

Article

Mobile_ViT: Underwater Acoustic Target Recognition Method Based on Local–Global Feature Fusion

Haiyang Yao ^{1,2} , Tian Gao ^{1,2} , Yong Wang ³, Haiyan Wang ^{1,4,*} and Xiao Chen ^{1,2}

¹ School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China; yaoHY1991@126.com (H.Y.); 211612052@sust.edu.cn (T.G.); chenxiao@sust.edu.cn (X.C.)

² Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China

³ Xi'an Microelectronics Technology Institute, Xi'an 710021, China; dyjay999@163.com

⁴ School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China

* Correspondence: hywang@sust.edu.cn

Abstract: To overcome the challenges of inadequate representation and ineffective information exchange stemming from feature homogenization in underwater acoustic target recognition, we introduce a hybrid network named Mobile_ViT, which synergizes MobileNet and Transformer architectures. The network begins with a convolutional backbone incorporating an embedded coordinate attention mechanism to enhance the local details of inputs. This mechanism captures the long-term temporal dependencies and precise frequency–domain relationships of signals, focusing the features on the time–frequency positions. Subsequently, the Transformer's Encoder is integrated at the end of the backbone to facilitate global characterization, thus effectively overcoming the convolutional neural network's shortcomings in capturing long-range feature dependencies. Evaluation on the Shipsear and DeepShip datasets yields accuracies of 98.50% and 94.57%, respectively, marking a substantial improvement over the baseline. Notably, the proposed method also demonstrates obvious separation coefficients, signifying enhanced clustering effectiveness, and is lighter than other Transformers.

Keywords: underwater acoustic target recognition; attention mechanism; feature fusion; MobileNet; Transformer



Citation: 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. <https://doi.org/10.3390/jmse12040589>

Received: 29 February 2024

Revised: 24 March 2024

Accepted: 27 March 2024

Published: 29 March 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

As human development and utilization of the ocean intensify, along with comprehensive management efforts, the ocean is no longer seen merely as a collection of islands, coastlines, and vast waters. It is now understood as a complex system, encompassing the marine environment, marine equipment, and human activities [1]. This complexity has made the development of deep-sea and distant-sea information sensing technologies a leading trend. Marine communication technology offers a direct and convenient means to comprehend this intricate system, making the establishment of a marine communication network an effective approach [2]. However, achieving marine communication requires a suitable carrier, and both light and electromagnetic wave propagation in water suffers exponential attenuation, rendering it impractical for underwater networks. Acoustic waves, capable of propagating over long distances underwater, emerge as the sole energy form suitable for this purpose, serving as the primary medium for underwater signal and information dissemination [3].

Underwater acoustic target recognition (UATR), which employs sonar to capture ship-radiated noise to determine the target's nature, is a vital research domain in hydroacoustic engineering [4]. Ship-radiated noise comprises both broadband continuous spectrum and narrowband line spectrum components, with the line spectrum being one of the most significant features. These noises, originating from the mechanical noise, propeller noise,

and structural vibration noise of ships, radiate periodic noise during navigation, rendering the ship's line spectrum stable and unique [5]. The challenge lies in efficiently isolating the line spectrum from the complex background for UATR.

Traditional signal processing methods in this area demand specific expertise and lack universality. The advent of deep learning has sparked interest in employing convolutional neural networks (CNNs) and recurrent neural networks (RNNs), specifically Long Short-Term Memory (LSTM) units, for classifying and identifying UAT features [6]. Despite some successes, these methods encounter limitations in fully characterizing targets with a single feature and in capturing global feature representations. Furthermore, while LSTM units address sequence dependencies, they struggle with long-term dependencies essential for underwater acoustic signals, leading to parallel computation inefficiencies and training challenges.

The homogenization of features and network isolation results in suboptimal recognition accuracy and clustering effectiveness, highlighting the necessity for more comprehensive information fusion across data, features, and decision levels. For the sensitivity of underwater acoustic data and the challenges of large-scale data acquisition, data-level fusion often relies on simulated data generation or random masks for enhancement. Conversely, feature-level and decision-level fusions offer more advanced integration techniques, promising improvements in UATR [7].

Leveraging the remarkable achievements of Transformer models in handling large datasets, alongside the proven efficiency of convolutional neural networks (CNNs) for local feature refinement, this paper introduces a hybrid network that synergizes the strengths of both architectures for local–global feature fusion. This approach is experimentally validated, confirming the effectiveness and rationale of the proposed method. The contributions of this paper are summarized as follows:

- The joint network of the MobileNet and Transformer marks a pioneering step in UATR. This innovative architecture leverages the MobileNet for the extraction of locally refined features and addresses the detailed information overlooked by the Transformer. By integrating convolutional operations with the self-attention mechanism, the network effectively fuses local and global features, solves the problem of feature homogeneity and network isolation, and significantly enhances feature representation. This tandem structure of the hybrid network is designed to optimize the preservation of both local details and global context, facilitating comprehensive local and global correction learning. The parameter quantity of the joint model is greatly reduced compared with other Transformers.
- We design the coordinate attention mechanism for local feature extraction. The coordinate attention mechanism matching acoustic signal time–frequency characteristics is introduced to compensate for the inaccuracy and inadequacy of feature extraction. This mechanism compensates for the shortcomings by capturing the long-term temporal dependencies and precise frequency–domain relationships of signals, focusing the features on the time–frequency positions where the target signals reside. It is used to enrich spatial channel information and enhance local feature learning. For global information processing, local features are forwarded to the global sensory field for advanced learning and correction via an Embedding layer and six stacked Encoder Blocks.
- The hybrid model demonstrates exceptional performance, achieving a recognition accuracy of 98.50% on the Shipsear dataset, which surpasses existing methodologies. It also exhibits a higher silhouette coefficient, characterized by larger interclass distances and smaller intraclass distances, culminating in superior clustering effects. Applied to the DeepShip dataset, known for its lower signal-to-noise ratio, the model still maintains robust recognition capabilities with an accuracy of 94.60%, showcasing its effectiveness.

This paper is organized as follows. Section 2 focuses on the related work in this field. In Section 3, the method is proposed in detail. In Section 4, the relevant experimental results are analyzed and discussed. Section 5 draws conclusions.

2. Related Work

2.1. Methods Based on Traditional Machine Learning

UATR technology originally depended on the manual analysis of acoustic signal characteristics from various perspectives. Skilled sonar operators could identify acoustic signals by analyzing beats, timbres, and other features in conjunction with various spectrograms. However, this approach is prone to environmental influences and subjective interpretations, leading to inconsistent accuracy and notable constraints. With advancements in science and technology, the underwater acoustic field has witnessed the introduction of machine learning methods based on statistical classification, as illustrated in Figure 1. This development marks a significant shift toward more objective and stable recognition techniques.

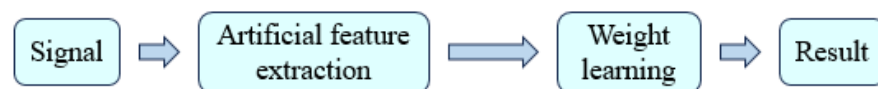


Figure 1. Machine learning method.

Since the 1990s, researchers have been integrating signal analysis theory with machine learning techniques, utilizing handcrafted feature identifiers to extract attributes from underwater acoustic signals. These attributes include the Zero-crossing Rate (ZCR), Wavelet Transform (WT) [8], Hilbert–Huang Transform (HHT), Higher-order Spectral Estimation, and Mel-frequency Cepstral Coefficient (MFCC) [9]. Machine learning classifiers such as Bayesian, Decision Tree, and Support Vector Machines (SVMs) [10] are then employed to identify and classify underwater targets. Notably, Moura et al. trained an SVM model using a dataset of ship-radiated noise collected from a real marine environment, employing LO-FAR images as input and achieving an accuracy of 73.18% [11]. In another study, a Gaussian Mixed Model (GMM) classifier, trained using the standard Expectation–Maximization algorithm, attained a classification accuracy of 75.4% [12]. The BAT algorithm optimizes kernel parameters and achieves higher classification accuracy, using MFCC features as input [13]. Compared with other parameter optimization algorithms, such as genetic algorithms (GAs) and particle swarm optimization (PSO), the BAT algorithm has the advantage of conducting global and local searches simultaneously to avoid falling into the local optimum. The results show that the accuracy of the classifier using the BAT optimization algorithm is six percentage points higher than the PSO algorithm. Yang et al. [14] introduced a novel AdaBoost SVM model based on a weighted sample and feature selection method (WSFSelect-SVME) to improve the accuracy of UATR, reducing extra computational and storage costs. The proposed model solved the limitations of traditional ensemble SVM methods: (1) Training data often have poor quality results in errors between actual and theoretical results. (2) Ensemble recognition systems usually have higher complexity and computational costs. The experimental results on the UCI sonar dataset and real-world underwater acoustic target dataset show that the WSFSelect-SVME model obtains better recognition performance and robustness than the Adaboost SVM ensemble algorithm. Kim et al. used synthesized sonar signals as input to avoid the problem of data acquisition and applied a multi-aspect target classification scheme based on a hidden Markov model for classification [15]. Meng et al. introduced an SVM classification method based on waveform structure, reaching an accuracy of 81.20% [16]. While traditional machine learning approaches have demonstrated commendable recognition capabilities in less complex marine settings, their capacity to accurately fit the sample distribution and generalize across datasets that require intricate feature extraction remains limited.

2.2. Methods Based on Deep Learning

In recent years, rapid advancements in deep learning technology have significantly impacted computer vision and pattern recognition, heralding a new era marked by self-optimization and deep feature mining capabilities. These advancements have found applications across a diverse range of fields [17–19]. UATR essentially falls under the umbrella of pattern recognition, and techniques based on deep learning are particularly pertinent and promising within this domain. Figure 2 delineates the recognition process, showcasing how deep learning methodologies can be effectively applied to UATR.

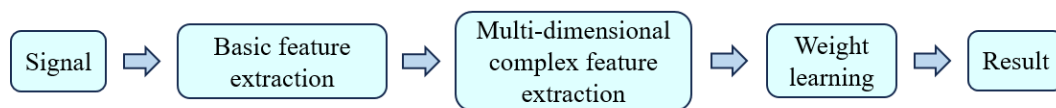


Figure 2. Deep learning method.

Recent advancements in deep learning have ushered in significant improvements in UATR, with various researchers employing innovative approaches to enhance recognition accuracy. Sabara et al. utilized spectrograms as inputs for aquatic target recognition and classification through convolutional neural networks (CNNs), achieving an accuracy of 80% [20]. Hu et al. distinguished the Shipsear dataset into three ship sizes (large, medium, and small) and utilized the original one-dimensional time-domain signals for input into a novel deep neural network model. This model, which combines depth separable convolution with time expansion convolution, attained a classification accuracy of 90.09% [21]. Zhao et al. introduced a multiscale residual unit (MSRU) to develop a deep convolutional stack network, demonstrating the MSRU algorithm’s effectiveness within a generative adversarial network framework and achieving an accuracy of 83.15% [22]. Li et al. proposed a method that leverages a deep neural network alongside an optimized loss function to reach 84.00% accuracy [23]. Ke et al. enhanced neural network performance through migration learning, achieving a recognition accuracy of 93.28% [24]. Luo et al. employed Restricted Boltzmann Machines (RBM) based on a stochastic neural network for recognition, achieving an accuracy of 93.17% [25]. A study extracted MFCC and LOFAR spectrogram features of underwater acoustic signals as network inputs, comparing CNN, LSTM, and SVM machine learning methods across different signal-to-noise ratios. The combination of LOFAR inputs with the CNN emerged as the most effective, reaching 95% accuracy. The three classifiers attained accuracies of 0.9914, 0.9892, and 0.9536, respectively, translating to a 22% recognition rate improvement. For ship-radiated noise simulation signals, both CNN and LSTM models were capable of nearly 80% recognition rates at a –10 dB signal-to-noise ratio [26].

While many of the methodologies previously outlined are somewhat basic and overlook comprehensive information integration, recent scholarly work has delved into UATR methods that leverage both feature-level and decision-level fusion. This approach has shown to significantly enhance recognition accuracy by synergizing different types of information. Han et al. adopted a feature-level one-dimensional fusion strategy, amalgamating feature vectors into a combined CNN and LSTM neural network, resulting in a classification accuracy of 92.14% [27]. Hong et al. implemented a feature-level three-dimensional fusion recognition method based on ResNet18, achieving a correct rate of 94.30% [28]. Feng et al. employed decision-level fusion by separately inputting three types of features, including MFCC, into the network, reaching a recognition accuracy of 98.34% [29]. These advancements underscore the efficacy of information fusion in improving recognition accuracy. However, while CNNs excel in local feature extraction, they fall short in global feature representation and the clear delineation of line spectra from background noise. Similarly, while LSTMs address certain sequence dependencies, they exhibit inefficiencies in processing time series due to a lack of long-term dependency and parallel computation capabilities, leading to inefficient training. Addressing these limitations, Li et al. pioneered the introduction of the Transformer model into UATR. Utilizing the Mel spectrum as input,

their STM model achieved an impressive accuracy of 97.70% [30]. This innovative approach, grounded in information fusion, demonstrates enhanced effectiveness in classification and recognition in the underwater acoustic domain, marking a significant advancement over traditional and singular feature-based methods.

In summary, firstly, CNNs excel at capturing local features, while Transformers are skilled in capturing global features. The combination could improve the local and global feature extraction capabilities of the model. Secondly, Transformers exhibit strong capabilities in modeling long-range dependencies in sequence data, and underwater acoustic signals exhibit richer features over long-range scales. By cascading Transformers after MobileNet, the model can effectively handle long-sequence data and capture long-range dependencies. Thirdly, the combination of MobileNet and Transformers is adaptable to various types of data, including images, text, and speech, and exhibits lighter weight than other Transformers. So, integrating the MobileNet and Transformer allows for leveraging their complementary strengths, facilitating effective processing of sequence data, and yielding improved model performance across diverse tasks.

3. Method

This section describes the Mobile_ViT hybrid network, which integrates the MobileNet convolutional network, enhanced by a coordinate attention mechanism, with the Transformer's self-attention architecture for global feature analysis. The architecture begins with a convolutional layer equipped with 32 kernels, succeeded by 14 residual Bottleneck layers. The Bottleneck layers' output connects to an Embedding layer, which then passes through a fully connected layer and 6 stacked Encoder Blocks for further processing. Figure 3 illustrates the complete Mobile_ViT structure, showcasing the seamless integration of MobileNet and Transformer as its core framework.

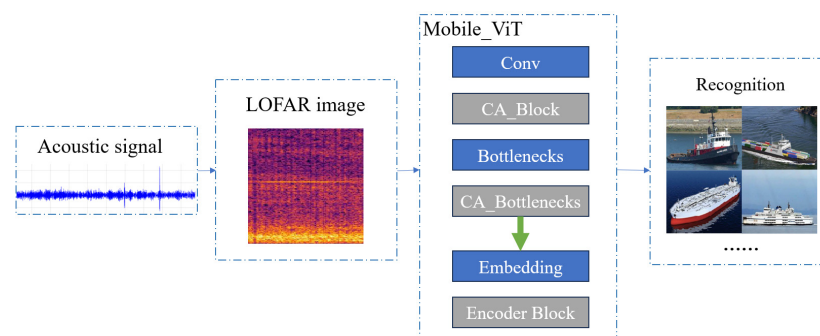


Figure 3. The process of underwater acoustic signal recognition.

3.1. The Structure of Mobile_ViT

MobileNet, a variant of convolutional neural networks, employs depth-separable convolutional units to drastically cut down the number of parameters, rendering it a lightweight network optimized for swift extraction of finely tuned local features. The hybrid network leverages the benefits of such lightweight convolutional architectures. Meanwhile, the Transformer, known for its fully attentional architecture, adeptly manages sequential issues involving dependencies. It utilizes self-attention mechanisms to identify long-range feature dependencies, facilitating global information exchange and enhancement within the hybrid framework. Figure 4 displays the architecture of the network.

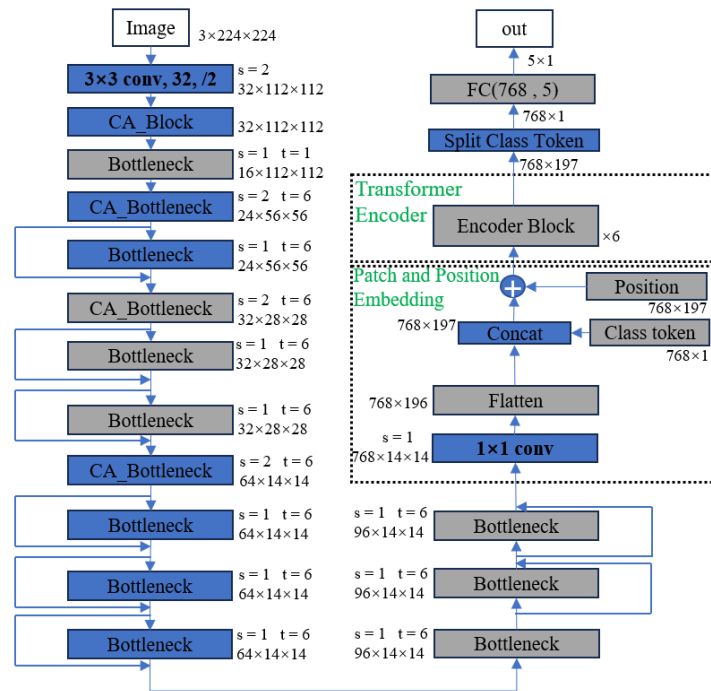


Figure 4. The architecture of Mobile_ViT.

3.2. Pretreatment

The Low-Frequency Analysis and Recording (LOFAR) spectrum is derived from the signal via the Short-Time Fourier Transform (STFT), focusing on the fact that high-frequency components undergo significant attenuation during underwater propagation [31]. Thus, LOFAR is selected as the preprocessing technique, with a frequency range set to 0–3000 Hz. Firstly, LOFAR breaks the signal into short overlapping segments and computes the Fourier transform of each segment. This results in a time–frequency representation of the signal. Secondly, LOFAR calculates the magnitude of the Fourier coefficients for each segment. This represents the intensity of different frequency components. Lastly, LOFAR normalizes the magnitude values to a suitable range (e.g., 0 to 255) to fit within the RGB color space. We create RGB images by combining the mapped colors for each segment. The LOFAR data are then converted into RGB values, and a selection of Y frames is used to generate the corresponding heatmap. The selection of Y depends on the frame length, frame shift, and segment time interval, aiming to mitigate the occurrence of the picket fence effect [32]. This visual representation effectively illustrates the frequency changes over time. By transforming 1D signals into 2D images, this approach allows for the application of computer vision strategies to the underwater target recognition task, offering a more intuitive analysis of frequency variations.

3.3. CA_MobileNet

CA_MobileNet serves as the foundation for local feature extraction, with the primary distinction from MobileNet being its selective retention of the model post-pruning and the integration of a coordinate attention mechanism during alterations in channel count and feature map dimensions. This mechanism reconstitutes the feature map to encode both channel and spatial information effectively, thereby amplifying the network’s representational power. The enhancement of the network’s representational capacity is achieved through the utilization of its intrinsic block structure, which notably includes CA_Block, Bottleneck, and CA_Bottleneck components.

3.3.1. CA_Block

The Coordinate Attention block is a computational module designed to augment the feature extraction capabilities of lightweight convolutional units. As a versatile plug-and-

play component, it accepts any intermediate feature tensor $X \in R^{C \times H \times W}$ (where C is the number of feature map channels, H is the height of the feature map, and W is the width of the feature map) as input and output X' with the same size as the input tensor. The output X' is enhanced by undergoing a transformation process that effectively encodes channel relationships and long-term dependencies using precise positional information, as depicted in Figure 5.

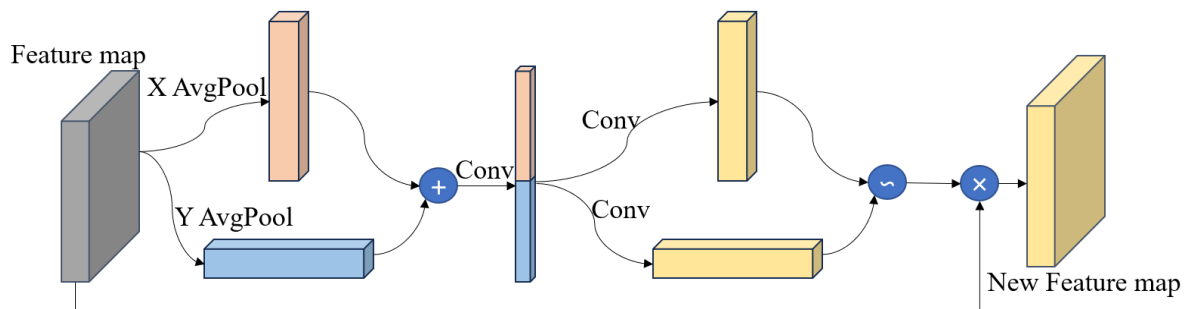


Figure 5. The architecture of CA_Block.

3.3.2. Bottleneck and CA_Bottleneck

The blocks in discussion are composed of two standard convolutions and one depth-separable convolution [33]. Initially, a 1×1 standard convolution is utilized to augment the dimensions, thereby mapping the feature extraction within a high-dimensional space. Subsequently, a 3×3 depth-separable convolution is applied for the purpose of feature extraction. To conclude, a 1×1 pointwise convolution is implemented to reduce the dimensions. It is within the CA_Bottleneck block that Coordinate Attention is integrated into the block's architecture. The configuration of these two blocks is depicted in Figure 6, while Table 1 displays their internal transformations from C_{input} to C_{output} with stride s and expansion factor t .

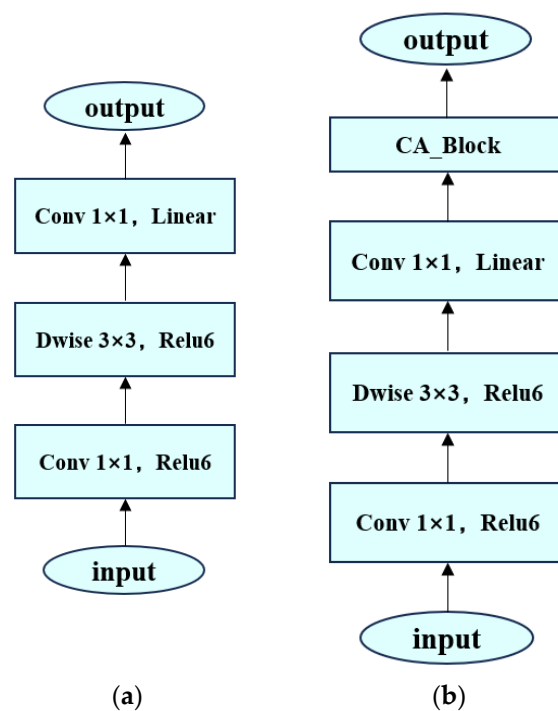


Figure 6. The architecture of two Bottlenecks. (a) The Bottleneck. (b) The CA_Bottleneck.

Table 1. Bottleneck residual block transforming.

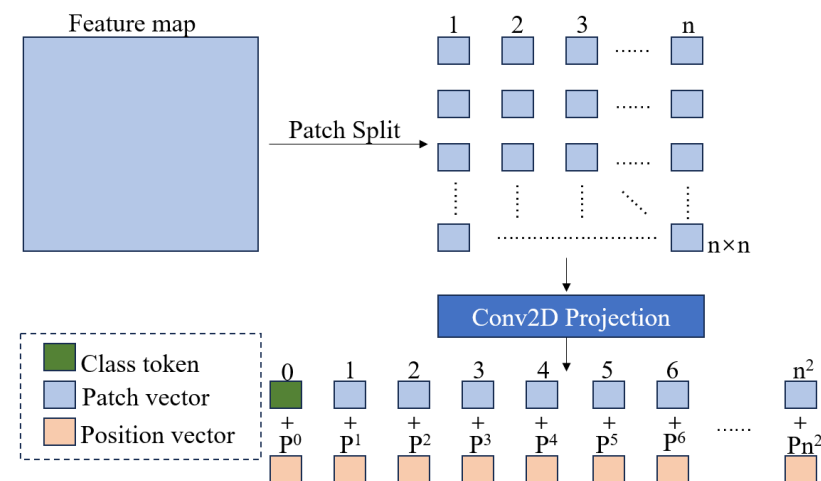
Input	Operator	Output
$H \times W \times C_{input}$	1×1 , conv2d, ReLU6	$H \times W \times C_{input}$
$H \times W \times (tC_{input})$	3×3 , Dwise $s = s$, ReLU6	$\frac{H}{s} \times \frac{W}{s} \times (tC_{input})$
$\frac{H}{s} \times \frac{W}{s} \times (tC_{input})$	Linear 1×1 , conv2d	$\frac{H}{s} \times \frac{W}{s} \times (tC_{output})$

3.4. Transformer

Transformer-cascaded self-attention blocks are employed to grasp feature dependencies across extended distances, addressing the shortcomings of convolutional neural networks (CNNs) in global feature acquisition. The meticulously refined local feature maps, produced via convolutional processes, are inputted into the self-attention mechanism. This step facilitates the fusion of local and global features, bolstering feature representation and effectuating the learning of local a priori with a global corrective approach. The architecture predominantly comprises Patch and Position Embedding, along with the Transformer Encoder Block, to achieve this comprehensive feature integration.

3.4.1. Patch and Position Embedding

We divide the convolved feature maps into uniformly sized patches before introducing them into the Transformer block. This step is crucial because the Transformer Encoder mandates a one-dimensional sequence of tokens for input, necessitating the flattening of each patch's height and width into a one-dimensional sequence. However, the Transformer inherently lacks the ability to discern the positional sequence of the input patches. To mitigate this, trainable position information is embedded within each token, granting the model the capacity to apprehend the features across the entire spectrum. Moreover, a learnable Class token, initialized randomly, is concatenated at the start of the patch sequence to serve in subsequent classification tasks. Figure 7 elaborates on the specifics of this procedure [30].

**Figure 7.** The process of Embedding.

3.4.2. Transformer Encoder

Upon processing through the Embedding layer, the complete sequence proceeds to the Encoder Block. Initially, it traverses a Layer Normalization (LN) layer before entering the Multi-Head Attention (MSA) mechanism for Multi-Head Self-Attention processing. A residual connection facilitates a summing operation, and following another pass through an LN layer, the sequence is directed into the Multi-Layer Perception (MLP) layer. A residual connection is then employed to derive the output from the Encoder Block. For optimal performance, this paper specifies the use of 6 stacked Encoder Blocks within

the Transformer architecture. The structural details of the Encoder Block are depicted in Figure 8.

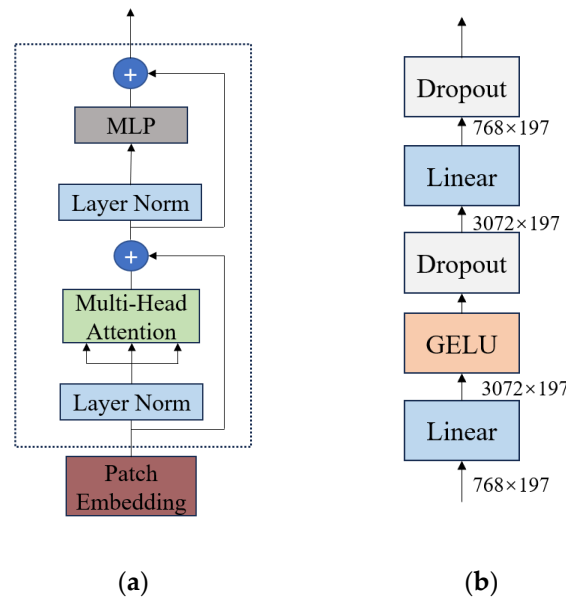


Figure 8. The architecture of Encoder Block and MLP. (a) The Encoder. (b) The MLP.

Explanations for LN, MSA, and MLP are provided below.

LN: Layer Normalization is a key part in the Transformer for stable training and faster convergence. LN is applied over each sample $X \in R^d$ as follows: $LN(X) = \frac{X - \mu}{\delta} \gamma + \beta$, where $\mu, \delta \in R$ are the mean and standard deviation of the features, respectively, and $\gamma, \beta \in R^d$ are the learnable affine transform parameters.

MSA: The MSA is an evolution of the self-attention mechanism. The model is divided into multiple heads to form multiple subspaces, enabling it to attend to different aspects of information. Self-attention transfers the inputs $X \in R^{n \times d}$ into three parts, then calculates the similarity between elements to facilitate the transformation of features. The formula is

$$Q, K, V = XW^q, XW^k, XW^v \quad (1)$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

where $W^q, W^k, W^v \in R^{d \times d_1}$ are the learnable projection matrices. $Q, K, V \in R^{n \times d_1}$ are the query, key, and value matrices, respectively. d is the embedding dimension. Multi-head self-attention splits the matrices into h parts and performs the attention function in parallel. The output values of each head are concatenated and projected linearly to form the final output.

MLP: The Multi-Layer Perception (MLP) architecture consists of two fully connected layers. Initially, patches are processed through the first fully connected layer, which expands the number of neuron nodes by fourfold. This expansion is succeeded by the application of a Gaussian Error Linear Unit (GELU) activation function. Subsequently, the output is directed through a Dropout layer, leading into the second fully connected layer, where the number of nodes is decreased back to the original count. A final Dropout layer is then applied to the output. The process can be summarized by the following formula:

$$MLP(X) = FC(\sigma(FC(X))) \quad (3)$$

$$FC(X) = WX + b \quad (4)$$

where W and b are the weight and bias terms of fully connected layer, respectively, and $\sigma(\cdot)$ is the activation function. FC is the fully connected layer.

4. Experiments

This section presents experimental evaluations conducted on two internationally recognized public datasets, Shippsear and DeepShip, to validate the efficacy of hybrid networks in UATR through the synergistic interaction of local and global feature information. The superiority of this method is demonstrated through comparative analysis with other pertinent approaches.

4.1. Datasets

The Shippsear dataset comprises a variety of ship-radiated noise signals, recorded off the coast of Spain, featuring a sampling frequency of 52,734 Hz [12]. It includes around three hours of audio data, equivalent to 90 audio recordings, with a hydrophone collection radius of 150 m. The dataset categorizes ships into four classes based on size and sailing speed: Class A includes small and medium-sized vessels; Class B encompasses small vessels; Class C consists of large passenger ships; and Class D comprises giant ocean-going vessels. Additionally, there is Class E dedicated to ambient ocean noise collected.

The DeepShip dataset contains radiated noise from 265 vessels, recorded under real sea conditions in the Strait of Georgia delta node [34]. This dataset, collected with a hydrophone radius of 2 km and a sampling frequency of 32 kHz, organizes vessels into four commercial categories: tankers, tugboats, passenger ships, and cargo ships. It features recordings across various sea conditions and noise levels, offering a comprehensive snapshot of the real-world marine environment. The dataset includes not only vessel signals but also natural background noise, marine mammal sounds, and noises from human activities.

In the Shippsear dataset, each audio recording is divided into 2 s segments, producing a total of 5269 data samples. Conversely, in the DeepShip dataset, each recording is segmented into approximately 6 s segments, resulting in 9646 data samples. For the experimental setup, 70% of the data are designated for training purposes, while the remaining 30% are reserved for validation, as outlined in Tables 2 and 3.

Table 2. Shippsear dataset.

Class	Total Size	Train	Val
A: Fish boats, Trawlers, Mussel boat, Tugboat, Dredger	5269	612	262
B: Motorboat, Pilot boat, Sai boat		507	216
C: Passengers		1395	597
D: Ocean liner, RORO		803	344
E: Background noise		374	159

Table 3. DeepShip dataset.

Class	Total Size	Train	Val
A: Tug	2341	1638	703
B: Tanker	2482	1737	745
C: Passengers	2624	1836	788
D: Cargo	2199	1539	660

4.2. Pre-Training Process

During the pre-training phase, the batch size is configured to 16, and the training extends over 100 epochs. Stochastic Gradient Descent (SGD) serves as the optimization algorithm, supplemented by a cosine annealing strategy for adjusting the learning rate.

The initial learning rate is established at 0.001. Model performance throughout the training process is evaluated using the cross-entropy loss function.

4.3. Evaluation Metrics

In this paper, we employ several metrics to evaluate the model's performance, including recognition accuracy (*Acc*), *Kappa* coefficient, *Recall*, *F1* Score, and the silhouette coefficient (*SC*).

If the predicted value is the same as the true value, the predicted value is a positive sample, denoted as *TP*; if the predicted value is a negative sample, it is denoted as *TN*; if not the same, the predicted value is a positive sample, denoted as *FP*, and if the predicted value is a negative sample, it is denoted as *FN*. The *Recall*, *Precision*, and *F1* Score are calculated as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad (5)$$

$$Recall = \frac{TP}{TP + FN}, \quad (6)$$

$$Precision = \frac{TP}{TP + FP}, \quad (7)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \quad (8)$$

The model bias is evaluated using the *Kappa* coefficient. It is calculated as follows:

$$P_0 = Acc, \quad (9)$$

$$P_e = \frac{a_1 \times b_1 + a_2 \times b_2 + \dots + a_c \times b_c}{n \times n}, \quad (10)$$

$$Kappa = \frac{P_0 - P_e}{1 - P_e}, \quad (11)$$

where $a_1, a_2, \dots, a_c$ indicate the number of actual samples for each category, and $b_1, b_2, \dots, b_c$ indicate the number of predicted samples for each category.

The *SC* evaluates the clustering effect of the model. It indicates the clarity of the contour of each category after clustering. The calculation is as follows:

$$S = \frac{1}{N} \sum_{i=1}^N s(i), \quad (12)$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (13)$$

where N represents the total number of sample points, $a(i)$ represents the average distance from other sample points in the same cluster as i , and $b(i)$ represents the minimum value of the average distance from sample points in different clusters that i belongs to.

4.4. Experimental Results and Analysis

The signals in the Shipsear dataset were collected within a 150 m range and exhibited a higher signal-to-noise ratio compared to those in the DeepShip dataset. Initially, the experiments were conducted using the Shipsear dataset. Table 4 provides a comparison between the method proposed in this study and other existing methods within the field. Additionally, Table 5 showcases a comparison of the reproduction results using some classical network architectures. The outcomes of these comparisons are detailed below:

Table 4. Shipsear comparison result.

Number	Method	Acc (%)
1	GMM + EM [12]	75.40
2	Waveform + SVM [16]	81.20
3	Graph + CNN [20]	80.00
4	Time-Dilated [21]	90.09
5	MSRU [22]	83.15
6	DNN + Loss [23]	84.00
7	ResNet18_3D [28]	94.30
8	Decision fusion [29]	98.34
9	Mel + STM [30]	97.70
10	Mobile_ViT	98.50

Table 5. Shipsear experiment result.

Model	Acc (%)	Kappa	Recall (%)	F1 Score	S	Size (M)
Resnet18 [28]	97.90	0.9722	97.64	97.80	0.6224	11.18
MobileNet	97.33	0.9646	97.26	97.16	0.5528	2.2
STM [30]	97.70	0.9670	97.02	96.92	0.7082	81.83
ViT	93.60	0.9151	93.52	93.40	0.4017	81.83
Mobile_ViT	98.50	0.9798	98.40	98.38	0.8817	21.61

Table 4 highlights the superiority of deep-learning-based methods over traditional machine learning techniques for underwater acoustic target recognition, demonstrating higher accuracy. As network models evolve, novel approaches utilizing feature-level or decision-level information fusion have further enhanced UATR capabilities. This study introduces a method based on local–global feature fusion, achieving an impressive classification accuracy of 98.50% on the Shipsear dataset, outperforming other methodologies. Table 5 showcases experimental results underscoring the efficacy of the proposed method. Utilizing Resnet18 and MobileNet convolutional neural networks for the local feature learning of UAT, this method achieves recognition accuracies of 97.90% and 97.33%, respectively. Moreover, employing cascaded attention blocks of the Transformer, recognition accuracies of 97.70% on STM and 93.60% on ViT were attained. The method introduced in this paper secures the highest recognition accuracy by initially extracting local refinement features through convolution operations for detailed information learning, followed by global information interaction via cascading attention blocks, epitomizing the novel local–global feature fusion approach. The parameter quantity of Mobile_ViT is greatly reduced compared with other Transformers.

Table 6 demonstrates the effectiveness of the modular methods proposed in this study, showing improved accuracies over the baseline MobileNet model, which is advantageous for real-world UATR applications. Incorporating the CA_Block results in a recognition accuracy improvement of 0.83%, with marginal enhancements observed in other performance metrics. The integration of both the CA_Block and Encoder Block elevates the recognition accuracy by 1.17%. Notably, the approach introduced in this paper achieves a higher silhouette coefficient, indicating a more distinct clustering effect. To vividly illustrate this impact, the experimental outcomes are visualized in Figure 9.

Table 6. Ablation experiment.

Model	CA_Block	Encoder_Block	Acc (%)	Kappa	Recall (%)	F1 Score	S
MobileNet	×	×	97.33	0.9646	97.26	97.16	0.5528
CA_Mobile	✓	×	98.16	0.9756	98.02	98.16	0.7881
Mobile_ViT	✓	✓	98.50	0.9798	98.40	98.38	0.8817

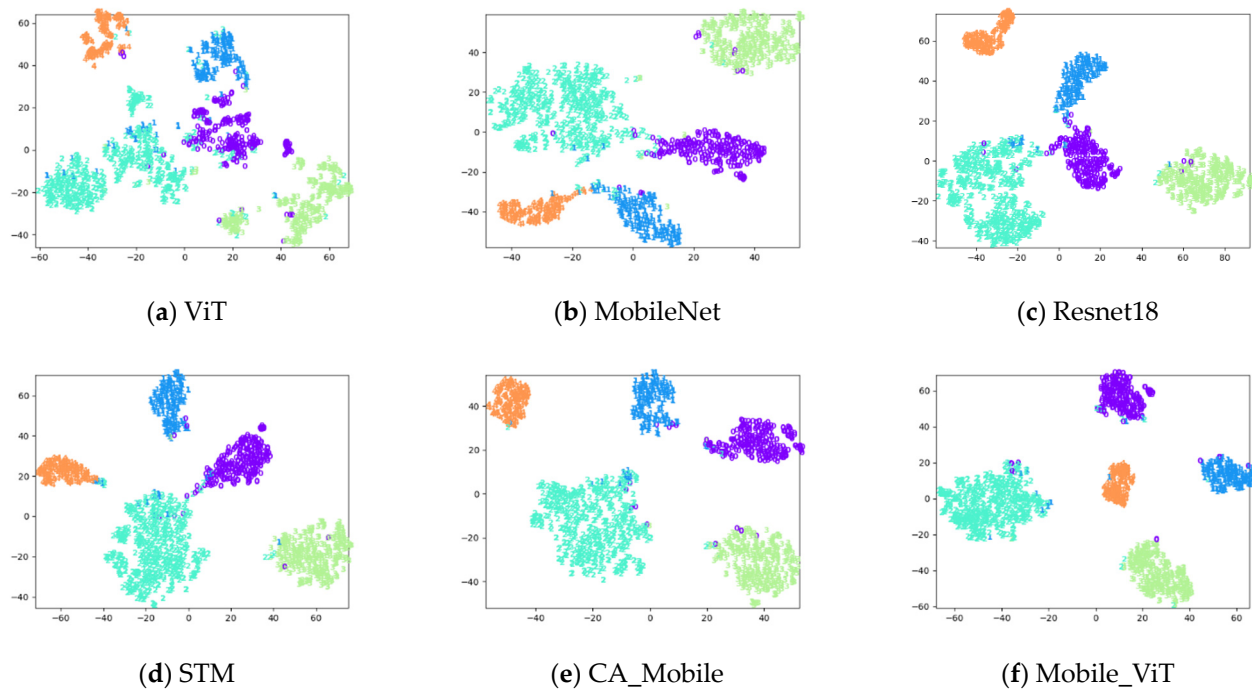


Figure 9. T-SNE visualization from the above experiments (different colors represent different classes).

Table 7 outlines the comparative performance of various methods on the DeepShip dataset, showcasing the adaptability and robustness of our approach in handling datasets with varied acoustic characteristics.

Table 7. DeepShip comparison result.

Number	Method	Acc
1	Wavelets + Inception	59.85
2	CQT+SCAE [34]	77.53
3	UALF [35]	81.39
4	Mel+ResNet18	91.12
5	Mobile_ViT	94.57

Table 7 demonstrates the superior performance of our proposed method in recognizing signals with low signal-to-noise ratios in real-world scenarios, significantly outperforming the approaches mentioned in [34,35]. The efficacy of the model's series structure is also confirmed, effectively preserving local detail information for subsequent global information interaction. Table 8 shows the enhanced recognition accuracies upon incorporating only the CA_Block and both the CA_Block and Encoder Block into the original MobileNet network, with improvements of 3.32% and 4.39%, respectively. Visualization in Figure 10 compares the classification effects among three methods: (a) showcases indistinguishable target categories with no clear boundaries; (b) exhibits slight improvement yet still lacks distinct boundaries; and (c) clearly distinguishes the four target categories with well-defined contours and spacing between different classes, indicating the introduced module's enhanced sensitivity to UAT signals in low-SNR and complex scenarios. The Mobile_ViT network, with its embedded coordinate attention mechanism and local–global information interaction, proves more adept at capturing low-SNR targets amidst the complex background noise of marine environments.

Table 8. DeepShip experiment result.

Model	Acc (%)	Kappa	Recall (%)	F1 Score	S
MobileNet	90.18	0.8689	90.10	90.12	0.3528
CA_Mobile	93.50	0.9132	93.55	93.49	0.5526
Mobile_ViT	94.57	0.9275	94.58	94.56	0.6868

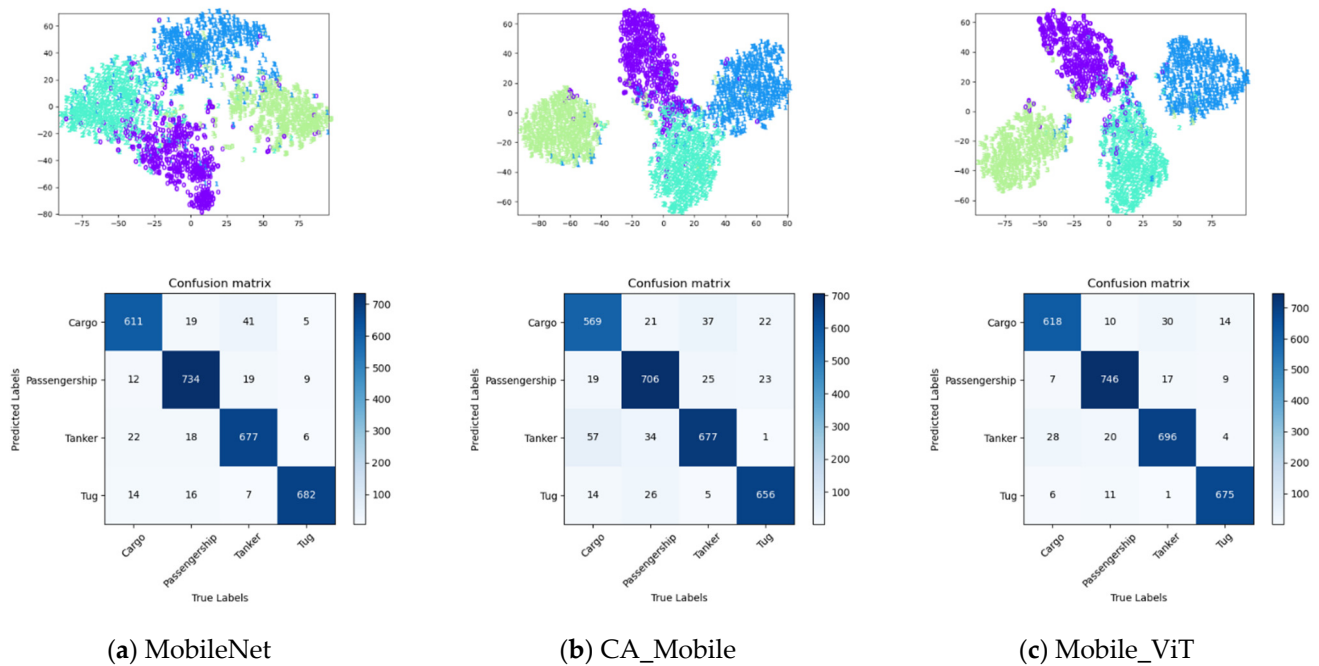


Figure 10. T-SNE visualization and confusion matrices from the above experiments (different colors represent different classes).

5. Conclusions

This study introduces Mobile_ViT, a hybrid network combining MobileNet and Transformer architectures, optimized for UATR in real scenarios. By incorporating a coordinate attention mechanism and local–global feature fusion, this network capitalizes on the benefits of integrating local detail enhancement with global information correction, demonstrating superior performance in marine environments. Particularly noteworthy is its capacity to discern targets with low signal-to-noise ratios amidst background noise, showcasing its effectiveness in detecting subtle underwater targets. The proposed method stands out in extracting features and classifying targets under conditions of vast distances and minimal signal clarity in deep-sea environments. Future work will focus on enhancing the model’s interpretability, continuing to study lightweight models, delving into the model’s learning mechanisms, and refining its decision-making processes.

Author Contributions: Conceptualization, H.Y. and T.G.; Formal analysis, T.G., H.Y. and H.W.; Funding acquisition, H.W.; Investigation, Y.W. and X.C.; Methodology, H.Y. and T.G.; Resources, Y.W. and H.Y.; Software, H.Y. and T.G.; Validation, H.Y. and T.G.; Writing—original draft, H.Y. and T.G.; Writing—review and editing, H.W. and X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Project of the National Natural Science Foundation of China, grant number 62031021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data and results supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

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

References

1. Kaiser, M.J. *Marine Ecology: Processes, Systems, and Impacts*; Oxford University Press: Oxford, UK, 2011.
2. Ali, M.F.; Jayakody, D.N.K.; Chursin, Y.A.; Affes, S. Recent advances and future directions on underwater wireless communications. *Arch. Comput. Methods Eng.* **2020**, *27*, 1379–1412. [\[CrossRef\]](#)
3. Urlick, R.J. Principles of underwater sound. *McGraw-Hill Google Sch.* **1983**, *2*, 2760–2766.
4. Vaccaro, R.J. The past, present, and the future of underwater acoustic signal processing. *IEEE Signal Process. Mag.* **1998**, *15*, 21–51. [\[CrossRef\]](#)
5. Arrabito, G.R.; Cooke, B.E.; McFadden, S.M. Recommendations for enhancing the role of the auditory modality for processing sonar data. *Appl. Acoust.* **2005**, *66*, 986–1005. [\[CrossRef\]](#)
6. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Mangai, U.G.; Samanta, S.; Das, S.; Chowdhury, P.R. A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Tech. Rev.* **2010**, *27*, 293–307. [\[CrossRef\]](#)
8. Li, X.; Zhu, F. Application of the zero-crossing rate, LOFAR spectrum and wavelet to the feature extraction of passive sonar signals. In Proceedings of the 3rd World Congress on Intelligent Control and Automation, Hefei, China, 26 June–2 July 2000; Volume 4, pp. 2461–2463.
9. Lim, T.; Bae, K.; Hwang, C.; Lee, H. Classification of underwater transient signals using MFCC feature vector. In Proceedings of the 2007 9th International Symposium on Signal Processing and Its Applications, Sharjah, United Arab Emirates, 12–15 February 2007; pp. 1–4.
10. Liu, J.; He, Y.; Liu, Z.; Xiong, Y. Underwater target recognition based on line spectrum and support vector machine. In Proceedings of the 2014 International Conference on Mechatronics, Control and Electronic Engineering (MCE-14), Shenyang, China, 27–29 August 2014; pp. 79–84.
11. 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.
12. Santos-Domínguez, D.; Torres-Guijarro, S.; Cardenal-López, A.; Pena-Gimenez, A. ShipsEar: An underwater vessel noise database. *Appl. Acoust.* **2016**, *113*, 64–69. [\[CrossRef\]](#)
13. 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.
14. 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.
15. Kim, T.; Bae, K. HMM-based underwater target classification with synthesized active sonar signals. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2011**, *94*, 2039–2042. [\[CrossRef\]](#)
16. Meng, Q.; Yang, S. A wave structure based method for recognition of marine acoustic target signals. *J. Acoust. Soc. Am.* **2015**, *137*, 2242. [\[CrossRef\]](#)
17. 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 IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
18. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
19. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
20. Sabara, R.; Jesus, S. Underwater acoustic target recognition using graph convolutional neural networks. *J. Acoust. Soc. Am.* **2018**, *144*, 1744. [\[CrossRef\]](#)
21. Hu, G.; Wang, K.; Liu, L. Underwater acoustic target recognition based on depthwise separable convolution neural networks. *Sensors* **2021**, *21*, 1429. [\[CrossRef\]](#)
22. Tian, S.; Chen, D.; Wang, H.; Liu, J. Deep convolution stack for waveform in underwater acoustic target recognition. *Sci. Rep.* **2021**, *11*, 9614. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Li, C.; Liu, Z.; Ren, J.; Wang, W.; Xu, J. A feature optimization approach based on inter-class and intra-class distance for ship type classification. *Sensors* **2020**, *20*, 5429. [\[CrossRef\]](#)
24. Ke, X.; Yuan, F.; Cheng, E. Underwater acoustic target recognition based on supervised feature-separation algorithm. *Sensors* **2018**, *18*, 4318. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Luo, X.; Feng, Y. An underwater acoustic target recognition method based on restricted Boltzmann machine. *Sensors* **2020**, *20*, 5399. [\[CrossRef\]](#) [\[PubMed\]](#)

26. Jian, Z. Research on underwater target recognition based on deep learning. Ph.D. Thesis, University of Electronic Science and Technology of China, Chengdu, China, 2020.
27. Han, X.C.; Ren, C.; Wang, L.; Bai, Y. Underwater acoustic target recognition method based on a joint neural network. *PLoS ONE* **2022**, *17*, e0266425. [[CrossRef](#)] [[PubMed](#)]
28. 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](#)]
29. Feng, H.; Chen, X.; Wang, R.; Wang, H.; Yao, H.; Wu, F. Underwater acoustic target recognition method based on WA-DS decision fusion. *Appl. Acoust.* **2024**, *217*, 109851. [[CrossRef](#)]
30. 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](#)]
31. 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](#)]
32. Li, Y.F.; Chen, K.F. Eliminating the picket fence effect of the fast Fourier transform. *Comput. Phys. Commun.* **2008**, *178*, 486–491. [[CrossRef](#)]
33. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
34. 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](#)]
35. Ren, J.; Xie, Y.; Zhang, X.; Xu, J. UALF: A learnable front-end for intelligent underwater acoustic classification system. *Ocean. Eng.* **2022**, *264*, 112394. [[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.