

Article

Ship Segmentation via Combined Attention Mechanism and Efficient Channel Attention High-Resolution Representation Network

Xiaoyi Li

Merchant Marine College, Shanghai Maritime University, Shanghai 201306, China; lixiaoyi@stu.shmtu.edu.cn

Abstract: Ship segmentation with small imaging size, which challenges ship detection and visual navigation model performance due to imaging noise interference, has attracted significant attention in the field. To address the issues, this study proposed a novel combined attention mechanism and efficient channel attention high-resolution representation network (CA2HRNET). More specially, the proposed model fulfills accurate ship segmentation by introducing a channel attention mechanism, a multi-scale spatial attention mechanism, and a weight self-adjusted attention mechanism. Overall, the proposed CA2HRNET model enhances attention mechanism performance by focusing on the trivial yet important features and pixels of a ship against background-interference pixels. The proposed ship segmentation model can accurately focus on ship features by implementing both channel and spatial fusion attention mechanisms at each scale feature layer. Moreover, the channel attention mechanism helps the proposed framework allocate higher weights to ship-feature-related pixels. The experimental results show that the proposed CA2HRNET model outperforms its counterparts in terms of accuracy (Acc_s), precision (P_c), F1-score ($F1_s$), intersection over union (IoU), and frequency-weighted IoU (F_{IoU}). The average Acc_s , P_c , $F1_s$, IoU, and F_{IoU} for the proposed CA2HRNET model were 99.77%, 97.55%, 97%, 96.97%, and 99.55%, respectively. The research findings can promote intelligent ship visual navigation and maritime traffic management in the smart shipping era.



Citation: Li, X. Ship Segmentation via Combined Attention Mechanism and Efficient Channel Attention High-Resolution Representation Network. *J. Mar. Sci. Eng.* **2024**, *12*, 1411. <https://doi.org/10.3390/jmse12081411>

Academic Editor: Weicheng Cui

Received: 14 June 2024

Revised: 8 August 2024

Accepted: 14 August 2024

Published: 16 August 2024



Copyright: © 2024 by the author. 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/>).

Keywords: ship segmentation; attention mechanism; smart shipping; visual navigation

1. Introduction

It is a big challenge to detect ships from visual sensory data sources when the ship imaging size is small due to a small signal-to-noise ratio and low signal intensity. The background in maritime images imposes additional ship detection and segmentation challenges. However, it is quite important to detect ships far away from your own ship in order for collision avoidance measures to be implemented. More specifically, the ship's course can be adjusted to avoid a potential maritime traffic collision by changing the ship's heading direction, lowering/increasing the ship's speed, etc. Ship identification far from the distance to closest point of approach (DCPA) can help a ship's crew take action to prevent ship collisions. From the perspective of ship visual navigation, ship identification can be roughly divided into ship detection from image sequences and ship recognition in terms of pixel-wise ship segmentation. Ship detection plays a crucial role in ship navigation and maritime-traffic surveillance-related tasks. More specifically, the on-board ship crew can take action to control a ship's speed and direction with the help of ship detection results (i.e., a potential ship collision can be identified via ship detection results and measures can be taken to avoid a maritime accident). In addition, maritime traffic regulators can determine illegal ship activities from ship detection results. Currently, the ship detection task is primarily implemented by finding the maximum similarity between input ship images and pre-trained models [1]. Ship detection models show satisfactory performance by labeling to-be-detected ships with bounding boxes in maritime images. Deep-learning-based models (such as the faster region-convolution neural network (R-CNN), you only

look once, etc. [2,3]) show better performance compared to traditional ship detection models (contour-based ship detection logic [4]). Ren et al. proposed a novel saliency-guided feature ship detection network to tackle distant ship detection under complex background interference [5].

A ship with a small imaging size can also be successfully detected by the saliency-guided detection model. Lee et al. utilized a deep learning model to detect ships from Korea Multipurpose Satellite (KOMPSAT) image sequences, which aimed to efficiently detect various ships sailing on the ocean and inland waterways [6]. Yasir et al. systematically analyzed a ship detection model constructed using synthetic aperture radar (SAR) data, and they conducted a comprehensive literature review by focusing on the merits and demerits of existing ship detection models to identify future directions in ship detection [7]. Jiang et al. developed a lightweight real-time ship detection model with the help of the YOLOv7-tiny model as the baseline model, and a partial convolution mechanism was employed to reduce parameters in the proposed framework [8]. The experimental results suggested that ship detection accuracy was larger than 97%, whilst the frame per second rate was 119.84 (i.e., the model can be implemented in real time). Wang et al. proposed an efficient convolutional-neural-network-based ship detection model by introducing an attention mechanism to enhance small ship detection performance, and a dilated convolution layer was further introduced to a traditional CNN to increase the CNN's receptive perception field. The proposed framework promoted ship detection in remote imagery data. However, ship detection in a complex maritime environment (e.g., large traffic volume, ships moving at high speed, ship occlusion) challenges ship detection model performance [9]. Similar studies can also be found elsewhere [10–12].

Much attention is also paid to recognizing ships from maritime images via traditional computer vision models and deep-learning-supported models. It is found that traditional ship-detection-related models may fail to design a distinct feature extraction mechanism, and thus, ship recognition and detection accuracy may sharply vary due to input image quality. CNN-based deep learning models have demonstrated powerful performance in ship-recognition-related tasks compared to traditional ship segmentation models. Ship semantic segmentation models can achieve pixel-wise ship identification and recognition, whilst background interference can be minimized as background and ship pixels can be successfully and accurately identified by the semantic models. In addition, semantic-segmentation-based models can clearly retain object boundary pixels in complex maritime traffic situations (i.e., when ships and background-related objects occupy the majority of maritime images).

A large number of studies have been conducted to accomplish ship segmentation via a multiscale attention ship instance segmentation mechanism, transformer, lightweight multi-layer perceptron, etc. [11,13,14]. George et al. proposed a novel framework to determine optimal ship speed under a fixed course by segmenting the overall path into small corridors, and the interdependency for each sub-trajectory optimization is also considered [15]. David et al. developed context information exploitation with the help of an automatic identification system (AIS) and a ship's trajectory is further identified and segmented by segmentation-related models [16]. Sun et al. suggested that instance segmentation for identifying small objects from maritime images plays quite an important role in the maritime visual navigation task, and small imaging size, low image coverage rate and blurry ship visual features challenge model performance. Sun et al. proposed a dual-branch activation network for small imaging size ship detection and segmentation [17]. Overall, previous studies mainly focus on ship detection and tracking by identifying ships with a bounding box (i.e., a bounding box indicating a ship). However, background-related pixels may be wrongly considered as ship pixels in previous studies. In addition, the semantic models try to identify ship-contour-related pixels, and trivial ship features and pixels may be discarded. To address the issue, this study proposes a novel ship segmentation framework via a combined attention mechanism and efficient channel attention high-resolution representation network (CA2HRNET). The remaining sections are organized as follows.

Section 2 describes the method used in the study, Section 3 describes the experimental results, and finally, conclusions are drawn in Section 4.

2. Methodology

2.1. Basic Ship Segmentation Structure

It should be noted that traditional convolutional neural network models implement ship feature extraction in a top-down manner. A traditional CNN employs down-sampling and max-pooling to compress the width and height of an input image. The feature extraction procedure can obtain different shapes with the help of feature layers. A traditional CNN implements top-down feature extraction, while a high-resolution representation network (HRNet) employs both high-resolution and low-resolution convolution operations (in a parallel manner) to extract object trivial and distinct features from images. It is found that both down-sampling and up-sampling mechanisms help the deep learning framework extract and merge ship shape features for the purpose of exploiting distinctive features. The HRNet model was first employed to detect human posture, via semantic segmentation, from video clips and showed quite powerful performance [18]. An HRNet-based model can efficiently learn global context and spatial information from input images, which mitigates the disadvantages of recovering low-resolution feature maps. The object shape and boundary features may be easily discarded during feature transfer from a lower encoder layer to a higher encoder layer. HRNet-network-generated prediction maps can obtain higher accuracy for achieving the object recognition task.

Overall, the basic HRNet network consists of a backbone layer, a feature integration layer, and a prediction head layer. The backbone layer uses four stages to perform multi-resolution feature fusion tasks via feature exchange in the layer. A schematic overview of the HRNet structure is shown in Figure 1. As can be seen, the HRNet network model consists of two key characteristics. First, the high-resolution and low-resolution layers are connected in a parallel manner. Second, feature map information is constantly exchanged between high- and low-resolution maps. The high-resolution feature map helps the model obtain accurate ship spatial information, while the low-resolution map ensures that boundary-related pixels can be clearly identified.

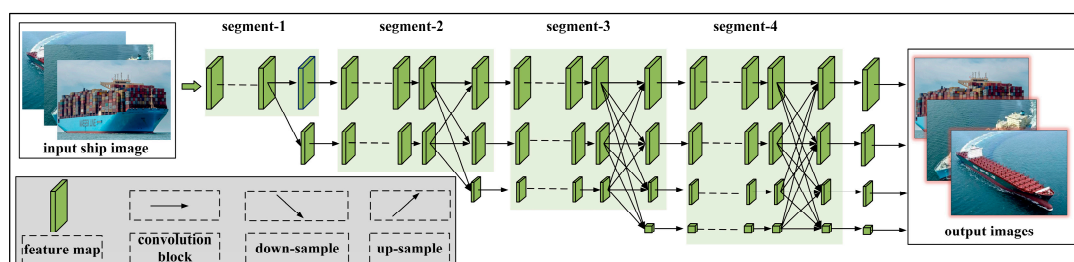


Figure 1. Overview of the HRNet Network.

2.2. CA2HRNET Framework

Traditional semantic segmentation models may lead to spatial and global contextual information loss due to resolution degradation during the feature extraction procedure. While HRNet-based models can alleviate this disadvantage with the help of high-resolution layer-based feature extraction, the information loss problem cannot be completely addressed. Our study proposes a novel CA2HRNET framework by tackling the information loss challenge with the help of an attention module. The attention module is introduced and embedded at the end of the high-resolution subnetwork convolutional layer. The proposed framework implements consistently high-resolution feature extraction by following the rule used in the HRNet network, and the feature resolution is simultaneously analyzed, which differs from traditional semantic segmentation models. The proposed network starts with a high-resolution sub-network as the first stage, and high-resolution layers and corresponding feature maps are added into low-resolution layers and networks to

further implement the feature extraction procedure. Moreover, multi-scale feature fusion is repetitively implemented to enrich feature maps.

2.2.1. Channel Attention Mechanism

Deep-learning-based models employ an attention mechanism to focus on a region of interest, and this attention mechanism can be adjusted to fulfill different visual tasks by changing channel features or spatial structures. The convolutional block attention mechanism (CBAM) consists of a channel attention mechanism and a spatial attention mechanism, which outperforms the conventional attention mechanism. Usually, channel attention (CA) is used to learn the importance of each channel, while spatial attention is used to learn the importance of each spatial location for an input video clip. This study enhances CBAM attention mechanism performance by introducing a correlation attention module (CAM) attention mechanism, which is abbreviated as C2BA2M2 in our study. The CAM mechanism helps the ship segmentation framework better identify ship trivial features while ship features are extracted by varied channel layers. More specifically, the proposed C2BA2M2 integrates channel attention, multi-scale spatial attention, and adaptive weight fusion, combines multiple modules, and enhances the ability of attention modeling from different perspectives. An overview of the proposed C2BA2M2 module is shown in Figure 2. The C2BA2M2 module better focuses on the important regions in the input image due to its ability to capture spatial information at different scales by introducing multiple attention units with different convolutional kernel sizes in the spatial attention module. In other words, the model receptive field can be significantly improved and target imaging scale variation does not impose an obviously negative impact on the C2BA2M2 module. Moreover, a weight fusion module is introduced to help the model adaptively determine the importance of channel attention and spatial attention based on the input features and dynamically adjust the weight setups. This mechanism makes attention allocation more flexible and efficient.

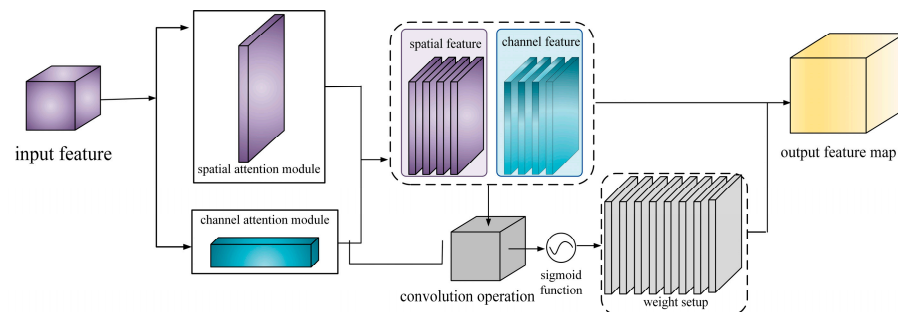


Figure 2. Schematic overview of the proposed C2BA2M2 module.

The spatial attention module introduces multiple attention units with different convolutional kernel sizes based on the original CBAM network. It then performs a weighted fusion of features obtained from different convolutions to obtain the spatial information of the multi-scale feature map. A schematic diagram of the spatial attention module is shown in Figure 3. Note that the multi-scale convolutional feature extraction is computed with Equations (1) and (2). The multi-scale fusion operation can output the ship features based on Equation (3). The fused features are subjected to global max-pooling and average pooling operations from the perspective of channel dimensions, and the convolution layer is implemented to reduce model dimensionality. Moreover, Sigmoid operations are used to generate ship spatial attention features.

$$F_i = \text{Conv}(K_i \times K_i) X_i, i = 0, 1, 2, \dots, S-1 \quad (1)$$

$$K_i = 2 \times (i + 1) + 1 \quad (2)$$

$$F = \text{Con}(F_0, F_1, \dots, F_{S-1}) \quad (3)$$

where F_i is the feature extracted by convolution at different scales, K_i is the convolution kernel size and X_i is the input feature. The Con symbol represents the concatenate function.

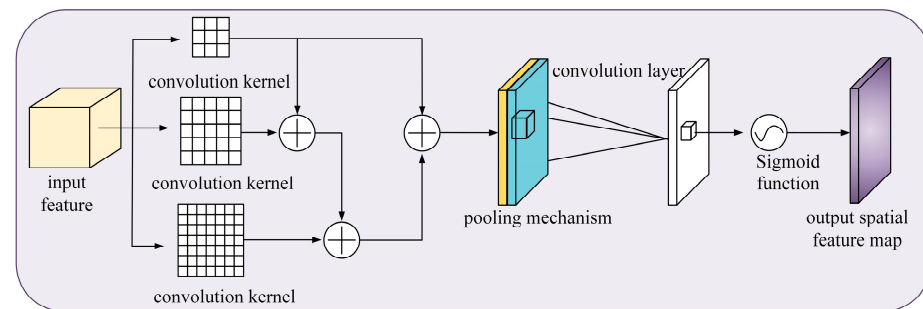


Figure 3. Spatial attention module used in the study.

The channel attention mechanism performs max-pooling and average pooling operations on the input feature maps in the width and height dimensions, respectively. The pooling layer outputs are then transmitted with a fully connected layer for weight parameter sharing with the help of weighting operation on the parameter-shared features. In addition, the sigmoid activation operation is applied to generate the final channel attention features. A schematic representation of the channel attention module is shown in Figure 4.

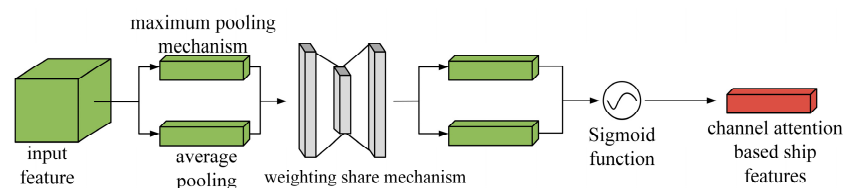


Figure 4. Channel Attention Module used in the study.

2.2.2. Details about CA2HRNET Framework

Utilization of the CAM attention mechanism allows the model to give more attention to the image spatial and channel scales in the feature extraction stage. Ship features will be further enhanced with the help of an efficient channel attention mechanism (ECA) after the up-sampling procedure. More important ship features will be assigned larger weights with channel-related information and mechanism. A schematic overview of the proposed CA2HRNET framework is shown in Figure 5. The entire CA2HRNET network consists of a backbone section, a feature integration section, and a prediction header section. In the backbone, the CA2HRNET network performs both down-sampling and up-sampling during the feature extraction process. In that way, different shapes of feature maps are obtained and feature fusion is performed. In the feature integration section, the network performs feature fusion of all the acquired feature maps by up-sampling the feature maps with smaller widths and heights. We can obtain a global feature map via this feature map fusion procedure, which primarily involves convolution, normalization, and activation operations. In the prediction header section, the network adjusts the number of channels to the number of categories with convolution operation (e.g., typical convolution kernel is 1×1). The potential width and height for the output layer are adjusted based on the input image size.

The proposed CA2HRNET network backbone is divided into four segments, segment-1, segment-2, segment-3, and segment-4. Segment-1 is the first part and performs the initial feature extraction. For a given input ship image, the network employs two convolution operations with step size (2, 2), convolution kernel size (3, 3) and 64 channels for height and width compression and feature extraction. Then, four bottleneck-block residual convolution operations are performed to merge ship features. The bottleneck-block layer involves two

convolution operators, a normalization operator, and an activation function, alongside one convolution and normalization in the trunk section. Moreover, the residual edge section either bypasses any processing or undergoes minimal processing before directly connecting to the output. The size of the feature map is obtained by implementing four bottleneck-block operations, while the feature map size is supposed to be (120, 120, 256).

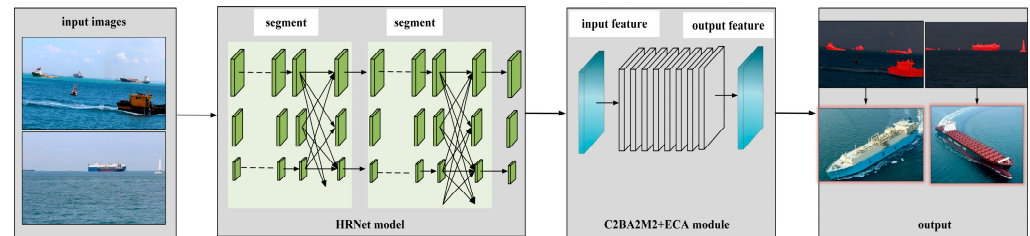


Figure 5. The proposed CA2HRNET framework overview.

Segment-2 is the second part and performs the feature extraction procedure. The feature layer input into segment-2 is processed using a convolution with a step size of (1, 1) and 32 channels and a convolution with a step size of (2, 2) and 64 channels to obtain a feature layer of (120, 120, 32) and a feature layer of (60, 60, 64). After that, four basic-block residual convolution operations are processed for each of the two feature layers. In the basic-block module, the backbone sections consist of a convolution, normalization, activation function and a convolution, normalization; The residual edge section either bypasses any processing or undergoes minimal processing before directly connecting to the output. After implementing four basic-block operations, a (120, 120, 32) feature layer and a (60, 60, 64) feature layer are obtained. The feature layer of (120, 120, 32) is then down-sampled and then added to the feature layer of (60, 60, 64); the feature layer of (60, 60, 64) is up-sampled and then added to the feature layer of (120, 120, 32). The potential output feature map sizes obtained are (120, 120, 32) and (60, 60, 64).

Segment-3 is responsible for further feature extraction. For the feature layer input into segment-3, convolution with a step size of (2, 2) and 128 channels is applied to process the (60, 60, 64) feature layer, and the output is a (30, 30, 128) feature layer. Therefore, there are three feature layers. The four basic-block residual convolution operations are performed on these three feature layers, and these three feature layers are utilized for up-sampling and down-sampling to establish a connection between the different feature maps, which can help achieve the high-level of feature fusion.

Segment-4 is responsible for the final feature fusion. For the feature layer input into segment-4, convolution with a step size of (2, 2) and 256 channels is applied to process the (30, 30, 128) feature layer, resulting in a (15, 15, 256) feature layer. Therefore, the four feature layers are processed with four basic-block residual convolution operations, which are used to establish connections with the up-sampling and down-sampling operations. Segment-4 outputs distinct yet discriminative features with the feature fusion procedure.

By constructing the above-mentioned segments, we can obtain four effective feature layers in segment-4, which are described as (128, 128, 32), (60, 60, 64), (30, 30, 128), and (15, 15, 256). These four feature layers enable network focus on the crucial yet non-trivial features in the image space and channel (i.e., the advantages of an attention mechanism), respectively. Then, feature fusion is performed on these four effective feature layers. The first three effective feature layers are first up-sampled, the height and width of the feature layers are adjusted to 128×128 , and then, the adjusted feature layers are stacked, before, finally, the feature integration of these four adjusted effective feature layers is performed using a convolution, normalization, and activation function. The features of the input image can be obtained through the above steps. The extracted features are allowed to pass through an ECA attention mechanism that allows the network to assign higher weights to the important channel information. Then, a 1×1 convolution is utilized to adjust the channels to the number of categories for the final feature layer, and up-sampling

is carried out by utilizing a resize operation to make the width and height of the final output layer consistent with that of the input image.

3. Experimental Section

3.1. Experimental Data

For the ship recognition study, three maritime video clips of typical maritime traffic situations were collected to validate the results of the semantic segmentation of ships. We used a similar dataset and dataset collection rule in our previous study [11]. Scenario 1 is a scene with lower visibility and the presence of occlusion by two large ships. Scenario 2 is a scene with good weather conditions and the presence of occlusion by two large ships. Scenario 3 is a scene with good weather conditions and the presence of a fast-moving near target. Part of the dataset is shown in Figure 6 and detailed maritime video parameters are shown in Table 1. For the purpose of model training, we have made a set of training image sequences with the help of annotation segmentation software labelMe (Version 5.0.1). The raw images in Figure 6a,c,e demonstrate typical ship image samples, while Figure 6b,d,f are the corresponding labeling images. To mitigate the image overfitting phenomenon, data augmentation was applied to the input training set in the dataset (e.g., image rotation, specifically in the form of left-right flipping and up-down flipping) to improve the stability and robustness of the model.

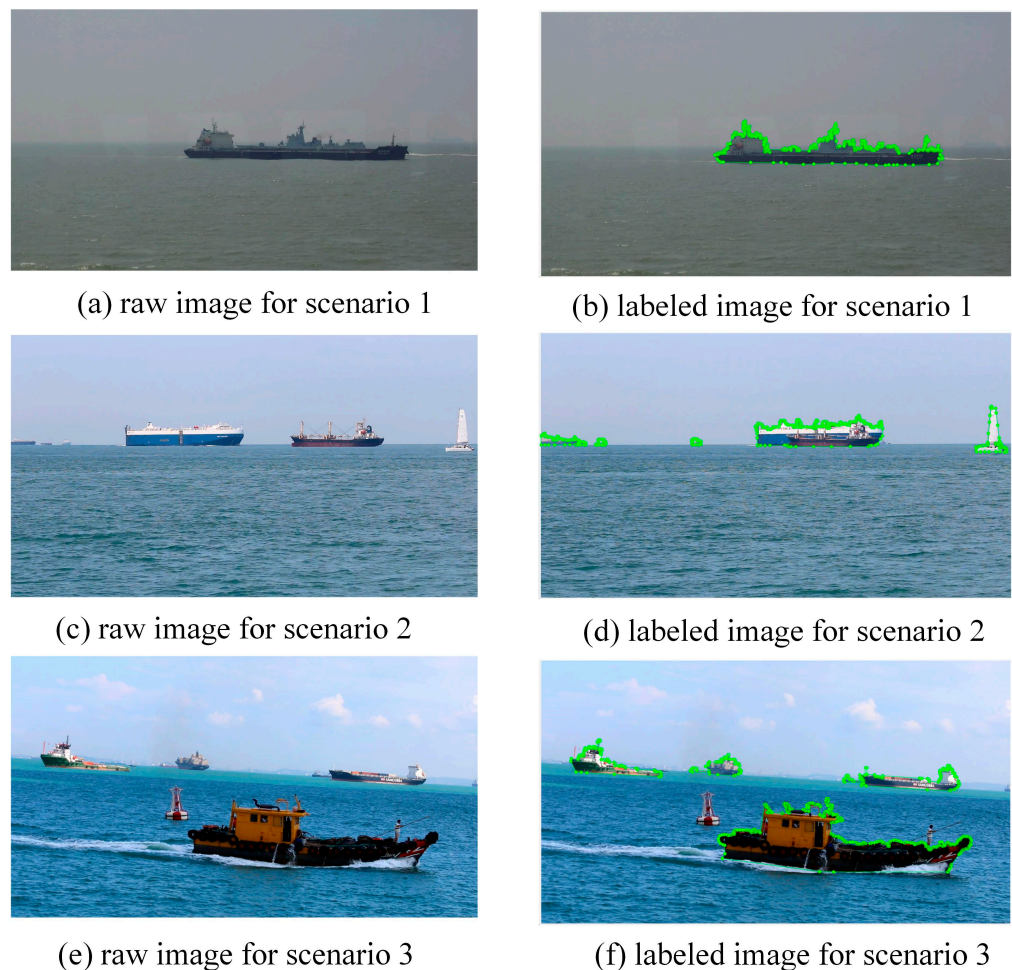


Figure 6. The input raw maritime images and training samples.

Table 1. Descriptions of the Maritime Video clips.

Video Number	Frame Rate	Image Resolution	Video Duration	Environmental Feature
scenario 1	30	1920 × 1080	15	slightly cloudy, fewer ships, no waves, low visibility
scenario 2	30	1920 × 1080	18	slightly cloudy, fewer ships, no waves, high visibility
scenario 3	30	1920 × 1080	9	cloudy, many ships, waves, high visibility

3.2. Experimental Setups

The experiments are implemented on the Ubuntu operation system (the corresponding version is Ubuntu 23.10), and the TensorFlow deep learning framework is used as a baseline for the experiment. The experimental platform has 16 GB memory, a 4 GHz CPU, and an RTX4000 GPU. Python 3.7 version is used to implement the various models used in the study. During the experiment, CuDNN, Cuda, OpenCV, and other third-party dependent libraries are used, and corresponding software packages are installed according to the requirements for each model to ensure the implementation of the network. The CuDNN version is 8.8.0, and Cuda toolkit is 12.2.2, and opencv version is 3.4.16.

3.3. Evaluation Index

Ship segmentation can be regarded as a type of binary classification of pixels. Overall, the to-be-detected ship target pixels are considered the foreground and the rest of the pixels are considered the background. To accurately evaluate the segmentation results, the true positions of the ship target pixel are manually marked in the experiment. The effectiveness of the algorithm is evaluated by calculating the difference between the detected and ground-truth values in the test set. The study employs accuracy (Acc_s), precision (P_c), F1-score ($F1_s$), intersection over union (IoU), and frequency-weighted IoU (F_{IoU}) indicators to quantify model performance. The formulas for calculating Acc_s , P_c , $F1_s$, IoU, and F_{IoU} can be found in Equations (4)–(8), respectively.

$$Acc_s = \frac{D_B + D_H}{D_B + D_H + P_B + P_H} \quad (4)$$

$$P_c = \frac{D_H}{D_H + P_H} \quad (5)$$

$$F1_s = 2 \times \frac{P_c \times R_e}{P_c + R_e} \quad (6)$$

$$IoU = \frac{D_H}{D_H + P_H + P_B} \quad (7)$$

$$F_{IoU} = \frac{D_H + P_B}{D_H + D_B + P_H + P_B} \times \frac{D_H}{D_H + P_H + P_B} \quad (8)$$

where D_B represents the number of pixels correctly classified as background; D_H represents the number of pixels correctly classified as a ship target; P_B represents the number of background pixels misclassified as ship targets; and P_H represents the number of ship target pixels misclassified as background.

3.4. Results and Analysis

For the purpose of verifying the ship segmentation accuracy of the proposed CA2HRNET network, we have also implemented other segmentation models that use an attention mechanism as well as other typical semantic segmentation networks. More specifically, the conven-

tional HRNet [19] is implemented, and we have enhanced the performance of the basic HRNet model by adding a squeeze and excitation mechanism (abbreviated as HRNet-SE). In addition, the basic HRNet model is also improved by incorporating a coordinate attention mechanism (abbreviated as HRNet-CA). Moreover, the PSP Net is also implemented to verify the segmentation performance of the proposed framework [20]. The UNet segmentation model [21] and DeeplabV3+ [22] are also implemented for the purpose of model performance comparison.

3.4.1. Scenario 1 Experimental Results

There are two large ships overlapping each other in scenario 1. The imaging size of the ships is quite large and the traffic volume is low. Since the ships are in the distance (i.e., far away from the camera), the contours of the ships are blurred, which challenges model segmentation accuracy. Despite this, our proposed framework achieves an excellent performance. The segmentation results of the different models are shown in Figure 7. It can be seen from the figure that each network has identified the ship contours; in particular, the HRNet with added attention mechanism exhibited better effect than the unimproved HRNet. Especially in the bow, CA2HRNET better expresses the details of the foremast of the forecandle and the superstructure and antenna of the stern are also well expressed. Compared with other networks, the UNet network performs well, and the corresponding contour of the ship is well expressed. DeepLabv3+ network performance is not good, from the enlarged detail map, it can be seen that the details of the bow part are not well expressed, with instances of misidentification of hulls and backgrounds. The PSPNet network displays the worst ship segmentation accuracy, with no corresponding contour of the hull at all, and the foremast being completely unrecognized in the zoomed-in detail image.

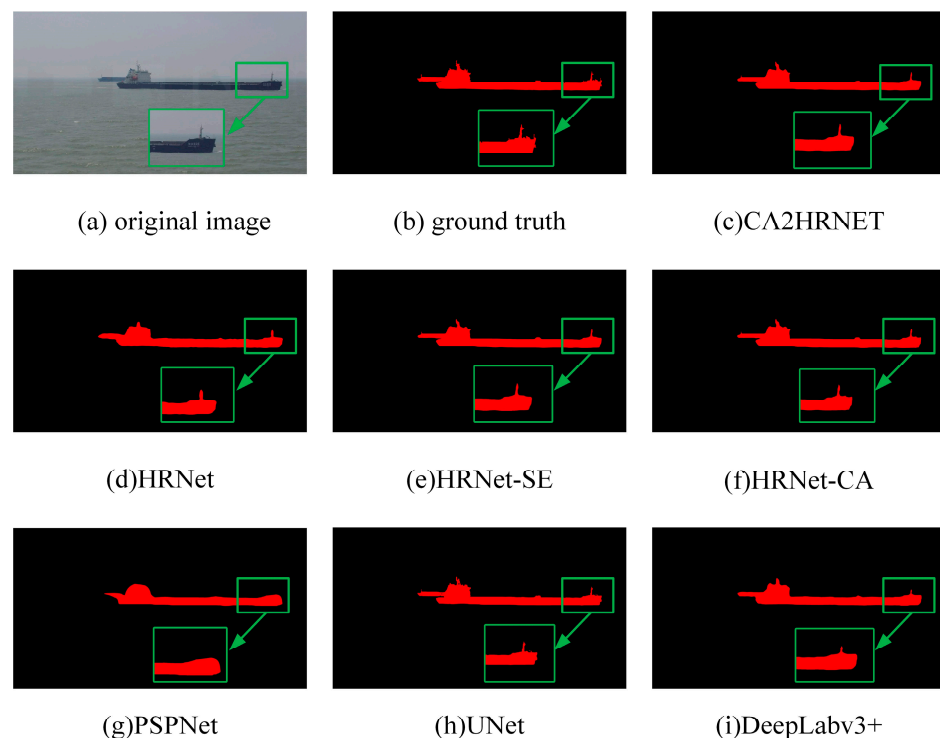


Figure 7. Ship Semantic segmentation of scenario 1.

The segmentation results of the different models are shown in Table 2. According to the results of the five evaluation indexes, the varied ship segmentation models have shown satisfactory results, as the accuracy values are larger than 99% for scenario 1. Compared with the original segmentation network, the networks with different attention mechanisms achieved better segmentation results. The proposed CA2HRNET model outperformed its counterparts with a segmentation accuracy of 97.81%. More specifically, the majority

of ship-related pixels can be correctly identified by the proposed CA2HRNET model. Compared with the HRNet-SE network and HRNet-CA network, the accuracy of the CA2HRNET network was 0.02% and 0.28% higher, respectively. This may be because the newly added attention mechanism results in a significant improvement.

Table 2. Ship segmentation performance comparison for scenario 1.

MODEL	Acc_s	P_c	$F1_s$	IoU	F_{IoU}
CA2HRNET	99.81%	97.81%	97.56%	97.53%	99.63%
HRNet-SE	99.81%	97.79%	97.53%	97.49%	99.63%
HRNet-CA	99.81%	97.53%	97.52%	97.50%	99.62%
HRNet	99.67%	95.31%	95.78%	95.78%	99.36%
PSPNet	99.41%	93.03%	92.34%	92.58%	98.87%
UNet	99.77%	98.47%	96.97%	96.94%	99.54%
DeepLabv3+	99.68%	94.22%	95.89%	95.88%	99.37%

The proposed framework achieved better ship segmentation performance because the spatial attention mechanism and channel attention mechanism helped the model obtain ship trivial visual features, and thus, the ship's bow-like pixels can be more accurately identified and segmented by the model. The segmentation accuracy and precision may be slightly or significantly improved based on the background interference severity in the maritime images.

3.4.2. Experimental Results in Scenario 2

The traffic volume in scenario 2 is obviously larger than that in scenario 1, and ships in scenario 2 can be clearly identified from the maritime images (i.e., ship imaging sizes in scenario 2 were significantly larger than those in scenario 1). For ships with large imaging areas, the network can better learn the hull features due to the presence of many semantic pixels of ships, which can significantly help the model recognize ship contours from images. However, for ships with small imaging areas, it is difficult for the network to classify the pixels of the ship from those of the background as there are fewer pixels for the network to learn. Therefore, it is more challenging for the network to classify ship and background pixels accurately and ship identification with small imaging sizes is more important.

We have also selected a series of image segmentation results to demonstrate varied model performance for the second scenario. The segmentation results of the different models are shown in Figure 8. It can be observed that for ships with larger imaging areas, there is more overlapping area between general cargo ships and bulk carriers. The CA2HRNET network better expresses the detailed contours of the superstructure of the overlapping part. The HRNet-SE network performed better than the other non-CA2HRNET networks, as only a few contour details were missed. For small target ships with small imaging areas, the various HRNet series networks successfully identified the small ships near the skyline. It can be inferred that the HRNet network retains the high-resolution feature map throughout the whole process, and the spatial information of the image is successfully captured by the segmentation model. The important information of the spatial domain and channel domain is extracted effectively by the introduced attention mechanism. The multi-level up-sampling mechanism of feature maps used by the network implemented better fusion and expression of feature maps with different resolutions.

Table 3 shows the segmentation performance of the different models according to five statistical indicators. It is found that the CA2HRNET network segmentation reaches 97.14% in terms of accuracy, which is 0.14% and 0.19% larger than the HRNet-SE network and HRNet-CA network, respectively. The CA2HRNET model outperformed the PSPNet, UNet, and DeepLabv3+ networks by 3.26%, 0.14%, and 1.69%, respectively, in terms of segmentation accuracy. It can be seen that the values of the statistical indicators in scenario 2 obtained by the various models are relatively smaller than those in scenario 1. After carefully checking the ship segmentation results, it was found that the false-detection

rate in scenario 2 was higher than that in scenario 1 due to interference caused by small and medium-sized ship targets in the maritime images (i.e., ship boundaries were also too blurred to be detected). Based on the values of five evaluation indexes, the CA2HRNET network achieved excellent results, and the accuracy of the HRNet network with attention mechanism reached 99.80%, which is 0.04% higher than the second-best UNet network.

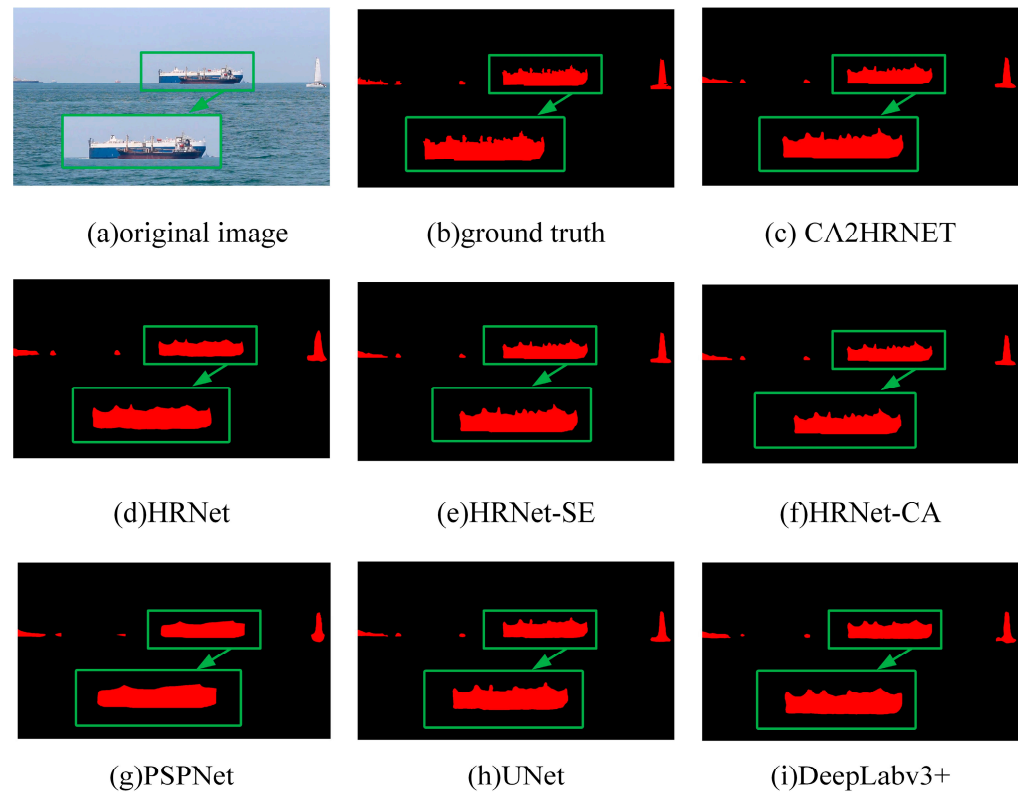


Figure 8. Ship Semantic segmentation of scene 2.

Table 3. Comparison of segmentation accuracy for Scenario 2.

MODEL	Accuracy	Precision	F1-Score	mIoU	FWIoU
CA2HRNET	99.80%	97.14%	96.59%	96.60%	99.61%
HRNet-SE	99.80%	97.00%	96.55%	96.57%	99.61%
HRNet-CA	99.80%	96.95%	96.53%	96.55%	99.60%
HRNet	99.63%	92.65%	93.61%	93.80%	99.28%
PSPNet	99.57%	93.88%	92.45%	92.76%	99.16%
UNet	99.76%	97.00%	95.74%	95.79%	99.52%
DeepLabv3+	99.65%	95.45%	94.00%	94.16%	99.33%

3.4.3. Experiment Results in Scenario 3

The traffic volume in scenario 3 is larger than that of both scenario 2 and scenario 1, and ship speeds in the third scenario were on average higher than in the previous two scenarios (it is noted that the fishing boat in the third video moves quite fast). Due to the fast speed of the fishing boat, the position feature of the fishing boat pixels showed a rapid change. This imposes a significant challenge for the learning of the network. However, at the same time, there are more abundant image pixels of ships near the skyline, which helps the network to learn and recognize ship contour details.

The segmentation results of the different models for the scenario 3 are illustrated in Figure 9. It can clearly be seen that each segmentation model can well separate ship pixels from those of the skyline. Similar to the segmentation performance in scenario 1 and scenario 2, the CA2HRNET network has a good performance in the segmentation of

the detailed contours of the prospective ships. Indeed, the proposed CA2HRNET network showed the best segmentation results and the overlap between the detail map and the real label is quite large. In addition, the HRNet networks that integrate other attention mechanism networks (i.e., HRNet-CA and HRNet-SE) also show better segmentation than the original HRNet network.

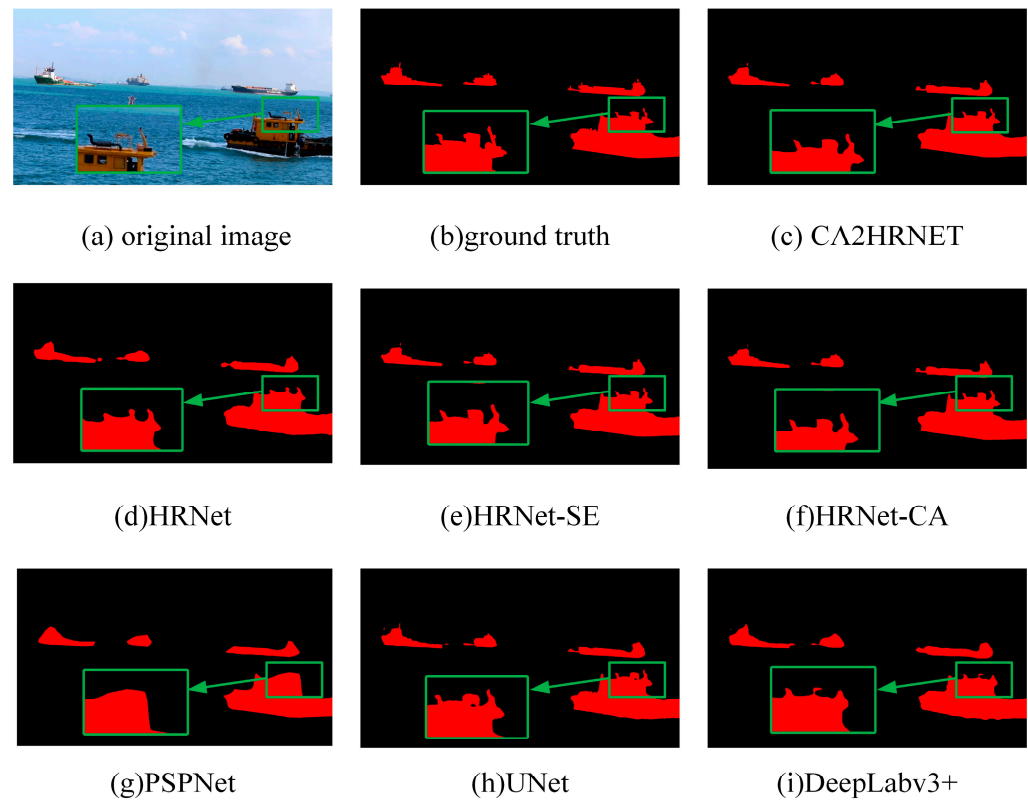


Figure 9. Semantic segmentation of scene 3.

Statistical evaluators for the third scenario are shown in Table 4. It can be seen that the accuracy of the CA2HRNET network reaches 97.69%, which outperforms the HRNet-CA network and HRNet network by 0.09% and 4.11%, respectively, in terms of the precision-related indicator. Moreover, the CA2HRNET network outperformed the PSPNet, UNet, and DeepLabv3+ networks by 3.69%, 0.33%, and 4.62%, respectively. Based on the above-mentioned analysis, the CA2HRNET network achieved better performance due to the advantages provided by the introduced hybrid attention mechanisms. The accuracy of the HRNet network with the introduced attention mechanism reaches about 99.680%, which is 0.02% higher than that of the UNet network.

Table 4. Comparison of segmentation accuracy for Scenario 3.

MODEL	Accuracy	Precision	F1-Score	mIoU	FWIoU
CA2HRNET	99.69%	97.69%	96.85%	96.79%	99.40%
HRNet-SE	99.68%	97.69%	96.72%	96.66%	99.37%
HRNet-CA	99.68%	97.60%	96.76%	96.70%	99.38%
HRNet	99.43%	93.58%	94.26%	94.27%	98.91%
PSPNet	99.33%	93.73%	93.06%	93.16%	98.70%
UNet	99.66%	97.36%	96.51%	96.45%	99.33%
DeepLabv3+	99.42%	93.43%	94.14%	94.16%	98.88%

4. Conclusions

To address small to medium-sized ship identification and recognition from maritime images, the study proposes a novel CA2HRNET model with the help of spatial attention and channel attention mechanisms. The proposed CA2HRNET model extracts effective feature maps in terms of channels and spatial dimensions from input maritime image sequences. The proposed ship semantic segmentation model generates and connects feature maps with high-resolution to low-resolution in the encoder layer. Overall, the CA2HRNET network retained the advantages of HRNet by generating four feature maps with different resolutions in a parallel connection manner rather than a cascaded manner. In this way, ship features with high resolution for each layer can be successfully retained, and thus, ship segmentation accuracy is significantly enhanced. The proposed model has shown good performance for maritime images overwhelmed with blurred ship contours. The CAM attention mechanism is further proposed to enrich the diversity of receptive fields due to the limited number of the four parallel layers. Channel attention, multi-scale spatial attention, and adaptive weight fusion are also integrated to improve ship segmentation model performance, which is robust against ship scale variation and visual shape variation challenges. Experimental results show that the proposed ship segmentation model showed accurate ship segmentation model performance (i.e., the average ship segmentation accuracy was larger than 97%). The following improvements can be explored to expand our study. First, we can further enhance our model performance utilizing a lightweight model, and model computational cost can also be evaluated. Second, we can collect maritime videos under additional adverse weather conditions (e.g., storm, low-visibility, night-time weather) for the purpose of improving model usage in real-world maritime activities. Third, it was found that ship detection performance for the various models was quite similar. The primary reasons can be ascribed to ship detection and segmentation model robustness and noise-free input maritime images. In future work, we may also verify ship detection and segmentation model performance when using low-quality maritime image clips.

Funding: This work was jointly supported by National Natural Science Foundation of China (52331012, 52102397).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can be accessed by sending an email to the author.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Cheng, S.; Zhu, Y.; Wu, S. Deep learning based efficient ship detection from drone-captured images for maritime surveillance. *Ocean Eng.* **2023**, *285*, 115440. [\[CrossRef\]](#)
2. Pinault, L.J.; Yano, H.; Okudaira, K.; Crawford, I.A. YOLO-ET: A Machine Learning model for detecting, localising and classifying anthropogenic contaminants and extraterrestrial microparticles optimised for mobile processing systems. *Astron. Comput.* **2024**, *47*, 100828. [\[CrossRef\]](#)
3. Gladis, K.A.; Madavarapu, J.B.; Kumar, R.R.; Sugashini, T. In-out YOLO glass: Indoor-outdoor object detection using adaptive spatial pooling squeeze and attention YOLO network. *Biomed. Signal Process. Control* **2024**, *91*, 105925.
4. Chang, H.; Fu, X.; Lu, J.; Guo, K.; Dong, J.; Zhao, C.; Feng, C.; Li, Z.; Zhang, Y. SPANet: A Self-Balancing Position Attention Network for Anchor-Free SAR Ship Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 8363–8378. [\[CrossRef\]](#)
5. Ren, Z.; Tang, Y.; Yang, Y.; Zhang, W. SASOD: Saliency-Aware Ship Object Detection in High-Resolution Optical Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [\[CrossRef\]](#)
6. Lee, K.-J.; Lee, S.-J.; Chang, J.-Y. A Study on Ship Detection and Classification Using KOMPSAT Optical and SAR Images. *Ocean Sci. J.* **2024**, *59*, 10. [\[CrossRef\]](#)
7. Yasir, M.; Jianhua, W.; Mingming, X.; Hui, S.; Zhe, Z.; Shanwei, L.; Colak, A.T.I.; Hossain, M.S. Ship detection based on deep learning using SAR imagery: A systematic literature review. *Soft Comput.* **2023**, *27*, 63–84. [\[CrossRef\]](#)
8. Jiang, X.; Cai, J.; Wang, B. YOLOSeaShip: A lightweight model for real-time ship detection. *Eur. J. Remote Sens.* **2024**, *57*, 2307613. [\[CrossRef\]](#)

9. Wang, W.; Fu, Y.; Dong, F.; Li, F. Semantic segmentation of remote sensing ship image via a convolutional neural networks model. *IET Image Process.* **2019**, *13*, 1016–1022. [\[CrossRef\]](#)
10. Manar, A.; Kim, S. IR/EO Ship Detection and Tracking Using SiamMask. In Proceedings of the 2022 22nd International Conference on Control, Automation and Systems (ICCAS), Jeju, Republic of Korea, 27 November–1 December 2022; pp. 1604–1606.
11. Chen, X.; Wang, M.; Ling, J.; Wu, H.; Wu, B.; Li, C. Ship imaging trajectory extraction via an aggregated you only look once (YOLO) model. *Eng. Appl. Artif. Intell.* **2024**, *130*, 107742. [\[CrossRef\]](#)
12. Chen, X.; Dou, S.; Song, T.; Wu, H.; Sun, Y.; Xian, J. Spatial-Temporal Ship Pollution Distribution Exploitation and Harbor Environmental Impact Analysis via Large-Scale AIS Data. *J. Mar. Sci. Eng.* **2024**, *12*, 960. [\[CrossRef\]](#)
13. Sharma, R.; Saqib, M.; Lin, C.T.; Blumenstein, M. MASSNet: Multiscale Attention for Single-Stage Ship Instance Segmentation. *Neurocomputing* **2024**, *594*, 127830. [\[CrossRef\]](#)
14. Zhang, Y.; Li, C.; Shang, S.; Chen, X. SwinSeg: Swin transformer and MLP hybrid network for ship segmentation in maritime surveillance system. *Ocean Eng.* **2023**, *281*, 114885. [\[CrossRef\]](#)
15. Tzortzis, G.; Sakalis, G. A dynamic ship speed optimization method with time horizon segmentation. *Ocean Eng.* **2021**, *226*, 108840. [\[CrossRef\]](#)
16. Sánchez Pedroche, D.; García Herrero, J.; Molina López, J.M. Context learning from a ship trajectory cluster for anomaly detection. *Neurocomputing* **2024**, *563*, 126920. [\[CrossRef\]](#)
17. Sun, Y.; Su, L.; Yuan, S.; Meng, H. DANet: Dual-Branch Activation Network for Small Object Instance Segmentation of Ship Images. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 6708–6720. [\[CrossRef\]](#)
18. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5686–5696.
19. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Liu, Q.; Wang, S.; Dai, Y.; Zhang, J.; Wang, Y.; Zhou, R. Improved PSP-Net Segmentation Network for Automatic Detection of Neovascularization in Color Fundus Images. In Proceedings of the 2022 IEEE International Conference on Visual Communications and Image Processing (VCIP), Suzhou, China, 13–16 December 2022; pp. 1–5.
21. Ortega-Ruiz, M.A.; Karabağ, C.; Roman-Rangel, E.; Reyes-Aldasoro, C.C. DRD-UNet, a UNet-Like Architecture for Multi-Class Breast Cancer Semantic Segmentation. *IEEE Access* **2024**, *12*, 40412–40424. [\[CrossRef\]](#)
22. Anilkumar, P.; Venugopal, P.; Maddikunta, P.K.R.; Gadekallu, T.R.; Al-Rasheed, A.; Abbas, M.; Soufiene, B.O. An Adaptive DeepLabv3+ for Semantic Segmentation of Aerial Images Using Improved Golden Eagle Optimization Algorithm. *IEEE Access* **2023**, *11*, 106688–106705. [\[CrossRef\]](#)

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.