a) only utilizes $\overline{f}_4$ for quality regression without feature fusion. The results indicate that the three correlation coefficients on the UIQE and UWIQA datasets decrease without the AFFM, performing worse than the JLSAU. This may be due to high-level features containing more global semantic information, while their ability to capture local structural information is inadequate, leading to reduced perception of local distortions. Through feature fusion, JLSAU effectively integrates the abundant local information from low-level features with the global information embedded in high-level features, enhancing the perception of distortions. Therefore, employing a suitable feature fusion method is advantageous for perceiving multi-scale distortions in enhanced underwater images.

5. Further Discussion

As mentioned previously, we visualize the comparison of enhanced underwater images of different qualities before and after histogram equalization in Figure 2. Generally, high-quality enhanced images exhibit smaller changes after equalization, resulting in greater similarity between $\boldsymbol{I}$ and $\overline{\boldsymbol{I}}$. Consequently, the mapping distance between them is smaller. The similarity between $\boldsymbol{I}$ and $\overline{\boldsymbol{I}}$ should have a positive correlation with image quality. Therefore, we employ indicators such as cosine similarity, SSIM, histogram similarity distance, and pHash to evaluate the similarity between $\boldsymbol{I}$ and $\overline{\boldsymbol{I}}$. Each indicator is utilized as a luminance feature in place of LFEM, and their ability to represent mapping relationships is compared in Table 4 to verify the superiority of LFEM. As shown in Table 4, experiments are conducted on the UIQE dataset. Compared to SSIM, histogram similarity distance, and cosine similarity, the proposed JLSAU using LFEM generally achieves the best overall performance in learning the correlation between mapping relationships and image quality. This indicates its advantage in learning high-dimensional mapping relationships, with the second-best result obtained by pHash. Therefore, to better model the consistent correlation with perceived quality, we propose using LFEM to learn quality-aware representations of luminance distribution.

Table 4. Comparison of different methods in representing mapping relationships.

Method	UIQE		
	SROCC	KROCC	PLCC
Cosine Similarity	0.9051	0.7291	0.9218
SSIM	0.8759	0.6879	0.8991
Histogram Similarity Distance	0.8841	0.7088	0.9125
pHash	**0.9164**	0.7405	0.9223
JLSAU	0.9102	**0.7433**	**0.9233**

To validate the role of the convolutional block in LFEM, ablation experiments are conducted while keeping all structures consistent except for the convolutional block. The results are shown in Table 5, where "W/O" represents without and "W/" represents with, indicating a notable overall performance decline after the removal of the convolutional block. This decline phenomenon is observed in both the UIQE and UWIQA datasets,

highlighting the significance of the convolutional block in LFEM. This importance may stem from the ability of CNN to complement missing spatial information in grayscale statistics through relative positional encoding, thereby enhancing the perception of luminance information comprehensively.

Table 5. Ablation study for LFEM.

Dataset	Method	SROCC	KROCC	PLCC
UIQE	W/O conv network	0.9072	0.7383	0.9068
	W/ conv network	**0.9102**	**0.7433**	**0.9233**
UWIQA	W/O conv network	0.7610	0.6166	0.7724
	W/ conv network	**0.7669**	**0.6193**	**0.7836**

To assess the impact of various saliency maps in the SWLM on the performance of JLSAU, we select four state-of-the-art salient object detection methods and conduct ablation experiments on the UIQE dataset. The overall structure of the model remains unchanged, and saliency maps obtained by different methods are used as inputs. The results are presented in Table 6. It can be observed that the saliency maps obtained by different methods indeed affect the overall performance of quality assessment. Fortunately, the proposed method consistently achieves the best PLCC and KROCC, demonstrating the superiority of the selected saliency map acquisition method. Compared to the selected method, these saliency methods used for comparison have a weaker ability to distinguish between different-quality underwater images, which may be due to their stronger generalization capabilities. Even in the presence of various distortions, the obtained saliency maps remain relatively consistent.

Table 6. Ablation experiment of saliency maps obtained by different methods.

Method	PLCC	KROCC	SROCC
GPONET [60]	0.9162	0.7364	0.9113
PEEKBOO [61]	0.9171	0.7407	0.9149
PGNet [62]	0.9167	0.7204	0.9018
ADM [63]	0.9210	0.7370	**0.9185**
(Proposed) VSFs [54]	**0.9233**	**0.7433**	0.9102

Furthermore, the impact of feature selection in the backbone network on the experimental results is explored. Specifically, while keeping the overall structure unchanged, some necessary hyperparameters within the module are modified to match the dimensions of features at different stages for training. The results, shown in Table 7, reveal that the model generally achieves optimal correlation indicators with the feature combination of $\overline{f}_2$, $\overline{f}_3$, and $\overline{f}_4$ (i.e., (4)), compared to other combinations, indicating superior performance. Additionally, within combinations containing $\overline{f}_4$ (i.e., (2), (3), (4)), generally better results are obtained compared to those obtained for (1). This is attributed to the superior semantic representation and richer global information provided by high-level features, which reflect the quality of enhanced underwater images at a deeper level. Therefore, the features $\overline{f}_2$, $\overline{f}_3$, and $\overline{f}_4$ are selected in the proposed JLSAU.

In addition, an effective UIQA method should have relatively low complexity. 50 images from UIQE dataset are utilized to test the running time of various methods, and the average time spent is taken as the experimental result, as shown in Table 8. The experiment is carried out on a computer with an Intel i5-12500 4.08 GHz CPU, NVIDIA RTX 3070ti GPU, and 32 GB memory using matlab2016b. It can be found that the average running time of the proposed JLSAU is less than that of DIIVINE, BMPRI, TReS, HyperIQA, CCF, FDUM, UIQI, Twice-Mixing, and Uranker. Compared with BRISQUE, CNN-IQA, UIQM, UCIQE, and CSN, JLSAU runs a little longer, but achieves better performance.

Table 7. Feature selection study for AFFM.

Method	$\overline{f}_1$	$\overline{f}_2$	$\overline{f}_3$	$\overline{f}_4$	UIQE			UWIQA		
					SROCC	KROCC	PLCC	SROCC	KROCC	PLCC
(1)	✓	✓	✓		0.8933	0.7247	0.9069	0.7455	0.6031	0.7648
(2)	✓	✓		✓	0.9045	0.7300	0.9095	**0.7729**	**0.6239**	0.7632
(3)	✓		✓	✓	0.9085	0.7326	0.9068	0.7455	0.6004	0.7824
(4)		✓	✓	✓	**0.9102**	**0.7433**	**0.9233**	0.7669	0.6193	**0.7836**

Table 8. Computational time (measured in seconds) comparison of IQA models.

Method	DIIVINE	BRISQUE	BMPRI	CNN-IQA	TReS	HyperIQA	UCIQE	UIQM
Time/s	2.0374	**0.0134**	0.2068	0.0572	0.0758	0.1677	0.0186	0.0500
Method	CCF	FDUM	UIQI	CSN	Twice-Mixing		Uranker	Proposed
Time/s	0.1060	0.2550	0.2358	0.0328	0.1065		0.0786	0.0733

6. Conclusions

In this paper, we present JLSAU, a multi-scale underwater image quality assessment (UIQA) method that joins luminance priors and saliency priors. By utilizing a pyramid-structured backbone network, it extracts multi-scale features that align with the perceptual characteristics of the Human Visual System (HVS), with a particular focus on distortions at different scales, such as texture details and haze effects. Then, to measure the enhancement effect of Underwater Image Enhancement (UIE) algorithms, we extract visual information such as luminance and saliency as supplements. The luminance priors obtained utilizing histogram statistics to learn the luminance distribution and by designing a convolutional network to encode positional information missing from statistical features. Furthermore, to improve the perception of structure and details locally, we separately learn saliency features in both the channel and spatial domains, reflecting changes in visual attention. Finally, by leveraging the attention mechanism, we effectively model the rich perceptual quality and enhancement information contained in the multi-scale features of enhanced images, thereby comprehensively improving the representation capability of image quality. Compared to existing UIQA methods, the proposed JLSAU demonstrates superior performance. However, there are still some limitations, such as the lack of a dedicated color prior for addressing color distortions, and the saliency detection methods used have yet to be optimized. In the future, we plan to further explore the incorporation of color priors and develop a saliency detection method tailored specifically to underwater images to better leverage the saliency prior.

Author Contributions: Conceptualization, Z.H. and Y.C.; Data curation, C.J.; Formal analysis, C.J.; Funding acquisition, Z.H., T.L. and Y.C.; Methodology, Z.L.; Validation, Z.H. and C.J.; Writing—original draft, Z.L.; Writing—review and editing, Z.H. and T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported, in part, by the National Natural Science Foundation of China under Grant (No. 61971247), in part by Natural Science Foundation of Zhejiang Province under Grant (Nos. LQ23F010011, LY22F020020, LQ24F010002), and in part by the Zhejiang Provincial Postdoctoral Research Excellence Foundation under Grant ZJ2022130.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sun, K.; Tian, Y. Dbfnet: A dual-branch fusion network for underwater image enhancement. *Remote Sens.* **2023**, *15*, 1195. [CrossRef]
2. Schettini, R.; Corchs, S. Underwater image processing: State of the art of restoration and image enhancement methods. *EURASIP J. Adv. Signal Process.* **2010**, *2010*, 1–14. [CrossRef]
3. Wu, J.; Liu, X.; Qin, N.; Lu, Q.; Zhu, X. Two-Stage Progressive Underwater Image Enhancement. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–18. [CrossRef]
4. Berga, D.; Gallés, P.; Takáts, K.; Mohedano, E.; Riordan-Chen, L.; Garcia-Moll, C.; Vilaseca, D.; Marín, J. QMRNet: Quality Metric Regression for EO Image Quality Assessment and Super-Resolution. *Remote Sens.* **2023**, *15*, 2451. [CrossRef]
5. Hao, X.; Li, X.; Wu, J.; Wei, B.; Song, Y.; Li, B. A No-Reference Quality Assessment Method for Hyperspectral Sharpened Images via Benford's Law. *Remote Sens.* **2024**, *16*, 1167. [CrossRef]
6. Li, Y.; Dong, Y.; Li, H.; Liu, D.; Xue, F.; Gao, D. No-Reference Hyperspectral Image Quality Assessment via Ranking Feature Learning. *Remote Sens.* **2024**, *16*, 1657. [CrossRef]
7. Cui, Y.; Jiang, G.; Yu, M.; Chen, Y.; Ho, Y.S. Stitched Wide Field of View Light Field Image Quality Assessment: Benchmark Database and Objective Metric. *IEEE Trans. Multimed.* **2024**, *26*, 5092–5107. [CrossRef]
8. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [CrossRef]
9. Moorthy, A.K.; Bovik, A.C. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Trans. Image Process.* **2011**, *20*, 3350–3364. [CrossRef]
10. Saad, M.A.; Bovik, A.C.; Charrier, C. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Trans. Image Process.* **2012**, *21*, 3339–3352. [CrossRef]
11. Min, X.; Zhai, G.; Gu, K.; Liu, Y.; Yang, X. Blind image quality estimation via distortion aggravation. *IEEE Trans. Broadcast.* **2018**, *64*, 508–517. [CrossRef]
12. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
13. Liu, X.; Van De Weijer, J.; Bagdanov, A.D. Rankiqa: Learning from rankings for no-reference image quality assessment. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1040–1049.
14. Kang, L.; Ye, P.; Li, Y.; Doermann, D. Convolutional neural networks for no-reference image quality assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1733–1740.
15. Yue, G.; Hou, C.; Zhou, T.; Zhang, X. Effective and efficient blind quality evaluator for contrast distorted images. *IEEE Trans. Instrum. Meas.* **2018**, *68*, 2733–2741. [CrossRef]
16. Zhu, H.; Li, L.; Wu, J.; Dong, W.; Shi, G. MetaIQA: Deep meta-learning for no-reference image quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 14143–14152.
17. Golestaneh, S.A.; Dadsetan, S.; Kitani, K.M. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 1220–1230.
18. Yang, C.; An, P.; Shen, L. Blind image quality measurement via data-driven transform-based feature enhancement. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [CrossRef]
19. Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; Zhang, Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3667–3676.
20. Zhao, W.; Li, M.; Xu, L.; Sun, Y.; Zhao, Z.; Zhai, Y. A Multibranch Network With Multilayer Feature Fusion for No-Reference Image Quality Assessment. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–11. [CrossRef]
21. Yang, M.; Sowmya, A. An underwater color image quality evaluation metric. *IEEE Trans. Image Process.* **2015**, *24*, 6062–6071. [CrossRef]
22. Panetta, K.; Gao, C.; Agaian, S. Human-visual-system-inspired underwater image quality measures. *IEEE J. Ocean. Eng.* **2015**, *41*, 541–551. [CrossRef]
23. Wang, Y.; Li, N.; Li, Z.; Gu, Z.; Zheng, H.; Zheng, B.; Sun, M. An imaging-inspired no-reference underwater color image quality assessment metric. *Comput. Electr. Eng.* **2018**, *70*, 904–913. [CrossRef]
24. Jiang, Q.; Gu, Y.; Li, C.; Cong, R.; Shao, F. Underwater image enhancement quality evaluation: Benchmark dataset and objective metric. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5959–5974. [CrossRef]
25. Yang, N.; Zhong, Q.; Li, K.; Cong, R.; Zhao, Y.; Kwong, S. A reference-free underwater image quality assessment metric in frequency domain. *Signal Process. Image Commun.* **2021**, *94*, 116218. [CrossRef]
26. Zheng, Y.; Chen, W.; Lin, R.; Zhao, T.; Le Callet, P. UIF: An objective quality assessment for underwater image enhancement. *IEEE Trans. Image Process.* **2022**, *31*, 5456–5468. [CrossRef]
27. Li, W.; Lin, C.; Luo, T.; Li, H.; Xu, H.; Wang, L. Subjective and objective quality evaluation for underwater image enhancement and restoration. *Symmetry* **2022**, *14*, 558. [CrossRef]

28. Liu, Y.; Gu, K.; Cao, J.; Wang, S.; Zhai, G.; Dong, J.; Kwong, S. UIQI: A Comprehensive Quality Evaluation Index for Underwater Images. *IEEE Trans. Multimed.* **2024**, *26*, 2560–2573. [CrossRef]
29. Zhang, S.; Li, Y.; Tan, L.; Yang, H.; Hou, G. A no-reference underwater image quality evaluator via quality-aware features. *J. Vis. Commun. Image Represent.* **2023**, *97*, 103979. [CrossRef]
30. Hou, G.; Zhang, S.; Lu, T.; Li, Y.; Pan, Z.; Huang, B. No-reference quality assessment for underwater images. *Comput. Electr. Eng.* **2024**, *118*, 109293. [CrossRef]
31. Guo, C.; Wu, R.; Jin, X.; Han, L.; Zhang, W.; Chai, Z.; Li, C. Underwater ranker: Learn which is better and how to be better. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 702–709.
32. Li, M.; Lin, Y.; Shen, L.; Wang, Z.; Wang, K.; Wang, Z. Human perceptual quality driven underwater image enhancement framework. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
33. Fu, Z.; Fu, X.; Huang, Y.; Ding, X. Twice mixing: A rank learning based quality assessment approach for underwater image enhancement. *Signal Process. Image Commun.* **2022**, *102*, 116622. [CrossRef]
34. Zhang, L.; Shen, Y.; Li, H. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Trans. Image Process.* **2014**, *23*, 4270–4281. [CrossRef] [PubMed]
35. Zhu, M.; Hou, G.; Chen, X.; Xie, J.; Lu, H.; Che, J. Saliency-guided transformer network combined with local embedding for no-reference image quality assessment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual Conference, 11–17 October 2021; pp. 1953–1962.
36. Jaffe, J.S. Underwater optical imaging: The past, the present, and the prospects. *IEEE J. Ocean. Eng.* **2014**, *40*, 683–700. [CrossRef]
37. Peng, Y.T.; Cosman, P.C. Underwater image restoration based on image blurriness and light absorption. *IEEE Trans. Image Process.* **2017**, *26*, 1579–1594. [CrossRef]
38. Zhao, X.; Jin, T.; Qu, S. Deriving inherent optical properties from background color and underwater image enhancement. *Ocean Eng.* **2015**, *94*, 163–172. [CrossRef]
39. Drews, P.L.; Nascimento, E.R.; Botelho, S.S.; Campos, M.F.M. Underwater depth estimation and image restoration based on single images. *IEEE Comput. Graph. Appl.* **2016**, *36*, 24–35. [CrossRef] [PubMed]
40. Li, Y.; Hou, G.; Zhuang, P.; Pan, Z. Dual High-Order Total Variation Model for Underwater Image Restoration. *arXiv* **2024**, arXiv:2407.14868.
41. Fu, X.; Fan, Z.; Ling, M.; Huang, Y.; Ding, X. Two-step approach for single underwater image enhancement. In Proceedings of the 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Phuket, Thailand, 24–27 October 2016; IEEE: Piscataway, NJ, USA, 2017; pp. 789–794.
42. Li, X.; Hou, G.; Li, K.; Pan, Z. Enhancing underwater image via adaptive color and contrast enhancement, and denoising. *Eng. Appl. Artif. Intell.* **2022**, *111*, 104759. [CrossRef]
43. Lu, S.; Guan, F.; Zhang, H.; Lai, H. Speed-Up DDPM for Real-Time Underwater Image Enhancement. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 3576–3588. [CrossRef]
44. Wang, Y.; Zhao, Y.; Pan, H.; Zhou, W. An improved reinforcement learning method for underwater image enhancement. In Proceedings of the 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Hangzhou, China, 4–6 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1077–1082.
45. Sun, S.; Wang, H.; Zhang, H.; Li, M.; Xiang, M.; Luo, C.; Ren, P. Underwater image enhancement with reinforcement learning. *IEEE J. Ocean. Eng.* **2022**, *49*, 249–261. [CrossRef]
46. Wang, H.; Sun, S.; Bai, X.; Wang, J.; Ren, P. A reinforcement learning paradigm of configuring visual enhancement for object detection in underwater scenes. *IEEE J. Ocean. Eng.* **2023**, *48*, 443–461. [CrossRef]
47. Wang, H.; Sun, S.; Chang, L.; Li, H.; Zhang, W.; Frery, A.C.; Ren, P. INSPIRATION: A reinforcement learning-based human visual perception-driven image enhancement paradigm for underwater scenes. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108411. [CrossRef]
48. Wang, H.; Zhang, W.; Ren, P. Self-organized underwater image enhancement. *ISPRS J. Photogramm. Remote Sens.* **2024**, *215*, 1–14. [CrossRef]
49. Wang, H.; Sun, S.; Ren, P. Meta underwater camera: A smart protocol for underwater image enhancement. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 462–481. [CrossRef]
50. Song, W.; Shen, Z.; Zhang, M.; Wang, Y.; Liotta, A. A hierarchical probabilistic underwater image enhancement model with reinforcement tuning. *J. Vis. Commun. Image Represent.* **2024**, *98*, 104052. [CrossRef]
51. Wang, H.; Zhang, W.; Bai, L.; Ren, P. Metalantis: A Comprehensive Underwater Image Enhancement Framework. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–19. [CrossRef]
52. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
53. Islam, M.A.; Jia, S.; Bruce, N.D. How much position information do convolutional neural networks encode? *arXiv* **2020**, arXiv:2001.08248.
54. Montabone, S.; Soto, A. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image Vis. Comput.* **2010**, *28*, 391–402. [CrossRef]
55. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional feature fusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual Conference, 5–9 January 2021; pp. 3560–3569.

56. Yi, W.; Dong, L.; Liu, M.; Zhao, Y.; Hui, M.; Kong, L. DCNet: Dual-cascade network for single image dehazing. *Neural Comput. Appl.* **2022**, *34*, 16771–16783. [CrossRef]
57. Chen, J.; Wen, Y.; Nanehkaran, Y.A.; Zhang, D.; Zeb, A. Multiscale attention networks for pavement defect detection. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–12. [CrossRef]
58. Yi, W.; Dong, L.; Liu, M.; Hui, M.; Kong, L.; Zhao, Y. MFAF-Net: Image dehazing with multi-level features and adaptive fusion. *Vis. Comput.* **2024**, *40*, 2293–2307. [CrossRef]
59. Hou, G.; Li, Y.; Yang, H.; Li, K.; Pan, Z. UID2021: An underwater image dataset for evaluation of no-reference quality assessment metrics. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–24. [CrossRef]
60. Yi, Y.; Zhang, N.; Zhou, W.; Shi, Y.; Xie, G.; Wang, J. GPONet: A two-stream gated progressive optimization network for salient object detection. *Pattern Recognit.* **2024**, *150*, 110330. [CrossRef]
61. Zunair, H.; Hamza, A.B. PEEKABOO: Hiding parts of an image for unsupervised object localization. *arXiv* **2024**, arXiv:2407.17628.
62. Xie, C.; Xia, C.; Ma, M.; Zhao, Z.; Chen, X.; Li, J. Pyramid grafting network for one-stage high resolution saliency detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11717–11726.
63. Zhou, X.; Shen, K.; Liu, Z. ADMNet: Attention-guided Densely Multi-scale Network for Lightweight Salient Object Detection. *IEEE Trans. Multimed.* **2024**, 1–14. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Remote Sensing Editorial Office
E-mail: remotesensing@mdpi.com
www.mdpi.com/journal/remotesensing

Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editors. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editors and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-3842-4