remote sensing

MDPI

Article

CSEF-Net: Cross-Scale SAR Ship Detection Network Based on Efficient Receptive Field and Enhanced Hierarchical Fusion

Handan Zhang and Yiquan Wu *

School of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; handan_z@nuaa.edu.cn
* Correspondence: imagestrong@nuaa.edu.cn; Tel.: +86-137-7666-7415

Abstract: Ship detection using synthetic aperture radar (SAR) images is widely applied to marine monitoring, ship identification, and other intelligent maritime applications. It also improves shipping efficiency, reduces marine traffic accidents, and promotes marine resource development. Land reflection and sea clutter introduce noise into SAR imaging, making the ship features in the image less prominent, which makes the detection of multi-scale ship targets more difficult. Therefore, a cross-scale ship detection network for SAR images based on efficient receptive field and enhanced hierarchical fusion is proposed. In order to retain more information and lighten the weight of the network, an efficient receptive field feature extraction backbone network (ERFBNet) is designed, and the multi-channel coordinate attention mechanism (MCCA) is embedded to highlight the ship features. Then, an enhanced hierarchical feature fusion network (EHFNet) is proposed to better characterize the features by fusing information from lower and higher layers. Finally, the feature map is input into the detection head with improved bounding box loss function. Using SSDD and HRSID as experimental datasets, average accuracies of 97.3% and 90.6% were obtained, respectively, and the network performed well in most scenarios.

Keywords: synthetic aperture radar; ship detection; deep learning; attention mechanism; feature fusion

Citation: Zhang, H.; Wu, Y. CSEF-Net: Cross-Scale SAR Ship Detection Network Based on Efficient Receptive Field and Enhanced Hierarchical Fusion. *Remote Sens.* **2024**, *16*, 622. https://doi.org/10.3390/rs16040622

Academic Editors: Mohammad Awrangjeb, Danfeng Hong, Shou Feng, Nan Su, Chunhui Zhao, Yiming Yan and Qingsheng Xue

Received: 28 December 2023
Revised: 30 January 2024
Accepted: 4 February 2024
Published: 7 February 2024

1. Introduction

SAR is a coherent imaging system that can produce high resolution remote sensing images without time constraints and regardless of extreme weather conditions, such as cloud cover [1]. Having high adaptation to ocean monitoring with fluctuating climate [2], SAR has growing importance for the detection of military and civilian ships [3], with roles such as controlling sea traffic, fisheries monitoring, protection of marine ecology, disaster relief, and other applications. SAR image ship detection can effectively improve ship transportation efficiency and reduce maritime traffic accidents. However, because of the disturbance of sea surface clutter and land building reflection, the ship features in SAR images are weakened. Coupled with the large differences in the size of ships, it is difficult to detect them all at the same time, so the detection of ships in SAR images has always been a research hotspot in remote sensing image processing.

Most of the early SAR image ship target detection methods used constant false alarm rate detector (CFAR) [4], which is based on the principle of constructing mathematical and statistical models of sea surface and ship targets, and the process is simple and efficient. However, in real situations different sea surfaces have different statistical properties, and objects that are not the ship targets can also affect the accurate modeling. Therefore, some improved CFAR algorithms have been proposed, one of which is based on the implementation of more optimal statistical distributions, such as Lognormal, Gamma, or Alpha distributions [5–7]. There is also one that utilizes wavelet transform and template matching [8,9] for ship target detection. These methods are more suitable for fixed sea surface scenes with less interference, and require a lot of preliminary work to analyze the

characteristics of the target and the background. Because of high parameter sensitivity, it is difficult to make satisfactory detection performance in complex environments. In addition to the CFAR method, a saliency-based ship detection method for SAR images has also been proposed [10]. Wang et al. [11] studied the dissimilarity and the similarity between the target and background pixels to construct a pattern recursion-based saliency ship detector, which was experimentally shown to be robust even in complex backgrounds. Although the saliency-based methods have achieved better detection results, they still have some shortcomings. For example, it is difficult to distinguish between the background and the target when they are weakly differentiated, so the detection accuracy is reduced. And the saliency map can only show the rough range of the target and cannot provide precise location and shape information.

In recent years, deep learning has been used in SAR image ship detection. Neural network models can automatically extract target feature information from images by reading a large amount of annotation data, carry out optimization learning, and finally output prediction results. Compared with the traditional methods, it has the advantages of high detection accuracy and efficiency, strong anti-interference, etc. which has become the commonly used method of SAR ship detection. Generally, it can be divided into one-stage and two-stage detection networks. The two-stage detection network gives the pre-selected box of the region of interest before performing the prediction box regression computation, and the representative network is Faster-RCNN [12]. Wang et al. [13] evaluated the regions generated by Faster R-CNN, using the maximum stability extreme region method, which optimized the original threshold decision method and reduced false positives of the model. Lin et al. [14] added the SE attention mechanism to the Faster-RCNN network, which gives different weights to different locations in the channel through the weight matrix to highlight important features and enhance the detection performance in the nearshore region. For the irregularity of the ship shape and to better extract the geometric features of the ship, Ke et al. [15] used deformable convolution to construct a Faster R-CNN network. Jiao et al. [16] proposed a thickly-linked network based on Faster-RCNN in order to improve the adaptability of the model in different scenarios by combining one feature map with each other and merging them together. However, constructing the network was time-consuming. Xu et al. [17] designed a grouped feature fusion module to achieve the information interaction between different polarization features. It not only improves the multi-scale feature extraction ability of the model, but also makes full use of the differences between various polarization features.

The one-stage detection network eliminates the region suggestion generation part, simplifying the object detection problem to a regression problem, which is characterized by a high rate, and the representative networks are YOLO series [18] and SSD [19]. Liu et al. [20] introduced coordinate attention to the YOLOv7-tiny model and improved the spatial pyramid pool (SPP) and SIoU loss function to strengthen detection performance. Cheng et al. [21] used non-local mean as a denoising method for SAR images to preprocess detected images, and then proposed a feature thinning module to suppress background interference and improve the positioning performance of the model. Zhang et al. [22] designed five modules to form a high-precision detection network called HyperLi-Net. Most of these approaches are aimed at improving the detecting precision, but ignore the fact that the complexity and computational load of the model will increase as the network goes deeper.

To lighten the model, some scholars introduce the lightweight network module. Drawing on the idea of GhostNet, Tian et al. [23] put the ghost module into the backbone of the RetinaNet network as a shallow convolutional layer to generate only some of the new channels, and at the same time decrease the number of deeper convolutional layers to reduce the overall computation of the network. Zhang et al. [24] built the lightweight ship detector ShipDeNet-20 for real-time detection, whose model size is less than 1 MB. Although it used fewer convolutional layers and smaller convolutional cores, it ensured the original detection accuracy through a feature enhancement module and a scale sharing

module. Kong et al. [25] chose YOLOx-Tiny as the baseline model and parallelized three different convolutions to form a lightweight feature extraction module, while using the more computationally efficient SPPF module to decrease the number of parameters. Most of the above methods use convolutional modules with fewer parameters to build models, which makes the backbone network inevitably lose some pixel information during feature extraction. For ship targets in SAR images, the pixel area difference between targets is large due to the different ship sizes, resulting in weak detection capabilities of small targets.

To overcome multiple targets in SAR images, more than one approach has been presented. Suo et al. [26] fused low-level spatial information with high-level semantic information across hierarchical levels to solve the problem of difficult detection due to large changes in ship size. Li et al. [27] designed multiple pyramid modules, and each contained a different combination of convolutional layers. They cascaded and juxtaposed them to obtain contextual fusion information. Zhu et al. [28] repeatedly concatenated two-pyramid networks, each with the same feature fusion structure, and finally connected the two outputs, which shortened the path of the same feature flow and facilitated multiscale feature fusion. Zhang et al. [29] constructed four different feature pyramid modules and cascaded them in a certain order to form Quard-FPN, which performs multi-scale feature fusion and improves the multi-scale detection ability of the model. In addition, they proposed the concept of balanced learning [30] for the first time, solving the imbalance problem of four different angles in SAR image ship detection and improving the detection accuracy.

Although advances in deep learning have significantly enhanced the accuracy and efficiency of detection, SAR image ship detection continues to present substantial challenges. In general, most of the current ship detection networks can achieve high-precision detection in a simple background. When sea conditions are variable, however, ships may be moored at the shore or entering and exiting the harbor, in this case, buildings in the near-shore area can cause interference. And detection will be more difficult if the ships are densely arranged. Consequently, most models fail to locate ships adequately, resulting in more missed detections. On the other hand, since the network performs multiple down sampling when extracting features, feature information of targets that occupy less pixel area is lost, which leads to a serious problem of missed detection of small-scale ships. Especially when large ships are present, it is easy to misidentify small ships as noise, so improving multi-scale detection accuracy is also a problem that needs to be solved.

To address the above issues, CSEF-Net, a new SAR image ship detection network based on large receptive field and cross-scale feature fusion is proposed. Compared with the baseline YOLOv7 network, the capability of feature extraction and fusion is improved, and the loss function of boundary box regression is optimized. The experimental results on the SSDD dataset, the HRSID dataset and the LS-SSDD dataset show the effectiveness of the proposed method. Our contributions can be categorized as follows:

1. The CSEF-Net network is proposed to enhance the accuracy of detecting ships across different scales in SAR images under complex scene conditions.
2. To improve the feature extraction ability of the backbone network without adding too many parameters, an efficient receptive field feature extraction backbone network (ERFBNet) is designed to enlarge the receptive field and retain more effective information. Meanwhile, an effective attention mechanism and a lighter convolutional aggregation module are introduced.
3. To promote the flow of features on different scales and merge contextual information, an enhanced hierarchical feature fusion network (EHFNet) is designed. This network aims to provide more accurate location and semantic information, which includes weighted fusion and a skip layer connection. Moreover, a new down sampling module is designed.
4. Based on the characteristics of cross-scale ship targets, a more effective loss function of boundary box regression is designed, which is conducive to the detection of target position and improves the overall detection accuracy of the network.

2. Methods

2.1. Overall Network Structure of the CSEF-Net

YOLOv7 [31] is a relatively mature single-stage detection network of the current YOLO series, which balances the detection accuracy and detection speed very well and shows excellent performance in many target detection tasks. There are several versions of YOLOv7, such as YOLOv7-tiny, YOLOv7x, YOLOv7e6, and YOLOv7w6. Basic YOLOv7 is selected as the baseline model to build the CSEF-Net network. The overall network framework is shown in Figure 1. According to the classical network architecture, it is divided into three parts: backbone network, feature fusion network, classification and prediction network. Firstly, the backbone network extracts the key characteristics of the input image, followed by the output of three different scale feature maps. Secondly, the three feature maps are multi-scale fused via the feature fusion network. Then, the detection head is used for training and prediction, and the prediction results along with the real labels are fed into the loss function for calculating and optimizing. Finally, the accurate ship target position is obtained by eliminating the redundant detection frame through non-maximum suppression.

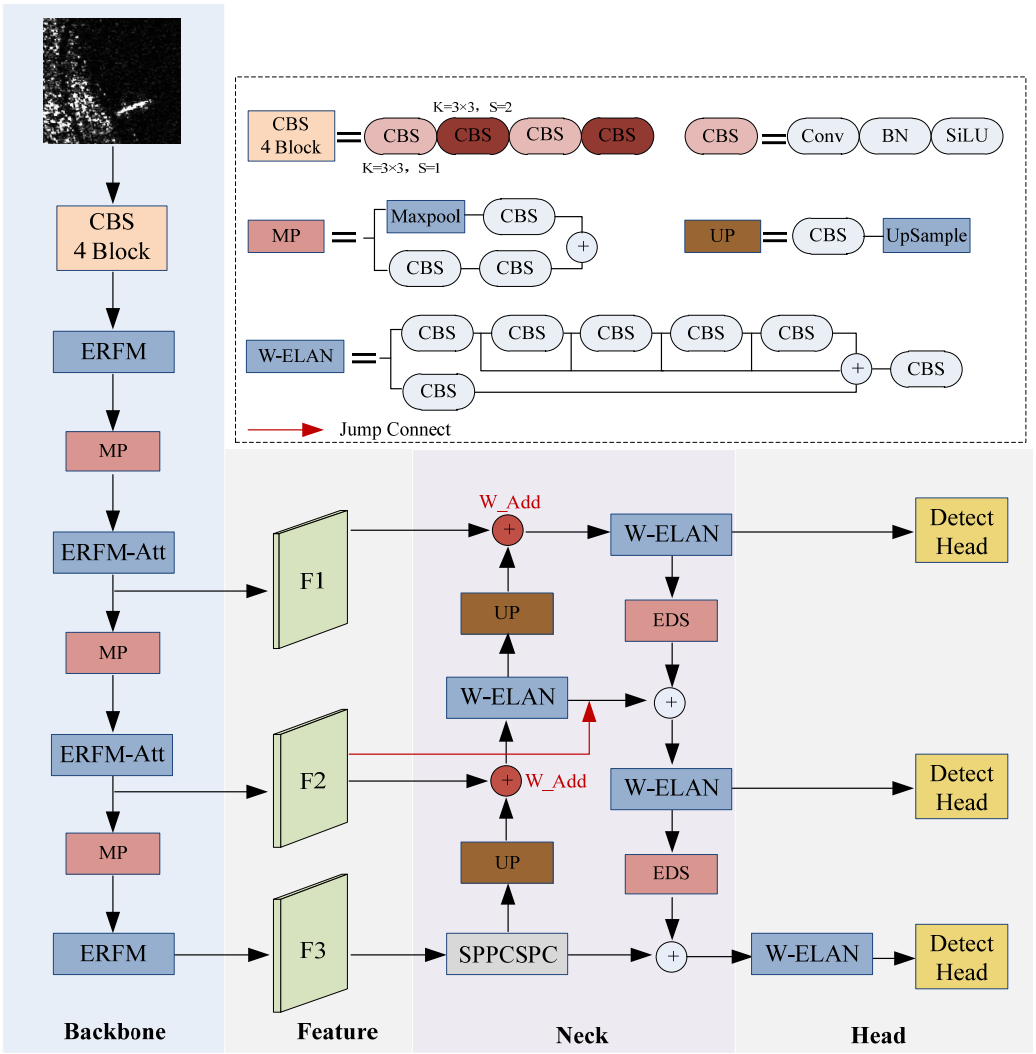


Figure 1. Overall network structure of CSEF-Net.

Specifically, CSEF-Net mainly includes three parts: efficient receptive field feature extraction backbone network ERFBNet, enhanced hierarchical feature fusion network EHFNet and detection head with improved bounding box loss function. Firstly, an efficient

receptive field module, ERFM, is designed to construct ERFBNet. By introducing depthwise separable convolution and partial convolution, the computation is reduced, making the model lightweight. At the same time, three dilated convolutions are parallel to expand the receptive field and retain more image information. A multi-channel coordinate attention mechanism (MCCA) is created, which has the ability to emphasize target features and minimize background noise. Then it is embedded in the ERFM module to improve the feature extraction capability of the backbone network.

Then, to improve the multi-scale feature fusion capability and cross-scale target detection accuracy of the network, a top-down and bottom-up enhanced hierarchical feature fusion network EHFNet is constructed. Specifically, weighted nodes and feature multiplexing connections are introduced in the downward feature fusion process to screen more important information, and improve the information mobility and shorten the path between different layers. To optimize the feature fusion upward process, an improved down sampling module is incorporated to effectively preserve more channel information and decrease the loss of target information, especially for small targets. After the input feature map passes through EHFNet, the fusion feature map with richer location and semantic information will be obtained.

Finally, the fused three feature maps of different scales are input into the detection head network for training to predict the ship target. The model parameters are optimized by the loss function, where the new bounding box regression loss function consists of Wise-IOU and NWD metrics. Moreover, non-maximum suppression (NMS) is used to filter the prediction boxes and suppress the redundant detection boxes.

2.2. Efficient Receptive Field Feature Extraction Backbone Network (ERFBNet)

The structure of ERFBNet is shown in Figure 1. The input image is first sampled twice by four CBS convolution modules. The image size becomes a quarter of its original size. Then, three different scale feature maps are output after stacking four efficient receptive field ERFM modules and three down sampling MP modules. The second and third ERFM modules are embedded with the multi-channel coordinate attention mechanism MCCA to form the ERFM-Att module.

2.2.1. Efficient Receptive Field Module (ERFM)

Belonging to an integral component of the backbone network, the ERFM module influences the feature extraction proficiency of the entire network. Its detailed structure is shown in Figure 2, incorporating two branches N1 and N2. In target detection, when the receptive field of the convolutional kernel exceeds the feature region of the target to be tested, that part of the features cannot be extracted, and the target is recognized as background. When the receptive field cannot cover the feature region of the target to be tested, the global information will be ignored, and the features extracted by the network are limited to the local space. Dilated convolution is a convolution kernel with expansion rate. With no augmentation in convolutional parameters, the convolution kernel size and receptive field can be enlarged. Hence, the N1 branch constructs convolutional kernels with diverse scale receptive fields through different dilated rates, utilized to obtain multi-scale feature maps with large receptive fields, thereby the detection network can adapt to targets of different sizes. The N2 branch reduces the number of parameters brought by ordinary convolutional stacking by lightweight convolutional modules. Firstly, the input features are routed to the N1 branch, which encompasses three parallel dilated convolutional layers with distinct sampling rates. The dilated rates are 2, 4 and 8, respectively, and the equivalent receptive field of the current layer are 5, 9 and 17. The features extracted from each convolutional layer are subsequently processed in a separate branch and fused to culminate in the final output. Concurrently, the input features are routed to the N2 branch, which successively passes through the conventional convolutional module CBS, the partial convolution [32] (PConv) module PBS, and the depthwise separable convolution [33] (DSC) module DBS to comprehensively extract features. There exists a residual connection

between PBS and DBS, thus enabling the features to detour to the output layer. Therefore, the latter layer can directly learn the residual to protect the integrity of information, and simplify the learning complexity of the network. Finally, all the output feature maps of N1 and N2 branches are spliced according to the channels. Then the number of channels is reduced by dimensionality reduction through the convolution layer with convolution kernel of one to get the final output. After the above steps, the ERFM-Att module further extracts the features through the multi-channel attention module MCCA.

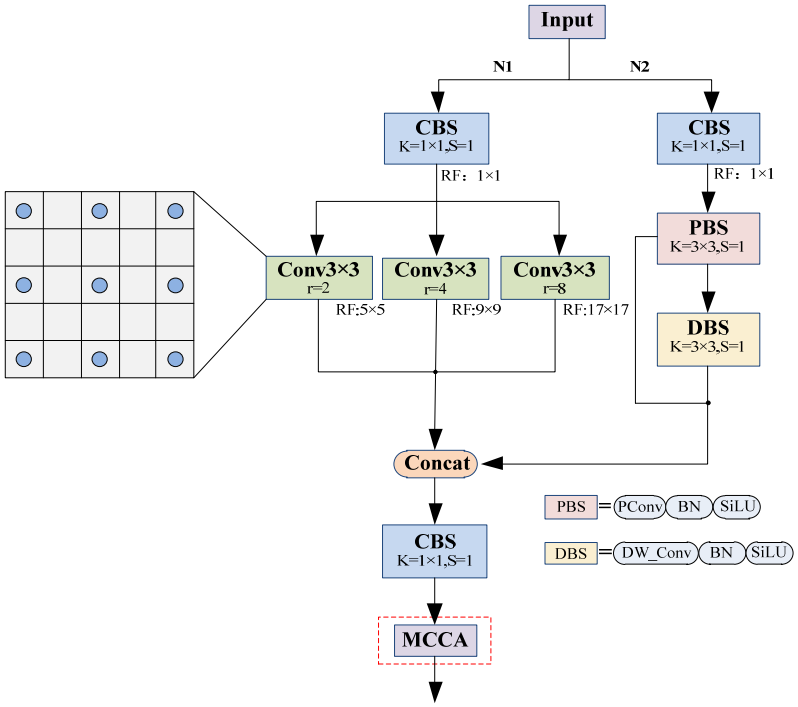


Figure 2. The structure of ERFM/ERFM-Att module.

The depthwise separable convolution consists of two segments, as shown in Figure 3. Initially, depthwise convolution is utilized to carry out operations with the input features. The implementation process is that one convolution kernel corresponds to one channel, that is, a channel only performs convolution operations on its corresponding convolution kernel. The quantity of convolution kernels and the count of output feature graph channels are congruent with the number of input feature graph channels. Subsequently, pointwise convolution is used to do operations with feature maps. Its operations are similar to conventional convolution operations, but its convolution kernel size is 1×1. Essentially, the feature maps of the previous step are weighted and combined in channel dimension to generate new feature maps, and the number of convolution kernels is identical to the number of channels in the output feature maps. But it may fluctuate from the number of input feature graph channels. Compared to conventional convolution, the computational volume ratio of DSC to conventional convolution is:

$$\frac{1}{N} + \frac{1}{D_F^2} \tag{1}$$

where D_F is the size of the convolution kernel of the depthwise convolution, and N is its number.

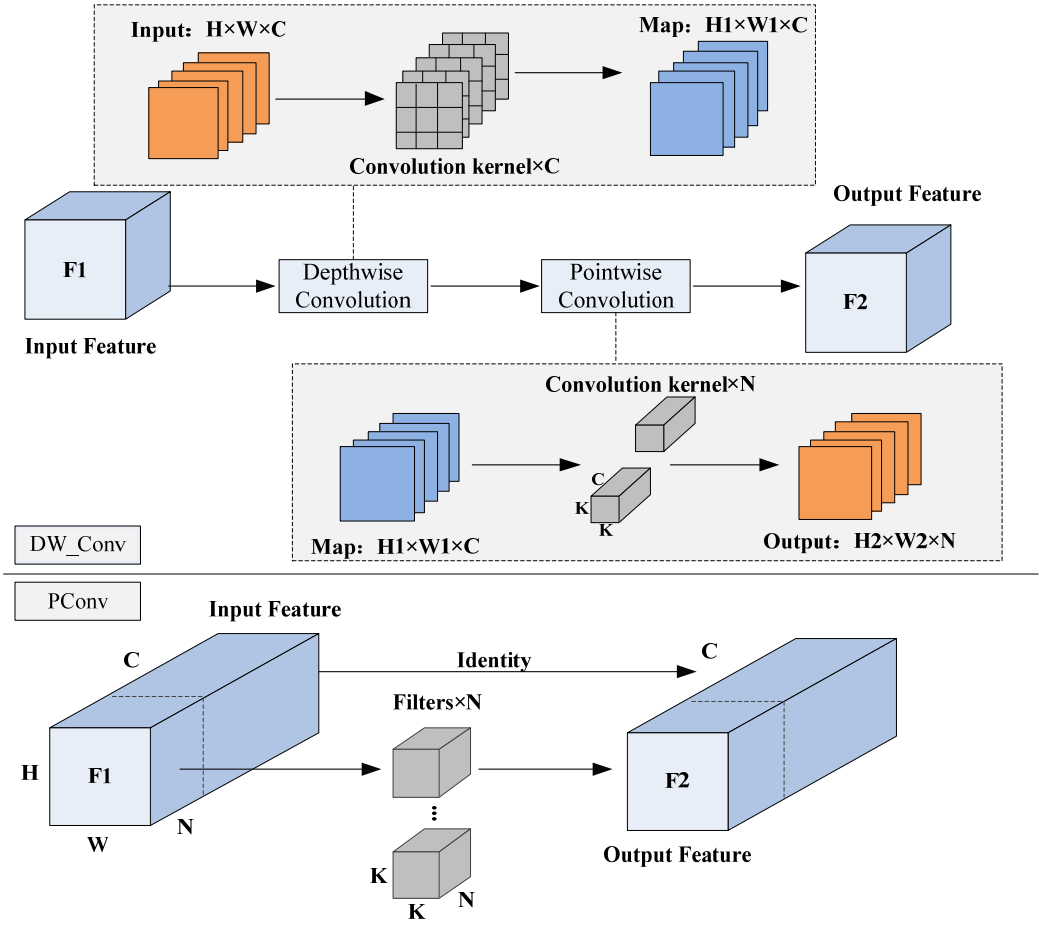


Figure 3. Depthwise separable convolution and partial convolution.

Partial convolution selects a portion of channels from the input feature map for convolution operation, and the remaining channel information is retained and merged with the new channels that have been computed, as shown in Figure 3. PConv usually considers the first or the last consecutive channel as representative of the whole feature map for calculation, which can avoid a large amount of information loss during the convolution process and retain the integrity of the channel information. Therefore, PConv can effectively decrease the model's parameter count and computational overhead by reducing the number of unnecessary computations and memory accesses, which makes the FLOPs only one sixteenth of the ordinary convolution and the memory accesses only one quarter of the ordinary convolution. In addition, the normalized BN layer and SiLU activation function are added to form the DBS and PBS convolution blocks, which can enhance the nonlinear representation of the network, maintain the diversity of features, and improve the generalization ability of the network.

2.2.2. Multi-Channel Coordinate Attention Module (MCCA)

Inspired by the fundamental coordinate attention mechanism [34], a multi-channel coordinate attention MCCA module is proposed with its detailed architecture illustrated in Figure 4. This module can learn to obtain the azimuth perception, concentrate on the significant coordinates, and disregard the invalid coordinates, thus effectively capturing features of ships of interest, suppressing the background clutter noise, and improving the efficiency of informational transmission. The basic coordinate attention module firstly calculates the average of the input features for every channel along the horizontal and vertical coordinate axes separately to yield a pair of direction-conscious attention maps,

which can not only capture the extended dependence along one spatial dimension, but also preserve the precise position information along the other spatial dimension.

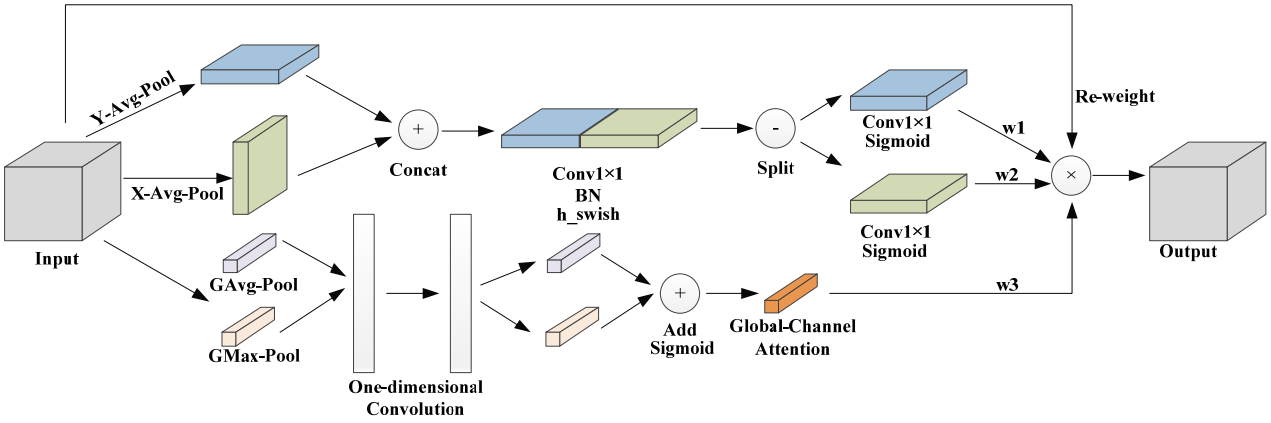


Figure 4. Multi-channel coordinate attention mechanism.

The direction-conscious attention can be expressed as:

$$X_h = YAvgPool(F_{in}) \tag{2}$$

$$X_w = XAvgPool(F_{in}) \tag{3}$$

where $AvgPool(\cdot)$ indicates average pooling based on vertical coordinates, $XAvgPool(\cdot)$ represents average pooling based on horizontal coordinates, and F_{in} denotes input features.

After that, the feature maps in two directions are cascades, and the channel dimension is reduced through convolution. Subsequently, the feature maps are divided back to the original two directions along the spatial dimension, restoring the number of channels to its original state through convolution. Finally, after sigmoid activation, the final coordinate attention weights are obtained, which can be expressed as:

$$W_1 = CS(Split_h(CBH(Concat(X_h, X_w)))) \tag{4}$$

$$W_2 = CS(Split_w(CBH(Concat(X_h, X_w)))) \tag{5}$$

$$h_{swish}(x) = x\frac{ReLU(x+3)}{6} \tag{6}$$

where $Concat(\cdot,\cdot)$ and $Split(\cdot)$ represent the join and the segmentation operation respectively. $CBH(\cdot)$ means that the input is convolved first, then normalized BN, and then passed through the h_swish activation function. $CS(\cdot)$ indicates the convolution and sigmoid activation function.

The integration of the coordinate attention mechanism allows the network to pay more attention to the spatial characteristics of features. However, the interaction information between the original channels of feature maps is also worth concentrating on. Therefore, another global channel attention branch is added. Firstly, it compresses the feature map in the global space dimension, including global average pooling and global max pooling. Then, after the one-dimensional convolution module, dimension reduction is avoided, and cross-channel interaction information is effectively captured. Finally, the output results are added and activated by sigmoid. The global channel weight can be expressed as:

$$W_3 = Sgm(Add(OConv(GAP(F_{in}), GMP(F_{in})))) \tag{7}$$

$$X_{out} = W_1 \times X_{in} + W_2 \times X_{in} + W_3 \times X_{in} \tag{8}$$

where $Sgm(\cdot)$ represents the sigmoid activation function, $Add(\cdot)$ indicates add by element operation, $OConv(\cdot)$ denotes one-dimensional convolution, $GAP(\cdot)$ and $GMP(\cdot)$ represent global average pooling and global maximum pooling respectively.

Firstly, the input features pass through the global channel attention submodule to obtain the feature map with the global channel interaction information weight, and then multiply with the coordinate attention weight. In this way, the features can be captured from multiple perspectives, and the model can focus on crucial information while discarding redundant data. Using the proposed ERFM module and MCCA attention mechanism to construct the backbone network can achieve the goal of extracting features efficiently without increasing the number of required parameters. It should be noted that among the four ERFM modules in the backbone network, only the middle two ERFM modules are embedded with the MCCA attention mechanism. The output features are passed into the subsequent feature fusion network to highlight the location information of the target and suppress the background noise.

2.3. Enhanced Hierarchical Feature Fusion Network (EHFNet)

As can be seen from Figure 1, the backbone network generates three feature graphs of different sizes C3, C4 and C5 as output, among which the low-level feature contains stronger location information and the high-level feature carry richer semantic information. In order to make full use of information at different levels, a hierarchical feature fusion network is needed to fuse different features. The enhanced multi-scale feature fusion network EHFNet is composed of two aggregation paths, top-down and bottom-up, which realizes bidirectional feature fusion, and fuses the feature graph of the previous path in the current path. Specifically, in the process of top-down feature fusion, the weighted node mechanism and feature reuse connection are introduced, which can efficiently aggregate important features and shorten the information path at the same level. In the process of bottom-up feature fusion, an improved down sampling module is used to avoid the loss of important target information, especially small target features. The specific feature flow diagram of EHFNet is shown in Figure 5.

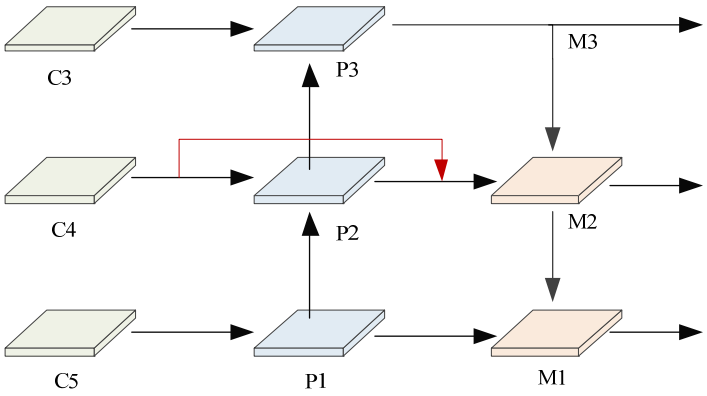


Figure 5. The feature flow diagram of EHFNet, where the red arrows represent jump connections.

2.3.1. Weighted Feature Fusion Nodes and Feature Multiplexing Connection

In order to facilitate subsequent classification and prediction, the fusion of features at different scales usually results in the direct and simple corresponding addition of different features. However, this will lead to the non-discriminative fusion of texture information at the lower level and semantic information at the higher level, ignoring the importance of information at different levels. Therefore, it is necessary to introduce a weighted feature fusion mechanism. This allows the model to learn the importance of input features at

different levels and to differentiate between input features at different resolutions. The weighted fusion mechanism used in this paper is as follows:

$$O = \sum_i \frac{w_i * I_i}{\in + \sum_j w_j} \tag{9}$$

where I_i is different input features, w_i is its corresponding weight, $\in$ is a small positive value to avoid the situation where the denominator is 0. Using this method, the output range can be reduced to between [0, 1], and the training speed is fast. At the same time, to reduce the amount of computation, EHFNet only adds a weighting mechanism in the top-down fusion process.

The sizes of the three different feature maps output by the backbone network are C3: 20 × 20, C4: 40 × 40, and C5: 80 × 80, respectively. Since the feature map of the middle layer, 40 × 40, is generated by the sub-sampling of the low-level feature map and the source of the generation of the high-level feature map, which contains more contextual information, we choose to add layer skipping connections to it, as shown in Figure 6. By making it skip the intermediate information path and directly fuse with the subsequent feature maps, a fast channel between layers of the same size is established, and the flow of multi-scale information is enhanced. In addition, the skip layer connection can recover part of the information lost in the subsequent down sampling process, which is beneficial to further improving the cross-scale detection performance of the model. After a series of operations, the processed output features contain a higher concentration of low-level local spatial and high-level global semantic information.

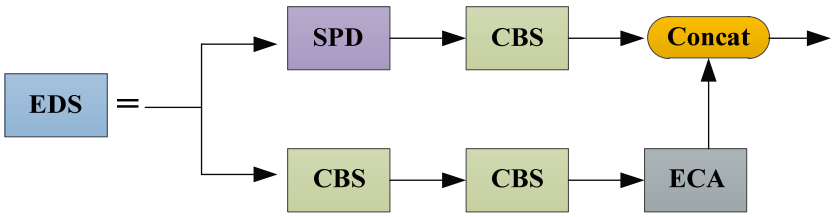


Figure 6. The structure of the efficient down sampling module.

2.3.2. Efficient down Sampling Module (EDS)

In the process of bottom-up feature fusion, two down sampling operations are performed, and the spatial information contained in the feature map decreases as it shrinks. When the resolution of the SAR image is too low or the detection of the target pixel is small, the fine-grained information will be lost and the model learning features will be insufficient. However, SAR images not only have much speckle noise, but also most ship targets occupy only a small part of the pixel value relative to the background. To avoid the loss of spatial position information of small targets during down sampling, a better down sampling module is proposed. The specific structure is shown in Figure 6.

As the general pooling layer is based on the pixel value in the local neighborhood for the overall calculation of subsampling operation, it will inevitably lose a lot of information in the feature map. To solve this problem, the SPD module is used instead of the general pooling to verify the current subsampling function. Inspired by image conversion technology, the SPD layer subsamples the input feature map, divides it into four different feature blocks with the same number of channels according to the length and width directions of the input feature, and splits it with the channel as the basis, so that the length and width of the original input feature map is scaled to one half, but the number of channels is increased to four times. In this way, all information in the channel dimension can be retained, so no information is lost, as shown in Figure 7. Then, a convolution module CBS without step size and with convolution kernel size of one is used to reduce the number of channels, reduce the number of channels increased by SPD layer, and carry out information fusion between channels. In addition, the original input feature map is passed through another

branch consisting of two convolutional modules, CBS, and the Efficient Channel Attention (ECA) mechanism, and the output information is fused with the output information passing through the SPD branch. The principle of ECA is shown in Figure 8. Input features are first homogenized globally to obtain one-dimensional features with the same size as the number of channels, and then one-dimensional convolution is used to capture cross-channel interaction information. In this way, not only is the resolution of down sampling guaranteed, but also the importance of features of different channels is preserved.

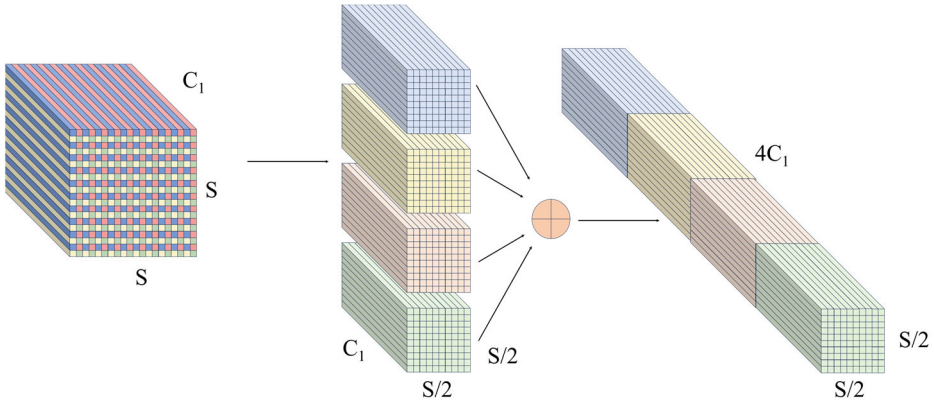


Figure 7. The structure of the SPD module.

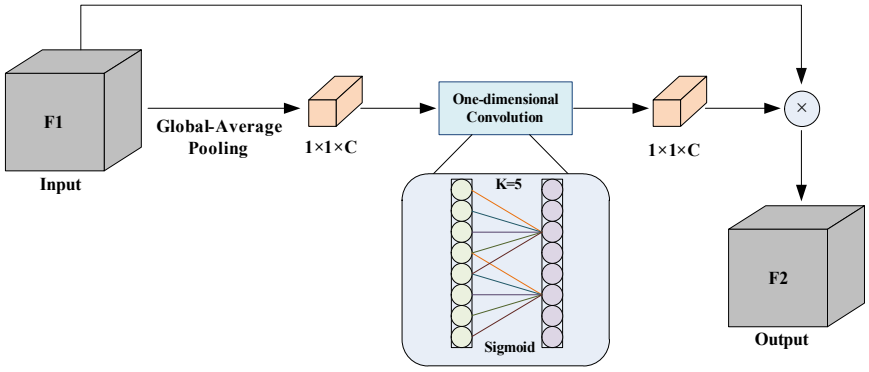


Figure 8. The ECA attention mechanism.

2.4. Bounding Box Loss Function

The loss function of bounding box regression (BBR) directly affects the localization performance of the model. The BBR loss function used in the original YOLOv7 model is the CIoU loss function. It comprehensively considers the overlap degree, center point distance, and aspect ratio difference between the predicted frame and the real frame. The formula is as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{10}$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}\right)^2 \tag{11}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{12}$$

where $\rho^2\left(b, b^{gt}\right)$ is the distance between the predicted box's center point and the actual box's center point, c is the diagonal distance of the smallest rectangular box containing two boundary boxes, v is utilized to assess the degree of alignment between the aspect ratio of the prediction box and the target box, and α is a balance factor that increases with IoU.

On the basis of CIoU, EIoU treats height and width as penalty terms respectively to avoid the limitations brought by fixed aspect ratio. The formula is as follows:

$$L_{EIoU} = L_{IoU} + L_{dis} + L_{asp} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{c_w^2} + \frac{\rho^2(h, h^{gt})}{c_h^2} \tag{13}$$

where c_w and c_h are the minimum width and height of the outer frame that encloses the two bounding boxes, w and h are the width and height of the predicted box, and w_{gt} and h_{gt} are the width and height of the real box.

Both CIoU and EIoU do not calculate the directional error between the true and predicted boxes, which leads to slower convergence of the model and deterioration of the results. SIoU adds the angle penalty metric, so that the predicted box and the true box can quickly align in the center, and only need to return to one coordinate, which greatly reduces the convergence time. Combined with the distance penalty term and the shape penalty term, the formula is as follows:

$$L_{SIoU} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{14}$$

$$\Delta = \sum_{t=x,y} \left(1 - e^{-\gamma \rho_t}\right), \ \rho_{x,y} = \left(\frac{b_{c_{x,y}}^{gt} - b_{c_{x,y}}}{c_{w,h}}\right)^2, \ \gamma = 2 - \Lambda \tag{15}$$

$$\Lambda = 1 - 2 * sin^2\left(\arcsin(x) - \frac{\pi}{4}\right), \ x = \frac{c_h}{\sigma} = \sin(\alpha) \tag{16}$$

$$\Omega = \sum_{t=w,h} \left(1 - e^{-w_t}\right)^{\theta} \tag{17}$$

where Λ is the angle penalty term, Δ is the distance penalty term, Ω is the shape penalty term, w_t Indicates the width/height difference of the two frames and the ratio of the maximum width/height of the two frames, α is the minimum angle between the central point of the two boxes and the X-Y axis, $\rho_{x,y}$ is the distance between the predicted box's center point and the actual box's center point.

Considering the existence of ship targets of different scales and the phenomenon of dense array and partial occlusion in SAR images, a new boundary box loss function WN-Loss is designed to improve the positioning accuracy of the model. WN-Loss is composed of Wise-IoU [35] and normalized Wasserstein distance (NWD) [36] measures. Wise-IoU defines a dynamic FM β by estimating the outlier value of the anchor box, and then assigns different gradient gains, so that BBR can focus on the anchor box with ordinary mass. Wise-IoU is divided into WIoUv1, WIoUv2, WIoUv3 versions, WIoUv1 version of the formula is as follows:

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{18}$$

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{\left(W_g^2 + H_g^2\right)^*}\right) \tag{19}$$

where W_g and H_g are the size of the minimum closed box. WIoUv2 uses Focal loss for reference to introduce the monotonic FM mechanism, effectively reducing the contribution of ordinary samples to the loss value, and setting the Focusing coefficient γ. The calculation formula is as follows:

$$L_{WIoUv2} = L_{IoU}^{\gamma *} L_{WIoUv1}, \ \gamma > 0 \tag{20}$$

$$L_{WIoUv2} = \left(\frac{L_{IoU}^*}{\overline{L_{IoU}}}\right)^{\gamma} L_{WIoUv1} \tag{21}$$

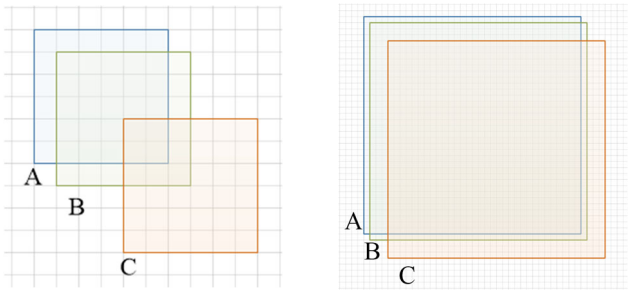
On the basis of the above WIoUv2 version, the WIoUv3 version introduces dynamic non-monotonic FM; the formula is as follows:

$$\beta = \frac{L_{IoU}^{*}}{\overline{L_{IoU}}} \in [0, +\infty) \tag{22}$$

$$L_{WIoUv3} = rL_{WIoUv1},\ r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \tag{23}$$

where β is the outlier of the anchor frame, and anchor frame mass increases as β decreases. When a smaller gradient gain is assigned to an anchor frame with a large outlier, it will effectively prevent large harmful gradients from low-quality samples.

In BBR, a common occurrence is to use IoU as the regression loss of the detect head, but the sensitivity of IoU to targets of different scales varies greatly. As shown in the Figure 9, assuming that the IoU threshold is 0.5, there are targets of different scales in one graph at the same time, and the position deviation of the targets is the same. For a small target of 6 × 6 pixels, a very small position deviation will also cause the IoU to drop from 0.53 to 0.06, which is a significant decrease, so that the target sample does not meet the threshold conditions and is incorrectly assigned as a negative sample. For a larger object of 36 × 36 pixels, the IoU decreases from 0.90 to 0.65, which is still within the threshold range.



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

Figure 9. Sensitivity of IoU to different scale targets. A is the reference box, B is a box offset by one pixel, and C is a box offset by 4 pixels: (**a**) small scale target; (**b**) large scale target.

However, the size of ship targets have multiple scales, and using only IoU measurement will cause the problem of unbalanced sample distribution. In order to alleviate the above problems, normalized Wasserstein distance (NWD) loss is introduced as the BBR loss measurement. The NWD metric replaces the standard IoU by the Wasserstein distance. Specifically, the bounding box is first modeled as a two-dimensional Gaussian distribution, and then the similarity of the Gaussian distribution of the bounding box is measured using the normalized Wasserstein distance. The primary benefit of the Wasserstein distance is that it allows for the measurement of distribution similarity even when the actual box and the predicted box do not overlap or when the degree of overlap is minimal. Moreover, NWD is not impacted by objects of different scales, making it more applicable for measuring the similarity between small ships. The NWD loss calculation formula is as follows:

$$L_{NWD} = 1 - NWD(N_p, N_g) \tag{24}$$

$$NWD(N_p, N_g) = \exp\left(-\frac{\sqrt{W_2^2(N_p, N_g)}}{C}\right) \tag{25}$$

$$W_2^2(N_p, N_g) = \| \left([cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2}]^T, [cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2}]^T\right) \|_2^2 \tag{26}$$

where $\| \cdot \|$ is the Frobenius norm, $W_2^2(N_p, N_g)$ is the Wasserstein distance between the real box and the predicted box, N_p and N_g are the corresponding two-dimensional Gaussian distributions, respectively, and $NWD(N_p, N_g)$ is the new metric normalized using the

exponential form. The final BBR loss function is composed of Wise-IoU and NWD loss multiplied by the corresponding weight coefficient respectively. The formula is as follows:

$$Loss_{bbr} = \lambda_1 L_{NWD} + \lambda_2 L_{WIoU} \tag{27}$$

When λ_1 is 1 and λ_2 is 0, it means that only Wise-IoU is utilized to the bounding frame loss function; when λ_1 is 0 and λ_2 is 1, it means that only NWD loss measure is utilized to the bounding frame loss function. In the experiment of this paper, λ_1 is 0.7 and λ_2 is 0.3.

3. Experiment and Results

3.1. Experimental Platform

The experimental environment relied on PyCharm 2022.3.3 IDE to build, and the deep learning framework consisted of PyTorch 2.0.1, CUDA 11.8 and CUDNN 8.9. The CPU was 12th Gen Intel(R) Core(TM) i7-12700 2.10GHz and the GPU was NVIDIA GeForce GTX 3060 12G. The PC operating system was Windows 11.

3.2. Datasets

(1) SSDD:

SSDD is a public ship dataset published by Li et al. [37]. It contains 1160 SAR images, including HH, HV, VV and other polarization modes, with a resolution between 1–15 m. A total of 2456 ships were built, the smallest being 7 × 7 and the largest 211 × 298, with a ratio of 60.2%, 36.8%, and 3% of small, medium, and large ships, respectively. There are multiple offshore simple scenes and offshore complex scenes in the dataset. In addition, the dataset provides annotation information in PASCALVOC format.

(2) HRSID:

The HRSID dataset [38] was released in 2020 and widely used. The original image was 136 large-scale SAR satellite images, which were then cropped into 5604 SAR images with a size of 800 × 800. A total of 16,951 ships were built, with a ratio of 54.5%, 43.5%, and 2% of small, medium, and large ships, respectively. And its image resolutions ranged from 0.5 m to 3 m. There are multiple offshore simple scenes and offshore complex scenes in the dataset. In addition, the dataset provides annotation information in MS COCO format.

(3) LS-SSDD:

The LS-SSDD dataset [39] is a large-scale SAR ship detection dataset published by Zhang et al. in 2020. It consists of 15 large 24,000 × 16,000 images from Sentinel-1, including both VV and VH polarization modes, which were then trimmed to 9000 smaller 800 × 800 images. There are a total of 6015 ships, with the ratio of small, medium, and large ships being 99.80%, 0.20%, and 0, respectively. In addition, the dataset provides annotation information in PASCALVOC format.

3.3. Model Evaluation

In this experiment, target detection evaluation indexes were used to evaluate the model performance, including Precision, Recall, Average Precision (AP), small target accuracy AP_S, medium target accuracy AP_M, large target accuracy AP_L, and F1 score. Accuracy and recall rates are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{28}$$

$$Recall = \frac{TP}{TP + FN} \tag{29}$$

TP, *FP*, and *FN* refer to the number of correctly detected vessels, false alarms, and missed vessels, respectively.

AP is the area surrounded by the curve of the relationship between precision rate and recall rate, which is defined as:

$$AP = \int_0^1 P(R)dR \tag{30}$$

F1 scores combine accuracy and recall as follows:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{31}$$

According to the number of pixels contained in the ship target position prediction box, ships are divided into small ships, medium ships, and large ships according to the COCO index definition to calculate their detection accuracy. Some COCO index definitions are shown in Table 1.

Table 1. The definition of some COCO indicators.

Metric	Meaning
AP	AP for IoU = 0.50:0.05:0.95
AP50	AP for IoU = 0.50
AP_S	AP for small targets (area < 32^2)
AP_M	AP for medium targets (32^2 < area < 96^2)
AP_L	AP for large targets (area > 96^2)
FPS	Frames per second

3.4. Experimental Results

3.4.1. Ablation Experiment

The ablation experiment was conducted on the SSDD dataset. To verify the validity of each module in the model and observe the detection performance of the CSEF-Net model, F1 score, mAP, and different scale accuracy indexes AP_S, AP_M, AP_L were used as evaluation indexes of the ablation experiment. All experiments adopted the same parameter settings, and the parameter values refer to the official YOLOv7 training settings. Specifically, epoch was set to 300, batch_size was set to 8, initial learning rate was set to 0.01, final OneCycleLR learning rate was set to 0.1, optimizer weight decay was set to 0.0005, SGD momentum was set to 0.937, and the performance data were test set results. The specific ablation experiment results are shown in Table 2.

Table 2. Results of ablation experiments.

Experiment	ERFBNet	EHFNet	WN-Loss	F1	mAP	AP_S	AP_M	AP_L
1	—	—	—	0.848	0.902	0.561	0.54	0.047
2	√	—	—	0.920	0.961	0.636	0.673	0.41
3	√	—	√	0.933	0.968	0.657	0.667	0.413
4	√	√	—	0.928	0.963	0.639	0.647	0.357
5	—	√	√	0.888	0.941	0.617	0.601	0.183
6	√	√	√	0.941	0.973	0.653	0.702	0.41

"√" indicates that the current module was used, and "—" indicates that it was not used.

Different module combinations were used in experiments 1–6. The baseline model was built in experiment 1, and the CSEF-Net model was built in experiment 6. It can be seen that the mAP of baseline model on SSDD dataset is 90.2%, F1 is 0.848, but AP_L is only 4.7%, indicating that the cross-scale detection ability is weak. In experiment 2, ERFBNet was integrated as the backbone network, and all indexes were improved compared with the baseline model. In experiment 3, after WN-Loss was added, only AP_M decreased slightly and AP_S reached the highest value of 65.7%, indicating that WN-Loss had a good effect on small target detection. Experiment 4 added EHFNet on the basis of experiment 2, the F1 and mAP were slightly improved while AP_M and AP_L were slightly reduced. The model

detection performance of experiment 5 was the worst except for the baseline model, indicating that after removing ERFBNet module, the feature extraction ability of the backbone network was significantly diminished leading to a poor detection effect. All modules were used in experiment 6, and a good balance effect was achieved in various indexes. Compared with the baseline model, mAP increased from 90.2% to 97.3%, F1 reached the highest value of 0.941, and the detection accuracy of ship targets of different scales was significantly improved, but AP_S and AP_L decreased by 0.4% and 0.3% compared with experiment 3.

The PR curve shows the relationship between the precision and recall of the model in training, and usually the closer the curve to the upper right the better performance of the model. Figure 10 demonstrates the comparison of PR curves for ablation experiments on the SSDD dataset. As can be seen from the figure, CSEF-Net performs best with the largest curve enclosure area relative to the baseline model YOLOv7.

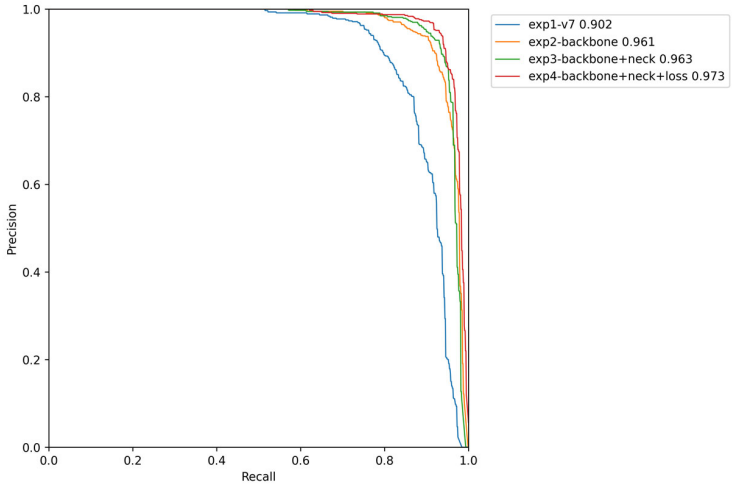


Figure 10. PR curves for the ablation experiments.

On the basis of experiment 4, we experimented with the weighting coefficients in WN-Loss to obtain the optimal settings and the results are shown in Table 3. After analyzing the experimental results, λ_1 was set to 0.7 and λ_2 was set to 0.3. We also conducted comparative experiments on the different boundary box loss functions mentioned in Section 2.4, including the performance results of three different versions of Wise-IOU without NWD measurement, and the performance results of the combination of NWD measurement and different IOU, as shown in Table 4. It can be seen that WN-Loss, composed of the NWD measure and Wise-IOU v2 version, performs best, with a mAP of 97.3%, and has the highest accuracy rate on small and medium ship targets.

Table 3. Experiments with different weighting coefficients for WN-Loss.

λ_1	λ_2	P	R	mAP
0.9	0.1	0.913	0.883	0.944
0.7	0.3	0.921	0.895	0.952
0.5	0.5	0.921	0.879	0.947

Table 4. Experimental results of different boundary box loss functions.

$Loss_{bbr}$		F1	mAP	AP_S	AP_M	AP_L
Wise-IOU	V1	0.9283	0.963	0.651	0.692	0.404
	V2	0.9348	0.974	0.640	0.680	0.310
	V3	0.9281	0.961	0.645	0.660	0.107
NWD	+CIOU	0.9078	0.952	0.632	0.673	0.503
	+EIOU	0.9081	0.951	0.632	0.616	0.177
	+SIOU	0.9145	0.961	0.642	0.665	0.369
	+V2	0.941	0.973	0.653	0.702	0.41

3.4.2. Comparison with Other Target Detection Algorithms

To further validate the performance of the model algorithm proposed in this paper, a comparison experiment was conducted on SSDD dataset and HRSID dataset under the same conditions with other detection algorithms. In the experiment, seven classical target detection algorithms were selected for comparison, these are the two-stage detection algorithm Faster R-CNN [12], the one-stage detection algorithms YOLOv5_n, YOLOv7 [31], SSD [19] and EfficientDet [40], and the anchor-free detection algorithms RetinaNet [41] and CenterNet [42]. In addition, five other SAR ship detection methods were also selected as comparative. They are I-YOLOv5 [43], Pow-FAN [44], Quad-FPN [29], BL-Net [30], and I-YOLOx-tiny [25], where Quad-FPN, BL-Net, I-YOLOx-tiny are multi-scale detection networks. The specific experimental results are shown in Table 5.

Table 5. Experimental results of comparative experiments.

Method	SSDD				HRSID				FPS	Params (M)	FLOPs (G)
	P	R	F1	mAP	P	R	F1	mAP			
Faster-RCNN	0.502	0.944	0.66	0.851	0.378	0.560	0.45	0.454	13	41.3	251.4
SSD	0.936	0.552	0.69	0.899	0.928	0.438	0.60	0.681	92	23.7	30.4
EfficientDet	0.959	0.533	0.69	0.713	0.969	0.331	0.49	0.484	29	3.8	2.3
YOLOv5_n	0.925	0.833	0.87	0.897	0.890	0.717	0.79	0.776	95	1.9	4.5
YOLOv7	0.928	0.782	0.84	0.902	0.847	0.724	0.78	0.819	56	37.1	105.1
RetinaNet	0.976	0.623	0.76	0.698	0.980	0.395	0.56	0.534	34	36.3	10.1
CenterNet	0.948	0.604	0.74	0.785	0.948	0.696	0.80	0.788	48	32.6	6.7
I-YOLOv5	0.883	0.934	0.90	0.950	0.843	0.845	0.84	0.851	13	-	-
Pow-FAN	0.946	0.965	0.95	0.963	0.885	0.837	0.86	0.897	31	136	-
Quad-FPN	0.895	0.957	0.92	0.952	0.879	0.872	0.87	0.861	11	-	-
BL-Net	0.912	0.961	0.93	0.952	0.915	0.897	0.90	0.886	5	47.8	417.8
I-YOLOx-tiny	0.960	0.930	0.94	0.961	0.936	-	-	0.867	49	1.4	5.7
CSEF-Net	0.967	0.918	0.94	0.973	0.927	0.801	0.85	0.906	43	37.3	104.1

The experimental results on SSDD and HRSID show that compared with the other seven classical methods, although the precision and recall of CSEF-Net are not the highest among all algorithms, the mAP is best. Compared with the baseline YOLOv7, the mAP of CSEF-Net is improved by 7.1% on the SSDD dataset and 8.7% on the HRSID dataset. It proved to be effective in ship detection using SAR images. The performance of the Faster-RCNN network is the worst, because the two-stage detection network first generates the preselected frame, and there is a lot of speckle noise in SAR images, which affects the accurate generation of the preselected frame. In addition, the FPS of the two-stage algorithm is also the lowest. Among the single-stage detection algorithms, SSD and EfficientDet have high precision, but the recall rate is not satisfactory, even below 50% on the HRSID dataset, and the F1 score of the YOLO series network indicates a better balance of precision and recall, and a higher mAP. In the anchor-free algorithm, there is no prior anchor box generation and only one box is predicted for each location, which may cause some overlapping or obscured areas to be undetected, so the recall rate and mAP of RetinaNet and CenterNet decreases. Compared to the five other methods, CSEF-Net has the highest mAP on both the SSDD and HRSID datasets. And it had the highest average accuracy on the HRSID dataset at 90.6%. What is more, the algorithms perform worse on the HRSID dataset than the SSDD dataset. The HRSID dataset contains more complex scenarios. For example, ships in ports are closely arranged, multiple small targets enter ports at the same time, and the cross-scale difference of ships is large, which makes the network performance challenging. However, CSEF-Net still performs better than other algorithms, and it can be seen that CSEF-Net can extract cross-scale features of ships as well as possible even in the case of more noise.

To evaluate the complexity and detection speed of the model, we used Params, FLOPs and FPS metrics. As shown in Table 5, the FPS, params, and FLOPs of CSEF-Net are 43, 37.3, and 104.1, respectively. This indicates that even though CSEF-Net achieves better accuracy at the cost of increasing complexity by introducing the attention mechanism and

adding additional parameters, it can still meet the requirements of real-time detection. However, the number of parameters and the number of floating-point operations of the model can be decreased, and we will consider further optimizing the complexity of the model in subsequent work.

3.4.3. Experimental Results in Different Scenarios

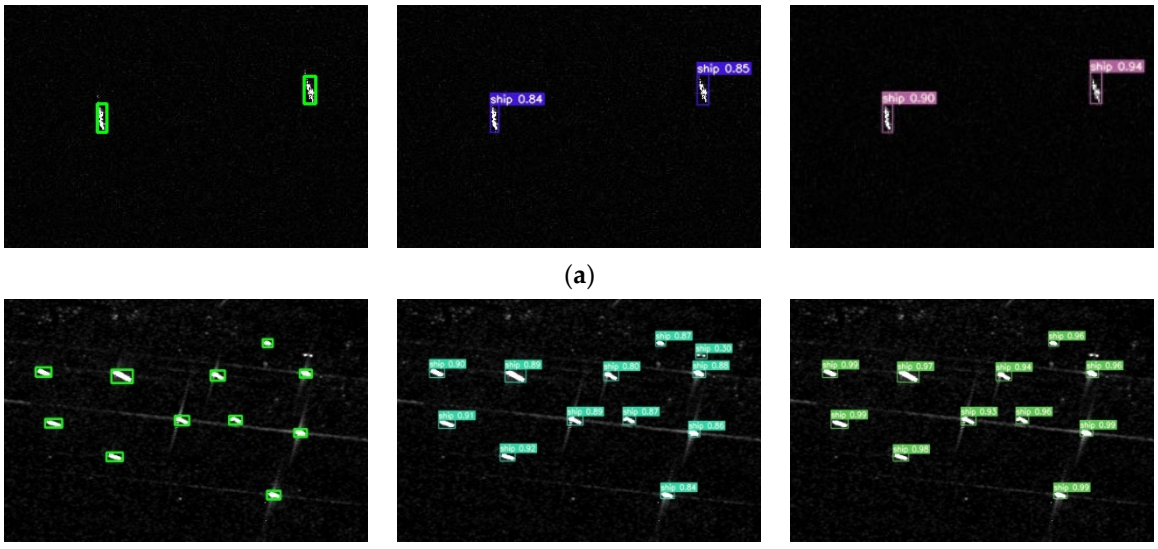
To validate the performance of the models in different scenes, the test set images of the SSDD dataset were divided into offshore and inshore scenes. In the offshore scene, ship targets are mostly located in the sea without interference from other objects, and the background is simple and easy to detect. In the inshore scene, ship targets are mostly docked at ports or coastlines, surrounded by interference from other objects, and the background is complex and difficult to detect. The CSEF-Net model was compared with the YOLOv7 model, and the specific performance results are shown in Table 6.

Table 6. Experimental results of different scenes.

Scene	Method	F1	mAP	AP_S	AP_M	AP_L
Offshore	YOLOv7	0.929	0.977	0.604	0.664	0.190
	CSEF-Net	0.982	0.993	0.672	0.767	0.700
Inshore	YOLOv7	0.647	0.683	0.443	0.327	0.026
	CSEF-Net	0.850	0.9	0.604	0.587	0.074

It can be seen that the performance of the CSEF-Net model in the offshore scene is better, F1 is 0.982, mAP is 99.3%, and the precision of various scales is relatively balanced. In the nearshore scenario, the model performance decreased somewhat, but it was still improved compared with the YOLOv7 model. F1 was 0.85 and mAP was 90%, indicating good robustness. However, AP_M and AP_L declined more than AP_S, this was especially true of AP_L, which was only 7.4%. This shows that the model loses more information to medium and large ships in complex scenes, and retains better information to small targets. The anti-jamming ability of the model can be further improved.

In order to visually demonstrate the detection effect of the proposed algorithm in different scenarios, six typical representative scenarios were selected from the test sets of the SSDD dataset and the HRSID dataset, respectively, for experimental comparison, as shown in Figures 11 and 12.



(**a**)

(**b**)

Figure 11. *Cont.*

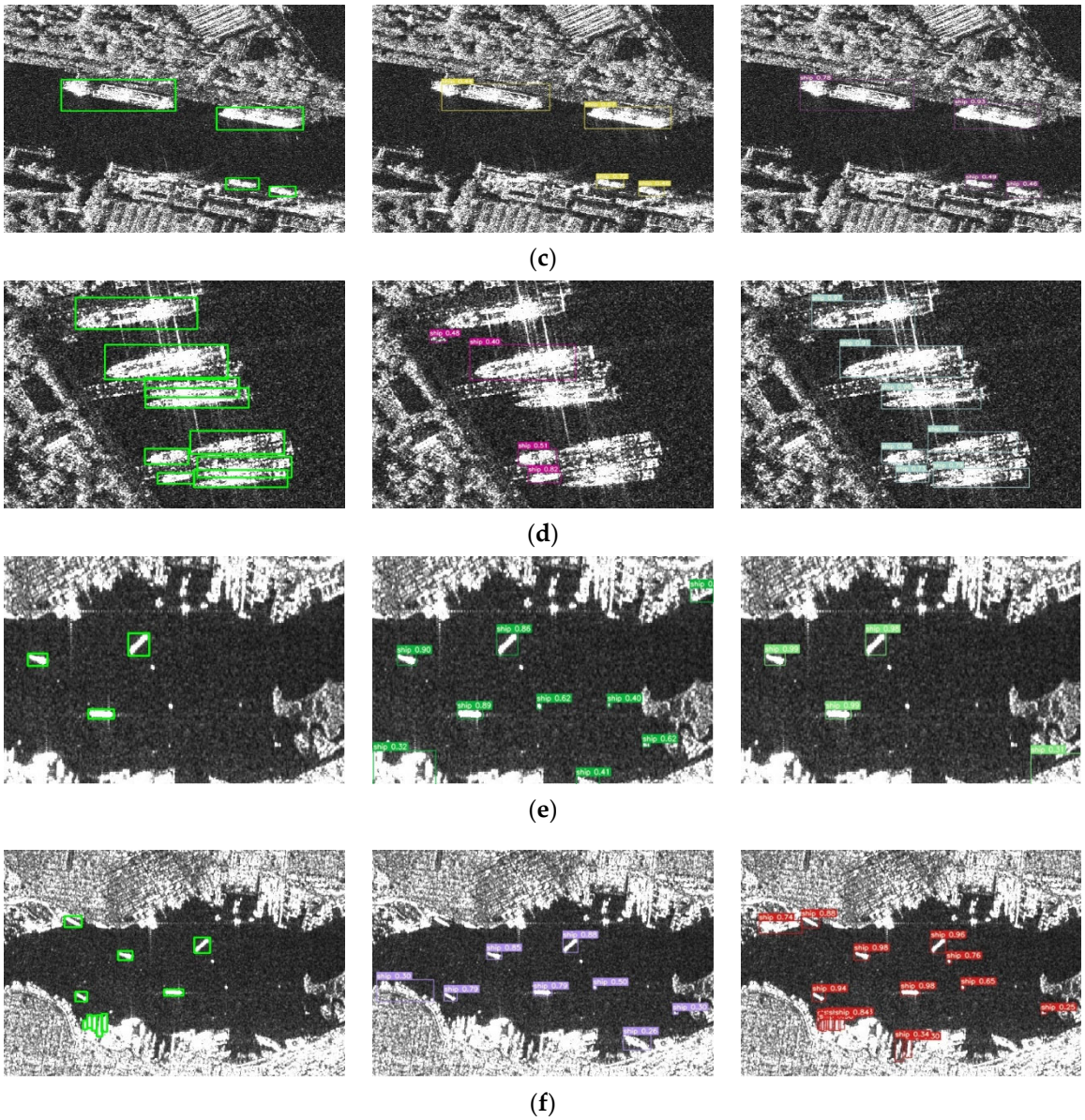


Figure 11. Detection results of typical scenarios in the SSDD dataset. The bounding box in the figure shows the real labels in the first column, the original YOLOv7 detection results in the second column, and the CSEF-Net detection results in the third column: (**a**) a small number of discrete small targets; (**b**) multiple discrete small targets; (**c**) targets that are more dispersed when docked; (**d**) overlapping targets that are distributed when docked; (**e**) small targets with discrete distribution in a complex background during port entry; (**f**) small targets with dense distribution in a complex background when entering port.

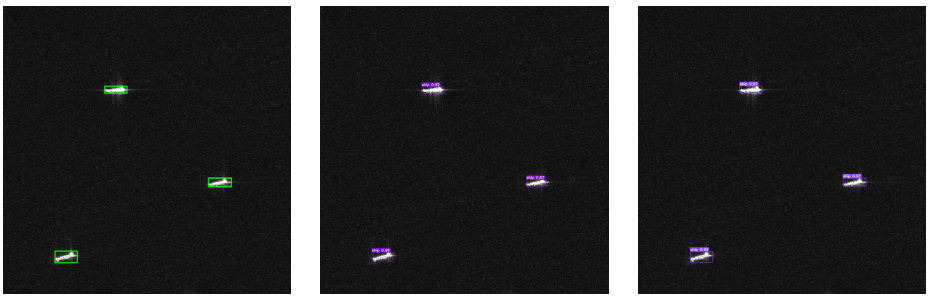


Figure 12. *Cont.*

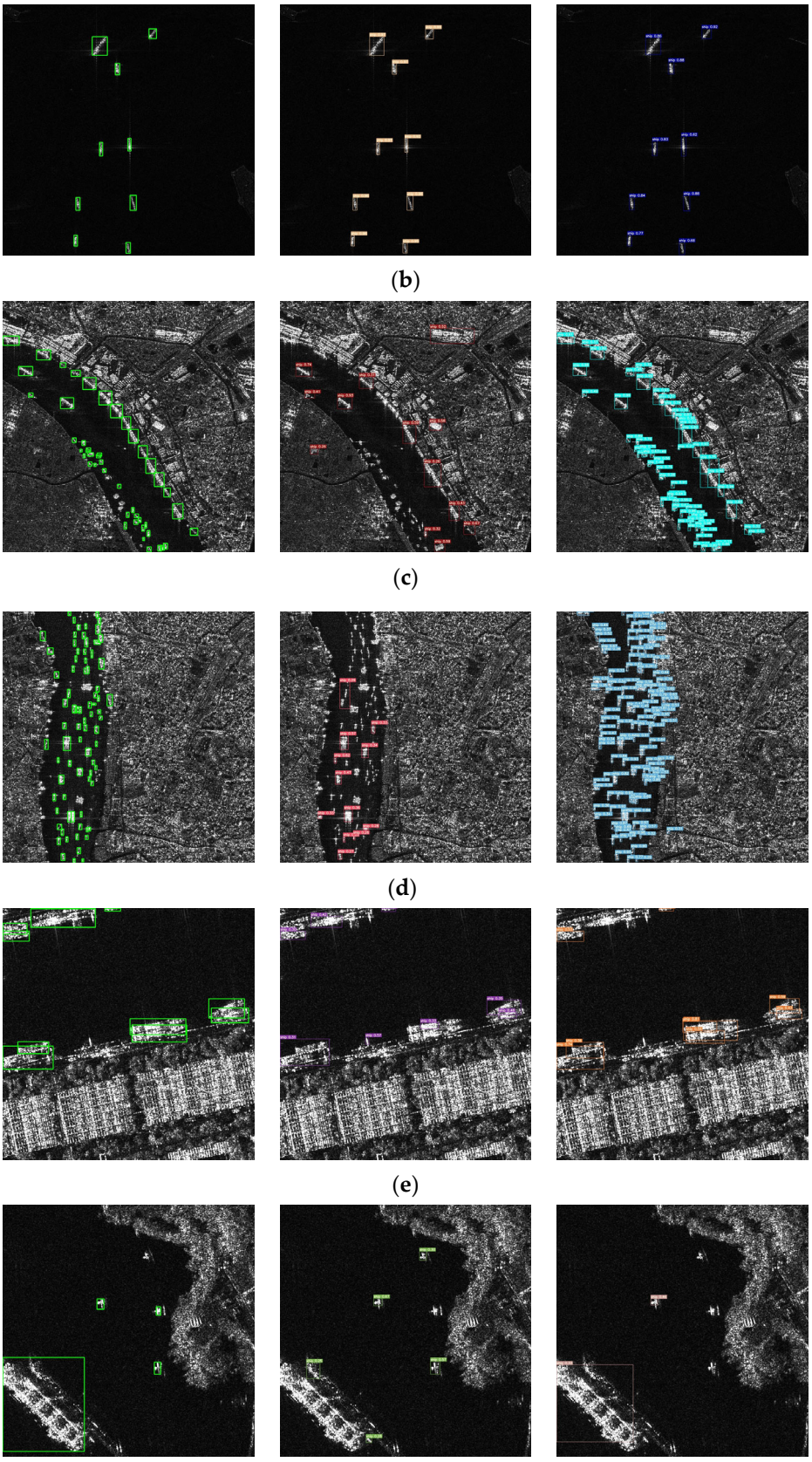


Figure 12. Detection results of typical scenarios in the HRSID dataset: (**a**) a small number of discrete small targets; (**b**) multiple discrete small targets; (**c**) multiple cross-scale targets in a dense array while docked at shore; (**d**) multiple small targets with dense distribution with a complex background when entering port; (**e**) the real box has a lot of overlap when docked on the shore; (**f**) a situation where the size of a large ship differs greatly from that of a small ship.

As can be seen from Figure 11a,b and Figure 12a,b, in the offshore scenario, the background is simple, the ship target is easy to detect, and the algorithm model in this paper has higher confidence for the same detection target. However, in the nearshore scenario, ships will be disturbed by complex shore buildings when docked or entering the port, and there are dense arrangements of ship targets of different scales at the same time, which makes detection more difficult. As can be seen from Figure 11d,f and Figure 12c,d,e, the detection performance of the YOLOv7 model is not good in the case of dense or overlapping ship targets, and the miss and error detection rate of the proposed algorithm model is lower, which indicates that the proposed algorithm model can better distinguish and extract ship features, and has strong anti-interference performance. As can be seen from the result diagram in Figure 11c, when the cross-scale difference of ship targets in the image is small, the model can still detect targets of different scales well; however, when there are targets with large cross-scale differences at the same time, as shown in Figure 12f, the YOLOv7 model fails to detect the large-scale ship in the lower left corner, resulting in an error detection situation. Although there are some omissions in the algorithm model, two different scales are correctly detected.

3.4.4. CSEF-Net's Performance in Other SAR Images

To verify the robustness and generalization of the model in other scenarios, we tested the performance of the CSEF-Net by conducting experiments directly on the untrained public dataset SAR-Ship-Dataset [45]. It is used for SAR ship detection in complex backgrounds with 39,729 images including 59,535 ship examples. Ultimately, CSEF-Net has a precision of 75.7%, a recall of 68.3%, and an average precision of 70.2% on this dataset. Figure 13 shows the partial visualization of test results. From these results, it can be seen that CSEF-Net has a small number of omissions and misdetections, and the generalization performance of the model can be further improved.

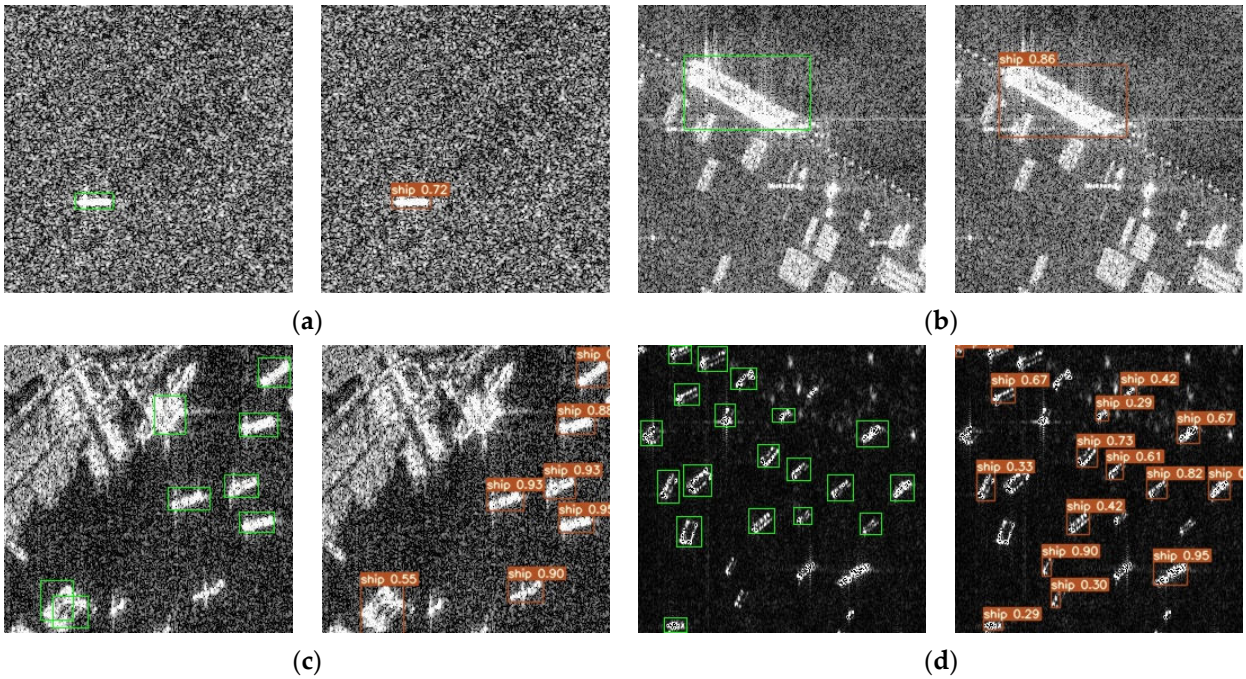


Figure 13. Detection results of typical scenarios in the SAR-Ship-Dataset. The image on the left is the real label and on the right is the detection result of CSEF-Net: (**a**) mass noise; (**b**) large scale ship targets; (**c**) nearshore interference; (**d**) large number of discrete targets.

In addition, we also selected a large-scale image from the Sentinel-1 satellite in the untrained LS-SSDD dataset and conducted experiments. The size of the original image is 24,000 × 16,000, limited by the hardware of the computer, the image size was compressed

to 8000 × 8000 when testing the whole image, and the segmented sub-images were tested separately. The detection results are shown in Figure 14. It shows that although some very small targets are missed, most ship targets can still be accurately detected.

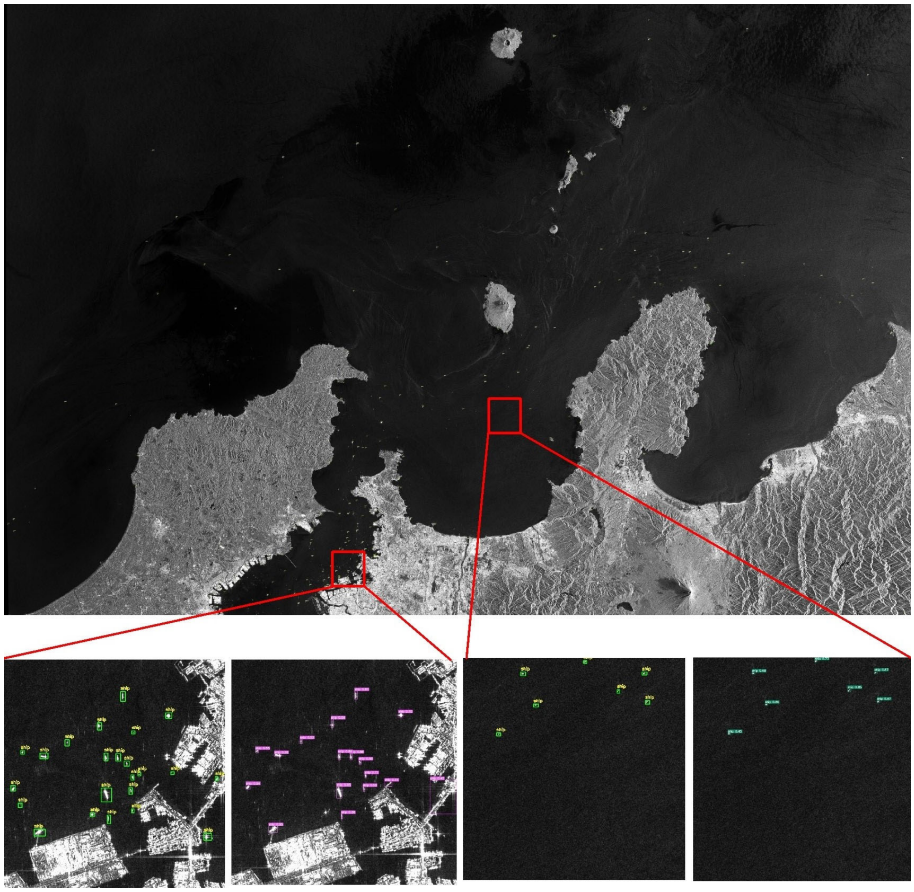


Figure 14. Detection results on the large-scale SAR images. In the same set of pictures, the left subgraph represents the real label, and the right subgraph represents the detection result of CSEF-Net.

3.5. Discussion of Error Detection and Leakage Detection

From the experiments in this chapter, it can be seen that although CSEF-Net has good detection performance in most cases, there are still some cases of missing and wrong detection. When there is a large number of overlapping boundary boxes, as shown in Figure 11d, the boundary boxes of three ships overlap each other, the boundary between features is not clear, and there are a large number of repeated redundant features in the extracted features, resulting in the inability of the middle ship to be detected, indicating that the model is not capable of distinguishing and refining features. When the ship targets are densely packed and there are many small targets, as shown in Figures 11f and 12c,d, the SAR image resolution is low and there is speckle noise on a similar scale to the small target ships, which is not conducive to detection. Coupled with the interference of land buildings and sea clutter, the model takes noise and interference as target error detection. This shows that the anti-interference performance of the model needs to be improved. To sum up, in order to improve the practical application effect of CSEF-Net, further improvements are needed in terms of enhanced feature extraction capability and refinement differentiation, as well as robustness and generalization in complex scenes.

4. Conclusions

In this paper, we note the problems of existing models, which have difficulty detecting cross-scale targets and are subject to complex background interference, and thus propose

the CSEF-Net ship detection algorithm. The algorithm utilizes ERFBNet to construct an efficient sensor field feature extraction backbone network, introduces the MCCA attention mechanism to highlight the target location, and promotes information flow by fusing different scale features through EHFNet. Experimental results on the SSDD and HRSID datasets demonstrate that the method shows effectiveness and better performance for ship target detection in SAR images. In addition, generalization experiments were performed on the untrained dataset. Despite the improvement in missed and false detections in the case of a dense arrangement and a large number of small targets, there are still some targets that are not detected correctly. Future research directions include extending the dataset to improve generalization, optimizing the model to improve detection accuracy and speed, and exploring the change from vertical to rotating detection boxes to better characterize ship positions. Moreover, we will also consider the tasks of ship classification and instance segmentation when ships are detected.

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

Funding: This research was funded by the National Nature Science Founding of China, grant number 61573183.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the need for future work.

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

References

1. Sun, G.C.; Liu, Y.B.; Xiang, J.X.; Liu, W.K.; Xing, M.D.; Chen, J.L. Spaceborne Synthetic Aperture Radar Imaging Algorithms: An overview. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 161–184. [CrossRef]
2. Zhao, Q.; Pepe, A.; Zamparelli, V.; Mastro, P.; Falabella, F.; Abdikan, S.; Bayik, C.; Sanli, F.B.; Ustuner, M.; Avşar, N.B.; et al. Innovative remote sensing methodologies and applications in coastal and marine environments. *Geo-Spat. Inf. Sci.* **2023**, 1–18. [CrossRef]
3. Yasir, M.; Wan, J.H.; Xu, M.M.; Hui, S.; Zhe, Z.; Liu, S.W.; Tugsan, A.; Colak, I.; Hossain, M.S. Ship detection based on deep learning using SAR imagery: A systematic literature review. *Soft Comput.* **2023**, *27*, 63–84. [CrossRef]
4. Qin, X.X.; Zhou, S.L.; Zou, H.X.; Gao, G. A CFAR Detection Algorithm for Generalized Gamma Distributed Background in High-Resolution SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 806–810.
5. Xing, X.W.; Ji, K.F.; Zou, H.X.; Sun, J.X.; Zhou, S.L. High resolution SAR imagery ship detection based on EXS-C-CFAR in Alpha-stable clutters. In Proceedings of the 2011 IEEE International Geoscience and Remote Sensing Symposium, Vancouver, BC, Canada, 24–29 July 2011; pp. 316–319.
6. Wang, C.K.; Wang, J.F.; Liu, X.Z. A Novel Algorithm for Ship Detection in SAR Images. In Proceedings of the 2019 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Dalian, China, 20–22 September 2019; pp. 1–5.
7. Madjidi, H.; Laroussi, T.; Farah, F. A robust and fast CFAR ship detector based on median absolute deviation thresholding for SAR imagery in heterogeneous log-normal sea clutter. *Signal Image Video Process.* **2023**, *17*, 2925–2931. [CrossRef]
8. Ma, W.; Achim, A.; Karakuş, O. Exploiting the Dual-Tree Complex Wavelet Transform for Ship Wake Detection in SAR Imagery. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6–11 June 2021; pp. 1530–1534.
9. Zhu, J.W.; Qiu, X.L.; Pan, Z.X.; Zhang, Y.T.; Lei, B. Projection Shape Template-Based Ship Target Recognition in TerraSAR-X Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 222–226. [CrossRef]
10. Arivazhagan, S.; Jebarani, W.S.L.; Shebiah, R.N.; Ligi, S.V.; Kumar, P.V.H.; Anilkumar, K. Significance based Ship Detection from SAR Imagery. In Proceedings of the 2019 1st International Conference on Innovations in Information and Communication Technology, Chennai, India, 25–26 April 2019; pp. 1–5.
11. Wang, H.P.; Xu, F.; Chen, S.S. Saliency Detector for SAR Images Based on Pattern Recurrence. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2891–2900. [CrossRef]
12. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]

13. Wang, R.F.; Xu, F.Y.; Pei, J.F.; Wang, C.W.; Haung, Y.L.; Yang, J.Y.; Wu, J.J. An Improved Faster R-CNN Based on MSER Decision Criterion for SAR Image Ship Detection in Harbor. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1322–1325.
14. Lin, Z.; Ji, K.F.; Leng, X.G.; Kuang, G.Y. Squeeze and Excitation Rank Faster R-CNN for Ship Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 751–755. [CrossRef]
15. Ke, X.; Zhang, X.L.; Zhang, T.W.; Shi, J.; Wei, S.J. SAR Ship Detection Based on an Improved Faster R-CNN Using Deformable Convolution. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Brussels, Belgium, 11–16 July 2021; pp. 3565–3568.
16. Jiao, J.; Zhang, Y.; Sun, H.; Yang, X.; Gao, X.; Hong, W.; Fu, K.; Sun, X. A Densely Connected End-to-End Neural Network for Multiscale and Multiscene SAR Ship Detection. *IEEE Access* **2018**, *6*, 20881–20892. [CrossRef]
17. Xu, X.W.; Zhang, X.L.; Zeng, T.J.; Shi, J.; Shao, Z.K.; Zhang, T.W. Group-Wise Feature Fusion R-CNN for Dual-Polarization SAR Ship Detection. In Proceedings of the 2023 IEEE Radar Conference (RadarConf23), San Antonio, TX, USA, 1–5 May 2023; pp. 1–5.
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
19. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the 2016 European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 10–16 October 2016; pp. 21–37.
20. Liu, Y.; Wang, X.Q. SAR Ship Detection Based on Improved YOLOv7-Tiny. In Proceedings of the 2022 IEEE 8th International Conference on Computer and Communications (ICCC), Chengdu, China, 9–12 December 2022; pp. 2166–2170.
21. Zha, C.; Min, W.D.; Han, Q.; Li, W.; Xiong, X.; Wang, Q.; Zhu, M. SAR ship localization method with denoising and feature refinement. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106444. [CrossRef]
22. Zhang, T.W.; Zhang, X.L.; Shi, J.; Wei, S.J. HyperLi-Net: A hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 123–153. [CrossRef]
23. Miao, T.; Zeng, H.C.; Yang, W.; Chu, B.; Zou, F.; Ren, W.J.; Chen, J. An Improved Lightweight RetinaNet for Ship Detection in SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4667–4679. [CrossRef]
24. Zhang, T.W.; Zhang, X.L. ShipDeNet-20: An Only 20 Convolution Layers and <1-MB Lightweight SAR Ship Detector. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1234–1238.
25. Kong, W.M.; Liu, S.W.; Xu, M.M.; Yasir, M.; Wang, D.W.; Liu, W.T. Lightweight algorithm for multi-scale ship detection based on high-resolution SAR images. *Int. J. Remote Sens.* **2023**, *44*, 1390–1415. [CrossRef]
26. Suo, Z.; Zhao, Y.; Hu, Y. An Effective Multi-Layer Attention Network for SAR Ship Detection. *J. Mar. Sci. Eng.* **2023**, *11*, 906. [CrossRef]
27. Li, Y.G.; Zhu, W.G.; Li, C.X.; Zeng, C.Z. SAR image near-shore ship target detection method in complex background. *Int. J. Remote Sens.* **2023**, *44*, 924–952. [CrossRef]
28. Zhu, H.; Xie, Y.; Huang, H.; Jing, C.; Rong, Y.; Wang, C. DB-YOLO: A Duplicate Bilateral YOLO Network for Multi-Scale Ship Detection in SAR Images. *Sensors* **2021**, *21*, 8146. [CrossRef]
29. Zhang, T.; Zhang, X.; Ke, X. Quad-FPN: A Novel Quad Feature Pyramid Network for SAR Ship Detection. *Remote Sens.* **2021**, *13*, 2771. [CrossRef]
30. Zhang, T.W.; Zhang, X.L.; Liu, C.; Shi, J.; Wei, S.J.; Ahmad, I. Balance learning for ship detection from synthetic aperture radar remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *182*, 190–207. [CrossRef]
31. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
32. Chen, J.R.; Kao, S.H.; He, H.; Zhuo, W.P.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.
33. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
34. Hou, Q.B.; Zhou, D.Q.; Feng, J.S. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
35. Tong, Z.J.; Chen, Y.H.; Xu, Z.W.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
36. Wang, J.W.; Xu, C.; Yang, W.; Yu, L. A Normalized Gaussian Wasserstein Distance for Tiny Object Detection. *arXiv* **2021**, arXiv:2110.13389.
37. Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. SAR Ship Detection Dataset (SSDD): Official Release and Comprehensive Data Analysis. *Remote Sens.* **2021**, *13*, 3690. [CrossRef]
38. Wei, S.J.; Zeng, X.F.; Qu, Q.Z.; Wang, M.; Su, H.; Shi, J. HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation. *IEEE Access* **2020**, *8*, 120234–120254. [CrossRef]
39. Zhang, T.; Zhang, X.; Ke, X.; Zhan, X.; Shi, J.; Wei, S.; Pan, D.; Li, J.; Su, H.; Zhou, Y.; et al. LS-SSDD-v1.0: A Deep Learning Dataset Dedicated to Small Ship Detection from Large-Scale Sentinel-1 SAR Images. *Remote Sens.* **2020**, *12*, 2997. [CrossRef]

40. Tan, M.X.; Pang, R.M.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
41. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
42. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
43. Yu, C.S.; Shin, Y. SAR ship detection based on improved YOLOv5 and BiFPN. *ICT Express* **2023**, *in press*. [CrossRef]
44. Xiao, M.; He, Z.; Li, X.Y.; Lou, A.J. Power Transformations and Feature Alignment Guided Network for SAR Ship Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
45. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S. A SAR Dataset of Ship Detection for Deep Learning under Complex Backgrounds. *Remote Sens.* **2019**, *11*, 765. [CrossRef]