remote sensing

MDPI

Article

DSF-Net: A Dual Feature Shuffle Guided Multi-Field Fusion Network for SAR Small Ship Target Detection

Zhijing Xu, Jinle Zhai *, Kan Huang and Kun Liu

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China; zjxu@shmtu.edu.cn (Z.X.); huangkan@shmtu.edu.cn (K.H.); kunliu@shmtu.edu.cn (K.L.)
* Correspondence: 202230310004@stu.shmtu.edu.cn

Abstract: SAR images play a crucial role in ship detection across diverse scenarios due to their all-day, all-weather characteristics. However, detecting SAR ship targets poses inherent challenges due to their small sizes, complex backgrounds, and dense ship scenes. Consequently, instances of missed detection and false detection are common issues. To address these challenges, we propose the DSF-Net, a novel framework specifically designed to enhance small SAR ship detection performance. Within this framework, we introduce the Pixel-wise Shuffle Attention module (PWSA) as a pivotal step to strengthen the feature extraction capability. To enhance long-range dependencies and facilitate information communication between channels, we propose a Non-Local Shuffle Attention (NLSA) module. Moreover, NLSA ensures the stability of the feature transfer structure and effectively addresses the issue of missed detection for small-sized targets. Secondly, we introduce a novel Triple Receptive Field-Spatial Pyramid Pooling (TRF-SPP) module designed to mitigate the issue of false detection in complex scenes stemming from inadequate contextual information. Lastly, we propose the R-tradeoff loss to augment the detection capability for small targets, expedite training convergence, and fortify resistance against false detection. Quantitative validation and qualitative visualization experiments are conducted to substantiate the proposed assumption of structural stability and evaluate the effectiveness of the proposed modules. On the LS-SSDDv1.0 dataset, the mAP_{50-95} demonstrates a remarkable improvement of 8.5% compared to the baseline model. The F_1 score exhibits a notable enhancement of 6.9%, surpassing the performance of advanced target detection methods such as YOLO V8.

Keywords: synthetic aperture radar(SAR); small ship detection; dual shuffle mechanism; multiple receptive fields; non-local structure

Citation: Xu, Z.; Zhai, J.; Huang, K.; Liu, K. DSF-Net: A Dual Feature Shuffle Guided Multi-Field Fusion Network for SAR Small Ship Target Detection. *Remote Sens.* **2023**, *15*, 4546. https://doi.org/10.3390/rs15184546

Academic Editors: Maximilian Rodger and Raffaella Guida

Received: 22 August 2023
Revised: 9 September 2023
Accepted: 12 September 2023
Published: 15 September 2023

1. Introduction

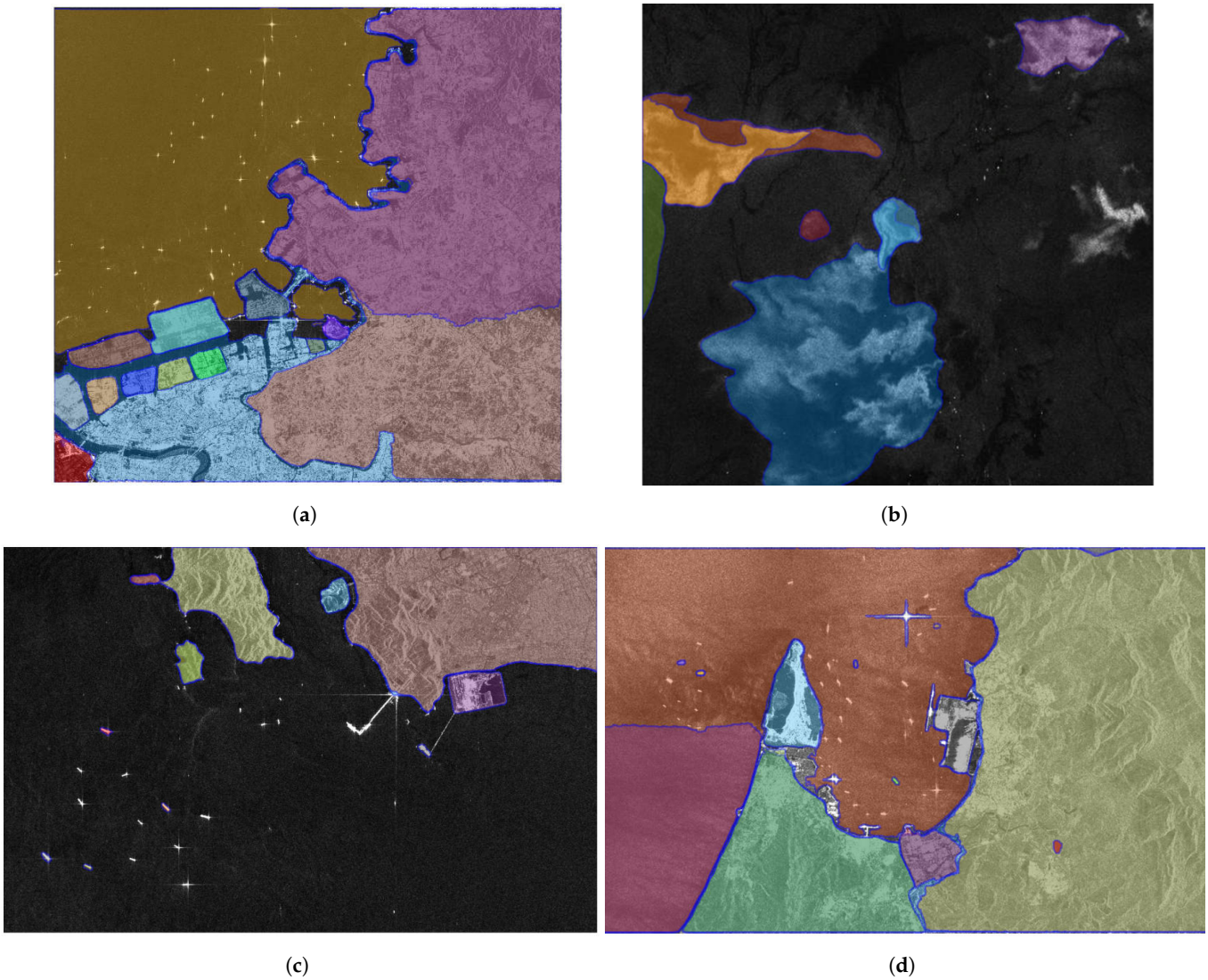
Over the past years, remote sensing images have gained considerable significance in various domains such as ocean monitoring, urban planning, resource exploration, and military defense [1]. Among these tasks, ship target detection is the most researchable [2], practical, and typical. Remote sensing ship images mainly include visible remote sensing images, infrared images, and SAR images. Visible remote sensing images provide a wealth of pixel-level information; however, their imaging process is susceptible to various environmental factors, such as lighting conditions, seasonal variations, and other perturbations. Infrared images, due to their low signal-to-noise ratio and limited structural information, are not conducive to the detection of ship targets. Therefore, they are not suitable for monitoring scenes with low light conditions, high cloud cover, and small temperature differences on the Earth's surface. Conversely, SAR images are extensively employed in real-time port monitoring, military target identification, and tracking due to their all-day and all-weather characteristics [3]. The identification of small ships in large-scale SAR images bears significant practical value and endures as a paramount research objective warranting attention.

The detection of small SAR ship targets presents challenges in terms of missed detection and false detection. As shown in Figure 1, even the state-of-the-art Segment Anything [4] suffers from these problems. The reasons can be attributed to:

(1) Environmental factors, such as sea clutter, introduce multiplicative speckle noise and blur details, thereby degrading the quality of SAR images [5]. On the one hand, the similarity in size between speckle noise and small-sized SAR ship targets can result in false detection of small-sized ship targets in SAR images. On the other hand, the presence of clutter and sidelobes can cause missed detection of small-sized ship targets in SAR images [6]. In addition, the moving small ship targets produce different degrees of geometric deformation, which in turn results in missed detections of small ship targets.
(2) In the inshore area, ship targets are characterized by a huge number, dense arrangement, and diverse scales. During the prediction process, a substantial overlap occurs among the generated bounding boxes. This leads to the loss of valid boxes after applying non-maximum suppression, consequently resulting in missed detection issues.
(3) Since the scales of the ship targets are small, the inshore docks and coastal islands can be wrongly detected as targets, resulting in the missed detection of small ship targets.

The traditional SAR ship target detection algorithms are based on the statistical characteristics of pixels and spatial distribution, such as the CFAR-Based method [7,8]. In addition, there are studies on SAR ship target detection based on sliding window method [9], global threshold segmentation method [10], extraction of hull texture and scale characteristics using manual features [11], and assisted detection using ship wake [12]. Although the traditional methods have made some progress, they still have the problems of tedious operation processes, complicated calculation procedures, poor generalization ability, and limited detection accuracy, etc.

We propose the DSF-Net to tackle the challenges of small target feature loss, background-target confusion, dense ship detection overlap, and network structure stability. The DSF-Net comprises two kinds of shuffle-guided attention mechanisms, from fine to coarse. It also includes a multi-scale feature enhancement fusion module with multiple receptive fields and a trade-off loss function that balances small target detection accuracy and training convergence speed. We design the Pixel-wise Shuffle Attention(PWSA) module, which consists of group convolution, dual feature enhancement, and the pivotal Pixel-wise Shuffle operation. This module employs group convolution to extract a more comprehensive range of semantic information and capture more complete features of small ship targets in SAR. It applies dual enhancement in spatial and channel domains to optimize the utilization of effective channel information and reinforce the spatial positioning information of SAR ship targets. Subsequently, the Pixel-wise Shuffle operation is conducted to enhance the fusion of different channel information, facilitate spatial position information exchange, and improve feature resolution, thereby enhancing the representation of fine details. Moreover, we propose the NLSA (Non-Local Shuffle Attention) module, which embeds the structure of the Non-Local in the group convolution and subdivides the features after grouping. We divide the subgroup features into shared features and unshared features, forming the Non-Local dual enhancement. Furthermore, we designed a Triple Receptive Field-Spatial Pyramid Pooling (TRF-SPP) structure to expand the receptive field of features and enhance the ability to capture global information. By employing TRF-SPP, we effectively address the issue of false detection, preventing the misidentification of docks, islands, and other backgrounds as ship targets, which may arise due to the limited context captured by local features. To leverage the advantages of the CIOU Loss and NWD Loss, we propose the R-tradeoff loss as a compromise between precision in bounding box regression and training convergence. The R-tradeoff loss is insensitive to fluctuations in the position of small targets, thereby enhancing the accuracy of small ship detection.



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

(**c**) (**d**)

Figure 1. Results of Segment Anything [4] in typical scenes. (**a**,**c**,**d**) Inshore Complex Background. (**b**) Background with cross-shaped bright spots.

The main contributions of this study can be summarized as:

We establish a connection between the shuffle operation, Non-Local structure, and attention mechanism. Furthermore, We delve into the integration of the multi-receptive field and feature fusion module. We address the challenge posed by ships of varying scales and propose the R-tradeoff loss to achieve an optimal balance.

(1) We propose the PWSA (Pixel-wise Shuffle Attention) module to address the issue of insufficient effective features. The main objective of this module is to enhance the feature extraction of small SAR ship targets and improve the Backbone's capability to extract features relevant to small SAR ships. PWSA enhances the extraction of cross-channel positional information, achieving data augmentation during the feature extraction process.

(2) We introduce a Non-Local Shuffle Attention (NLSA) module to enhance the long-range dependency between different spatial locations. The NLSA module fosters better spatial relationships among pixels at various locations and ensures feature stability during fusion across different dimensions. Compared to traditional attention methods, NLSA, through the Non-Local structure, is more adept at capturing long-range dependencies between small-scale SAR ships. Moreover, the feature shuffle within NLSA enhances the feature representation capability for small SAR ship targets.

Another notable advantage of NLSA is that it ensures the stability of the feature extraction structure but merely increases a slight number of parameters.

(3) We design a multi-scale feature fusion module called TRF-SPP (Triple Receptive Field-Spatial Pyramid Pooling). Compared to previous multi-scale fusion approaches such as SPP and SPPF, TRF-SPP boasts a larger receptive field and a more comprehensive contextual understanding. We propose an SCTS (Squeeze Concatenate and Triple Split) technique, which effectively enhances the features of small SAR ship targets across various receptive field dimensions, reducing the likelihood of false detection.

(4) To achieve precise detection of multi-scale SAR ship targets while ensuring a rapid convergence in the training process, we propose an innovative R-tradeoff loss specifically tailored for small targets of ships. The R-tradeoff loss exhibits strong resilience to scale variations in SAR ship targets, enhancing the overall robustness of DSF-Net.

The remaining part of this study is organized as follows: in Section 2, the relevant research works are briefly described. The proposed DSF-Net is described in Section 3, and in Section 4, the experimental protocols and results are reported and analyzed. Finally, the conclusion of this study is drawn in Section 5.

2. Related Work

2.1. Deep Learning in SAR Ship Detection

With the increasing computational power, deep learning methods, known as GPU technology aesthetics [13], have become prominent in object detection methods. In the field of general-purpose object detection, the overall model architecture can be divided into anchor-based methods, including two-stage object detection methods [14] and one-stage object detection methods [15–18]; anchor-free detection methods [19], and Transformer structure-based detection methods [20]. Notably, deep learning techniques have also found applications in the SAR ship target detection domain, where researchers have successfully leveraged them to tackle specific challenges. For instance, Zhou et al. [21] proposed a two-stage detection model incorporating a Doppler feature matrix to mitigate image blurring and focusing difficulties caused by ship motion. Ma et al. [22] employed an anchor-free detection framework to address the problem of missed detection of small ship targets in dense scenes through keypoint estimation. Furthermore, Xia et al. [23] combined the strengths of convolutional neural networks and visual Transformers to effectively handle issues such as background interference and unclear ship edges. Guo et al. [24] introduced a real-time ship target detection method that leverages the speed advantages of one-stage detection algorithms. Although different model architectures have shown improvements in the accuracy of ship target detection, they often neglect a crucial aspect: the detection of small-scale ship targets, which is a challenging task in ship target detection. Small-scale ship targets suffer from issues such as being easily confused with the background and loss of distinctive features.

To address these challenges, some researchers have proposed a multi-scale fusion structure and an embedded attention module to enhance small target features' capturing capability and detection accuracy. For example, Zhang et al. [25] employed SE attention [26] as a regulatory module to balance the contribution of each feature. Although the SE module primarily focuses on weight assignment among different channels, it fails to consider the positional information of the ship target. To address this limitation, Su et al. [27] proposed a channel-location attention mechanism (CLAM) module that combines spatial and channel information to enhance the feature extraction capability of the Backbone network. However, the CLAM module is limited in its ability to capture long-range dependencies between features at different locations. In another study, Chen et al. [28] utilized shape similarity distance as a metric to optimize the Feature Pyramid Network (FPN) structure using the k-means clustering method. Nevertheless, this approach only conducted clustering analysis for the anchors and overlooked the possibility of losing the location information of small targets during the upsampling process of FPN. Although it does yield stronger semantic information, it may sacrifice location details for smaller targets within low-dimensional

feature maps. In order to enhance the fusion of multi-scale information and quantify the disparity between predictions and ground truth, several studies have sought to improve the loss function. For example, Zhang et al. [29] introduced the Global Average Precision (GAP) loss to address the issue of prediction score shifting, while Xu et al. [30] proposed the TDIOU loss to more accurately measure the positional relationship between predictions and actual ship targets. Despite the considerable efforts devoted by previous researchers to SAR ship target detection, several crucial aspects have been overlooked. For instance, the extraction of location and feature information from small ship targets remains insufficient. Furthermore, the current feature extraction methods primarily focus on local information, neglecting the extraction and fusion of global information features. Additionally, the existing loss functions predominantly concentrate on the positional alignment of predictions and ground truth while disregarding the variations in scale and position for small pixel targets. These unresolved aspects warrant further investigation and attention.

2.2. Attention Mechanism

Over-extraction of redundant features makes the network deviation from the focus. In order to refocus the network on the salient features, the attention mechanism comes into being. Although earlier attention mechanisms merely focused on the channel information of features, such as SE-Net [26], it provided a new perspective to effectively enhance the feature representation capability. Inspired by SE Attention, Woo et al. [31] started to consider the problem from both channel and spatial perspectives and proposed CBAM Attention. The CBAM attention enhances the capture of features with greater contribution to the prediction and suppresses the over-extraction of redundant features, leading to superior performance. Since then, the two dimensions of the spatial and channel have become the Evergreen tree of attention module design. With the emergence of Shuffle Net [32], the practical operation of channel shuffle started to be noticed and gradually integrated with the attention module [33–35]. For instance, Zhang et al. [35] used group convolution with channel shuffling to enhance the features in both spatial and channel dimensions, resulting in a notable improvement in feature extraction efficiency.

2.3. Non-Local structure

In the image classification, object detection, and semantic segmentation tasks, the extracted features often suffer from limitations imposed by the local receptive field. Additionally, the detection or classification based on local features will produce different results, resulting in category confusion. To address this issue, capturing long-range dependencies becomes crucial. To this end, Wang et al. [36] proposed Non-Local structure, which is a new paradigm for capturing long-range dependencies without changing the network structure. Building upon this, Fu et al. [37] further developed the application of a Non-Local structure in computer vision tasks by combining it with the attention module, which reduces the probability of category confusion.

2.4. Multi-Scale Feature Fusion

The extraction of effective features is hindered by challenges such as small image size, low resolution, and insignificant texture. Leveraging semantic information is a crucial approach to tackling these challenges. However, the conventional convolutional blocks often increase computational parameters significantly, and the recurrent use of pooling and downsampling modules may lead to the diminishment of features associated with small targets.

To address this problem, some researchers propose to perform feature extraction and fusion from multiple scales, which avoids the stacking of a large number of convolutional blocks and extracts more effective contextual information at the same time. To acquire multi-scale features, He et al. [38] proposed spatial pyramid pooling for acquiring multi-scale features. Building upon this concept, Glenn [39] further refined and consolidated the approach, successfully applying it to object detection. Similarly, Liu et al. [40] designed

the RFB module by replacing the convolution module with atrous convolution to obtain a larger receptive field. Not coincidentally, Chen et al. [41] developed an encoder-decoder architecture employing atrous separable convolution, yielding favorable results in contextual information extraction for image segmentation tasks.

In the field of SAR ship target detection, multi-scale feature fusion also finds applications. For instance, Wang et al. [42] employed a method that combines Dilated-Res Units with multi-scale features to expand the receptive field and integrate contextual information. Yang et al. [43] proposed a multi-scale adaptive feature pyramid network (MSAFPN), which achieves cross-scale fusion. Liu et al. [44] improved the structure of the RFB and applied it to SAR ship target detection. Hong et al. [45] proposed a Gaussian-YOLO layer to address the issue of multi-scale object detection variations.

2.5. Design of the Loss Function

With the improvement of neural network detection performance, the traditional loss function can no longer meet the requirements for accurate localization. Therefore, some researchers proposed the IOU loss based on the proximity of the bounding boxes to the predictions in the shape. However, the IOU loss has disadvantages, particularly its insensitivity to the prediction position. To address this issue, Hamid et al. [46] proposed the GIOU loss function and introduced the concept of minimum closure, but there is a problem that the loss is the same while the quality of the prediction is different. To overcome this, Zheng et al. [47] introduced penalties based on the GIOU loss, including overlap area, central point, and aspect ratio, and proposed the DIOU and CIOU losses. Additionally, [48,49] proposed EIOU loss and SIOU loss improve the prediction precision from the perspective of bounding box regression quality and bounding box angle, respectively. Moreover, [50] utilized the property that Wasserstein distance is insensitive to small target's scale fluctuation and proposed the Normalized Wasserstein Distance loss to improve the detection performance of small targets.

3. Methods

In this section, we introduce the overall architecture of the DSF-Net, the structure of the PWSA module, which is embedded in the Backbone of the DSF-Net, the structure of the NLSA module, which is located in the lateral connection between Backbone and Neck, the structure of the TRF-SPP module, which is placed at the bottom of the Backbone, and the design of the R-tradeoff loss.

3.1. Overall Architecture of DSF-Net

The main Architecture of DSF-Net is depicted in Figure 2. CSPDarknet53 [39] is chosen as the baseline Backbone. We modularize the entire one-stage detection network into three components: the Backbone, responsible for extracting the coarse features; the Neck, responsible for refining and fusing the features; and the Detection Head, which completes the final stage of the object detection task. At the same time, we consider that the lateral connection part between the Backbone and the Neck is crucial. It determines whether the original features for fusion have a good representation ability. Additionally, the stability of the lateral connection determines the effectiveness of the final feature fusion. Given its bridging function, connecting the Backbone and the Neck like a bridge, we refer to the starting point of the lateral connection on the Backbone as the Bridge Node. We add the PWSA module to the Backbone and replace the conventional SPPF with TRF-SPP at the tail of the Backbone network. For the Neck part, we add the NLSA module at the Bridge Node and also adopt the Neck with PANet [51] structure for multi-scale feature fusion. The output of the Neck is divided into three inputs, which are subsequently fed into the Detection Head for the final object prediction.

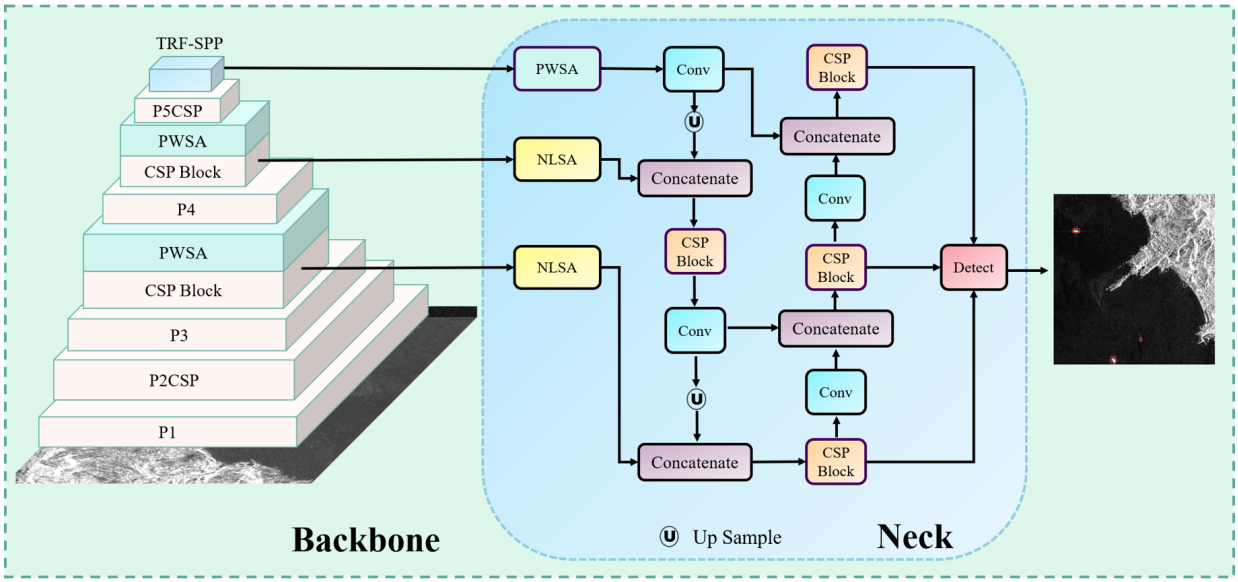


Figure 2. Illustration of the framework of DSF-NET.

3.2. PWSA Module

During the SAR imaging process, small ship targets are susceptible to the multiplicative speckle noise, owing to their inherent size characteristics and imaging properties. Given the similarity in size between speckle noise and small ship targets in SAR, there is a higher probability of confusion, which can result in false detections. Furthermore, it is worth noting that generic Backbone networks exhibit a limited capacity for extracting features from small ship targets, thereby impeding the capture of their complete characteristics. With the deepening of the network layers, there will be a problem of losing the features of small ship targets. To further enhance the feature extraction capability of the backbone network for small ship targets, we have designed the PWSA (Pixel-Wise Shuffle Attention) module, as illustrated in Figure 3.

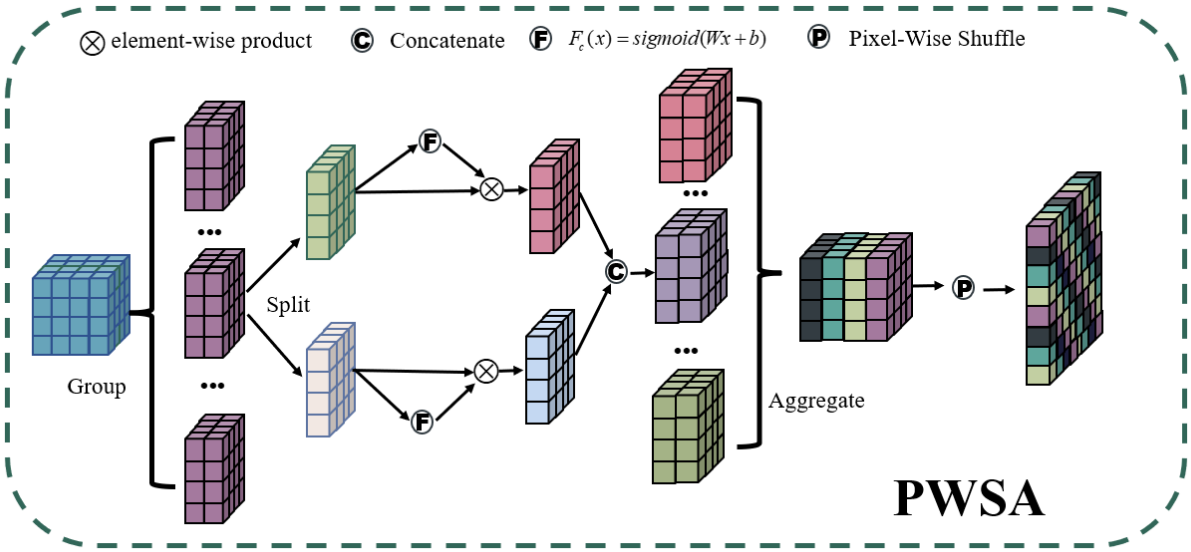


Figure 3. The structure of Pixel-Wise Shuffle Attention module.

The PWSA module performs group-wise convolution on the input features, reducing the computational complexity and improving the computational efficiency of the model. To focus the network on the features of small ship targets, we employ dual enhancement in both spatial and channel dimensions. On the one hand, channel attention is utilized to

adaptively learn the importance and contribution of different channels for SAR ship target detection, thereby enhancing the mining and utilization of effective channel information. On the other hand, spatial attention is employed to help the model better understand the spatial relationship of ship targets in SAR images, aiding in their detection and localization. In addition, we introduce the Pixel-Wise Shuffle operation, which redistributes information across different channels. This operation not only enhances the fusion of information from different channels but also improves the communication of spatial positioning information for SAR ship targets.

Compared to other forms of Shuffle Attention, PWSA is more suitable for detecting small SAR ship targets. Due to the use of group convolution in PWSA, there is no need to be concerned about a significant increase in computational complexity when incorporating PWSA. Therefore, PWSA can be integrated into deep networks, enabling more efficient performance. The dual enhancement in both spatial and channel dimensions guarantees a more comprehensive extraction of features for small targets within the network. This effectively eliminates the restrictions posed by the traditional single-dimensional attention enhancement, which often leads to redundant feature extraction for small ship targets. Unlike channel shuffle, Pixel-Wise Shuffle offers a cross-channel coordinate interrelation, facilitating information communication for small SAR ship targets from both the channel and position perspectives. This can be comprehended from the following perspective: Pixel-Wise Shuffle can be regarded as a unique form of data augmentation within the feature extraction process. It achieves precise feature localization for small SAR ship targets by enhancing the extraction of cross-channel positional information.

The flow of features in the PWSA module is as follows: Firstly, the input features are processed in a group-wise manner. Specifically, the input features are divided into g groups in the channel dimension, where the channel dimension of each group is c/g. Then, the features of each group are equally divided, resulting in a channel dimension of $c/2g$ for each part. The two parts of features are separately enhanced in the spatial dimension and channel dimension, and then aggregated by concatenating them to reduce the number of parameters. Subsequently, each group of features is aggregated. Finally, the Pixel-Wise Shuffle operation is adopted to integrate the features into the form of $(C/4, 2H, 2W)$, enabling the communication of information across different channels while improving the resolution of the target.

The group enhancement process in PWSA can be expressed by Equations (1) and (2):

$$G_1 = \sigma\left(\frac{W_1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_1(i,j) + b_1\right) \otimes X_1 \tag{1}$$

$$G_2 = \sigma(W_2 \cdot GN(X_2) + b_2) \otimes X_2 \tag{2}$$

where X_1, X_2 are the two part features after splitting, G_1, G_2 are the output of X_1, X_2 after enhancing from spatial and channel, H, W are the height and width of the output features, W_1, W_2, b_1, b_2 are the weights and biases, respectively, GN is Group Normalization, $\otimes$ denotes element-wise product.

The output of PWSA can be expressed as follows:

$$\begin{aligned} P &= S(E(G_1, G_2), r) \\ &= E(\underset{i=1}{\overset{c/r^2}{E}}(G_1, G_2), \underset{j=1}{\overset{r \cdot H}{E}}(G_1, G_2), \underset{k=1}{\overset{r \cdot W}{E}}(G_1, G_2)) \end{aligned} \tag{3}$$

where S denotes the Pixel-Wise Shuffle operation, and E denotes the Concatenate operation. The superscript indicates the channel size, while the subscripts i, j, k represent the dimensions of the features.

3.3. NLSA Module

Figure 4 illustrates the structure of the NLSA module, which receives the features extracted from the Backbone. Firstly, we divide the features into groups to improve the efficiency of feature extraction. Then, we embed the structure of Non-Local [36] into the feature enhancement part, dividing each group of features into two parts of shared features. One part is used for dual enhancement, and the other part constructs Non-Local connections to capture long-range dependencies. The fusion is performed by a Non-Local structure. We replace the initial element-wise sum with concatenate to better preserve multiple features and increase the stability of the structure. Finally, multiple features are aggregated, and channel shuffling is performed to increase the information exchange among channels. For the choice of the location to make feature enhancement, we find that the structure of dual feature enhancement is the best choice. If we make any of the dual feature enhancement and Conv1 × 1 multiplex a set of feature maps, it does not play a better detection effect. We argue that multiplexing in a scattered manner may dilute the effect of feature enhancement. At the same time, replacing Conv1 × 1 with a feature enhancement module will destroy the construction of long-range relationships. We describe this situation as the feature dilution phenomenon.

In addition, the precise placement of the PWSA module and the NLSA module within the feature extraction network requires careful consideration. The Bridge Node holds significant importance for two key reasons. Firstly, it represents the most feature-rich layer among its counterparts in the Backbone network. Secondly, it serves as the intersection point between the feature fusion Neck and the Backbone. Consequently, enhancing features at the Bridge Node becomes particularly essential. To bolster the Backbone network's capability in capturing the distinguishing characteristics of small ship targets, we strategically position the PWSA module after the Bridge Node. This arrangement aims to augment the Backbone's capacity for feature extraction. Simultaneously, we position the NLSA module on the lateral connection of the Neck. This placement facilitates the fusion of multi-scale information and ensures the stability of the lateral connection structure.

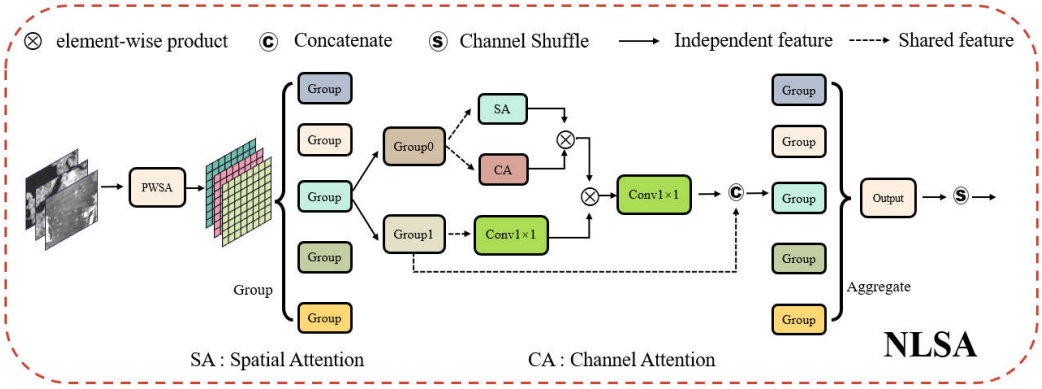


Figure 4. The architecture of Non-Local Shuffle Attention module.

3.4. TRF-SPP Module

In the task of small SAR ship target detection, the issue of false detection presents a significant challenge. Especially in the inshore region, it is easy to incorrectly detect piers and islands as small ship targets. Moreover, the presence of light spots generated by ship movement can also contribute to false detection. We believe that the false detection in the inshore region is mainly due to the weak ability to capture contextual information and the absence of global information on the captured features. Discriminating the inshore background solely based on local information is a daunting task, comparable to attempting to judge the grandeur of Hercules solely from his foot without a comprehensive view. To address these challenges, we propose a novel module called Triple Receptive Field-Spatial Pyramid Pooling (TRF-SPP), depicted in Figure 5, which aims to enhance detection accuracy by incorporating contextual information and capturing global dependencies.

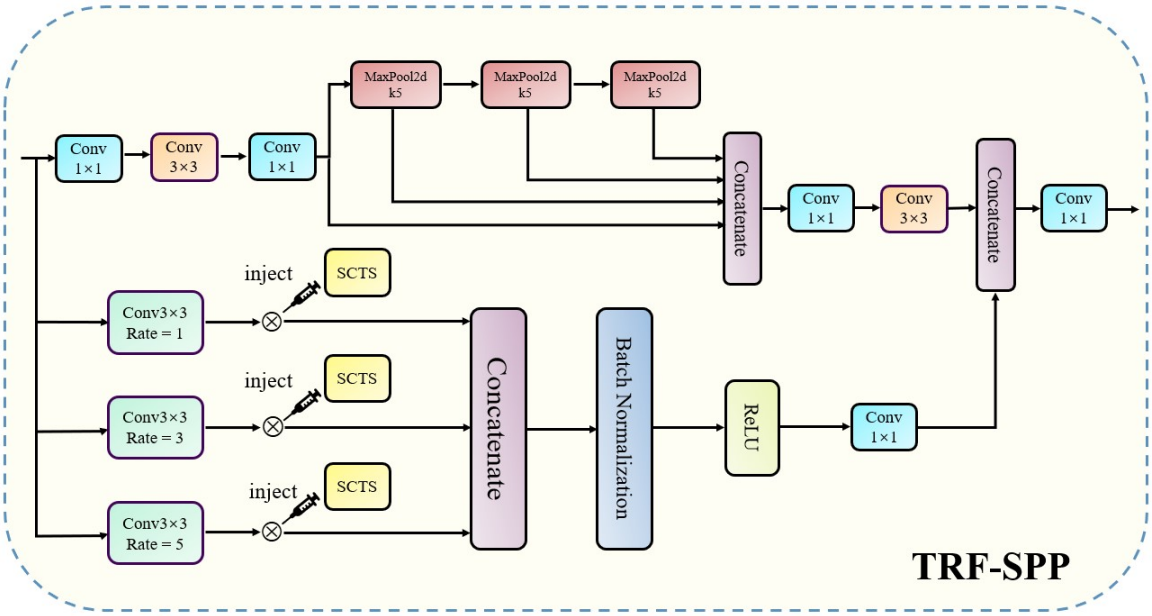


Figure 5. The architecture of Triple Receptive Field-Spatial Pyramid Pooling module.

The input features are divided into two branches. The first branch goes through the SPPF structure to obtain the basic multi-scale features. In the second branch, we incorporate the feature capture structure with atrous convolution to increase the dimensionality of the receptive fields and obtain more contextual information. Meanwhile, we apply the SCTS module to enhance the multiple receptive field features and fuse them with the basic receptive field features extracted by SPPF.

The structure of SCTS is shown in Figure 6. First, the input features are processed by three branches of atrous convolution. Subsequently, the dimensions of the three kinds of convolution are adjusted by using Conv1 × 1. The obtained features are squeezed together to form a feature fusion, and the dimension is changed to 3 by Conv1 × 1. Then, the features are transformed to the weight value domain by Softmax and split into triple weights. Finally, inject these triple weights into TRF-SPP to achieve the enhancement for multiple receptive field features.

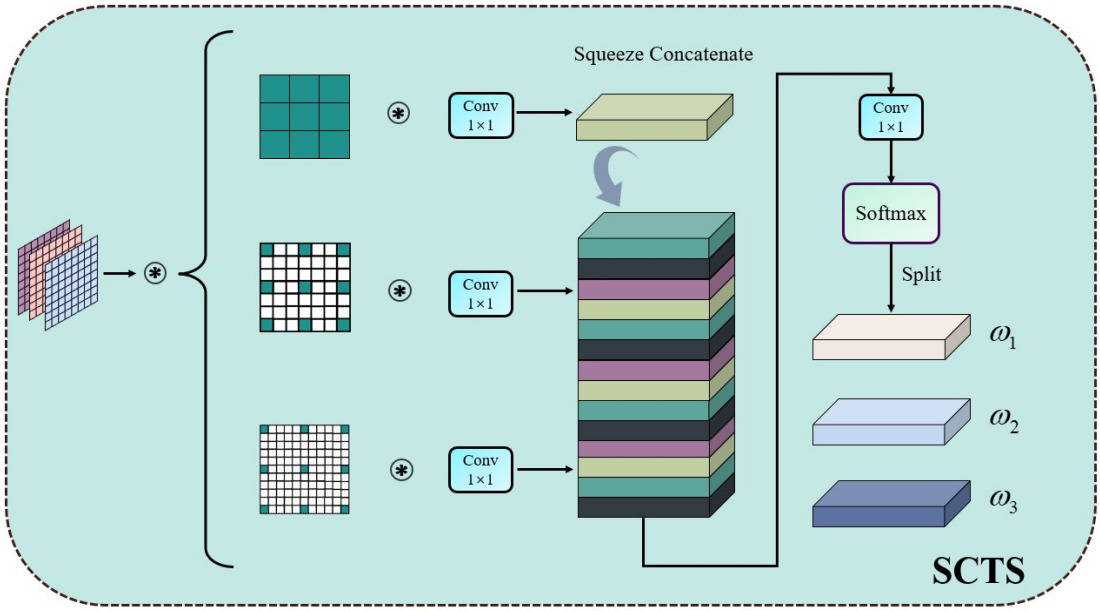


Figure 6. The structure of Squeeze Concatenate-Triple Split module.

3.5. R-Tradeoff Loss Design

With the improvement of the neural networks in capturing features, the traditional loss function is no longer suitable for predicting the degree of similarity between the Bounding boxes and the predictions. To solve this problem, some researchers have proposed IOU loss [46–48], and CIOU loss is one of the pioneers. As shown in Figure 7, CIOU mainly considers the overlap area, central point distance, and aspect ratio [47]. The CIOU loss function can be expressed as:

$$\mathcal{R}_{CIOU} = \frac{D^2}{C^2} + \alpha v \tag{4}$$

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2 \tag{5}$$

$$\alpha = \frac{v}{(1 - IOU) + v} \tag{6}$$

$$L_{CIOU} = 1 - IOU + \mathcal{R}_{CIOU} \tag{7}$$

where D is the Euclidean distance of the center point of boxes. C is the diagonal length of the smallest enclosing box covering two boxes. Additionally, w^{gt} is the width of the ground truth, h^{gt} is the height of the ground truth, w is the width of the bounding box, and h is the height of the bounding box.

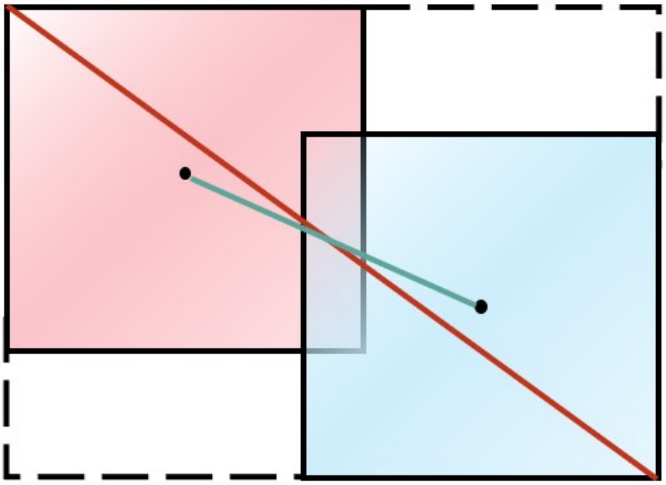
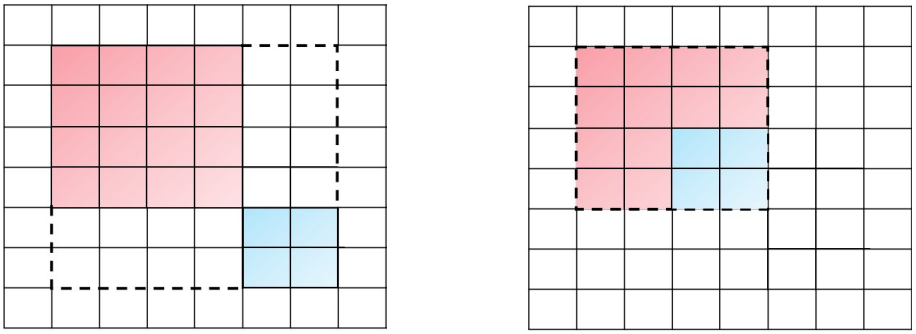


Figure 7. Illustration of the CIOU, where the dashed line represents the minimum closure, the red box indicates the ground truth, the blue box indicates the anchor box, the solid red line represents the diagonal length of the smallest enclosing box covering two boxes, which is expressed by C, and the solid blue line represents the central distance of central points of two boxes, which is expressed by D.

CIOU Loss is extremely sensitive to the movement of small pixel target positions. Moving a very small distance may lead to a large change in loss during the regression of small targets. With CIOU loss and NMS working together, the number of positive samples will be reduced. Although CIOU Loss integrates several metrics, we believe that it produces a large number of missed detections for small targets.

As shown in Figure 8, each box represents a pixel, and the dashed line indicates the minimum closure of the two bounding boxes. The red box represents the ground truth, and the blue box represents the anchor. For a more intuitive presentation, we set the aspect ratio of the two boxes to 1 to represent the bounding box regression for small pixel targets. The CIOU of the two bounding boxes is calculated according to Equations (4)–(7). The CIOU value of Figure 8a is 1.25, and Figure 8b is 0.8125; the increment of the IOU value is 0.4375. It can be explicitly seen that the variation of CIOU is one-third of the initial value after moving only two pixels.

To solve the problem of CIOU missed detection, the first step is to overcome the sensitivity of CIOU to the position of small-sized targets. Previous studies proposed NWD Loss, utilizing the Normalized Gaussian Wasserstein Distance, which exhibited insensitivity to targets of varying scales [50]. However, we believe that NWD Loss is only suitable for small target scenes, and its scale insensitivity presents a dual nature. On the one hand, it mitigates missed detection of small targets, but on the other hand, it may lead to false detection and slower convergence for medium and large-scale targets.



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

Figure 8. Illustration of the CIOU variation instance. (**a**) Cases without intersection, CIOU = 1.25. (**b**) The case of moving two pixels, CIOU = 0.8125.

To address the above problems, we take full advantage of the CIOU and NWD loss and propose R-tradeoff Loss, where R is the tradeoff factor. By adjusting the R-tradeoff factor, the loss function can effectively balance rapid convergence and precise localization when faced with fluctuations in target positions at different scales. When R is large, the NWD loss serves as a relatively small penalty term, and the R-tradeoff loss exhibits properties similar to CIOU. When R is small, BBox is modeled as a Gaussian distribution, and the Wasserstein Distance is used for measurement, enhancing resistance to size fluctuations in small-scale SAR ship targets. Therefore, the R-tradeoff loss demonstrates good robustness for small-scale SAR ship target detection.

The calculation formula can be expressed as:

$$W_2^2(\mathcal{N}_a, \mathcal{N}_b) = \left\| \left([cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2}]^T, [cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2}]^T \right) \right\|_2^2 \tag{8}$$

$$NWD(\mathcal{N}_a, \mathcal{N}_b) = e^{\left(-\frac{\sqrt{W_2^2(\mathcal{N}_a, \mathcal{N}_b)}}{c}\right)} \tag{9}$$

$$\mathcal{L}_R = (1 - R) * (1 - NWD) + R * L_{CIOU} \tag{10}$$

where $\mathcal{N}_a = (cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2})$ represents bounding box A, $\mathcal{N}_b = (cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2})$ represents bounding box B, $\| \cdot \|_2$ represents the 2-Norm and C is a constant factor, L_{CIOU} is the equation shown in Equation (7).

4. Experiments & Discussion

In this section, we describe the experiment dataset, evaluation metrics, experiment details, and experimental results. At the end, we give the corresponding detection instances with comparisons.

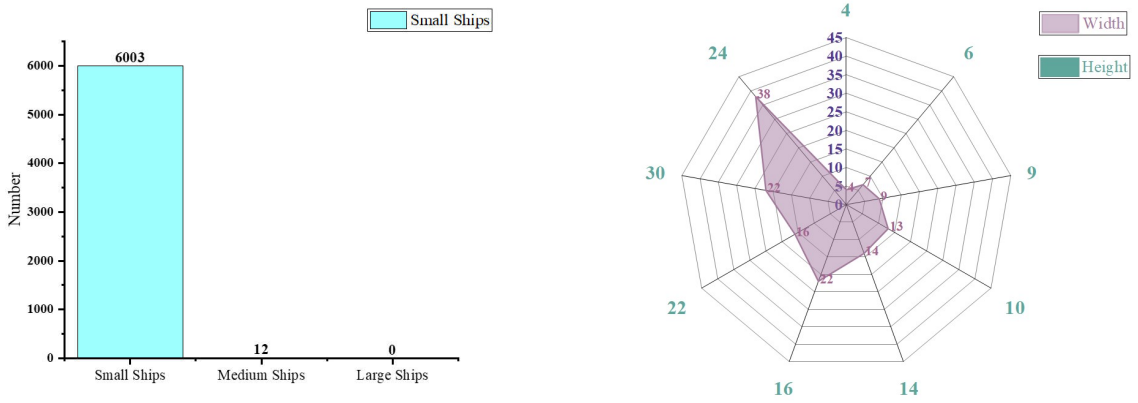
4.1. Dataset

In our experiments, we used LS-SSDDv1 [52] as the benchmark dataset. LS-SSDDv1.0 is a large-scale small SAR ship detection dataset. The relevant information can be referred to Table 1. Due to the excessive aspect ratio of individual images, they are cropped to 800 × 800 in the official dataset to facilitate computer processing and readability. The ratio of the training set and validation set is 2:1, with a total of 9000 images. We counted the

scales of the ships in LS-SSDDv1.0, and the statistics are shown in Figure 9a. At the same time, we utilize the k-means method in all labels to determine the hyperparameters of the anchors, and the results are depicted in Figure 9b. Based on the results, it can be observed that the dataset primarily consists of small-sized ships, with a majority of them being below 32 × 32 pixels. Consequently, detecting these small ships poses significant challenges.

Table 1. The basic parameters of LS-SSDDv1.0 [52].

Parameter	LS-SSDDv1.0
Satellite	Sentinel-1
Sensor mode	IW
Location	Tokyo, Adriatic Sea, etc.
Resolution(m)	5 × 20
Polarization	VV, VH
Image size(pixel)	24,000 × 16,000
Cover width(km)	250
Image number	15



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

Figure 9. The statistics of dataset information. (**a**) LS-SSDDv1.0 target scale statistics. (**b**) Anchor hyperparameters obtained by K-means clustering.

4.2. Evaluation Metrics

The evaluation metrics we adopted were Precision (P), Recall (R), mean Average Precision (mAP), F1 score, and GFLOPs. The calculation formulas are as follows:

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

where TP represents true positives, indicating the correctly detected targets. FP represents false positives, representing the incorrectly detected targets. FN represents false negatives, representing the missed targets.

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{13}$$

where R is the Recall in Equation (11), P is the Precision in Equation (12), and the F_1 score is a composite metric of Recall and Precision.

The curve drawn with the Recall as the horizontal axis and the Precision as the vertical axis is the P-R curve. The area of the P-R curve is the value of Average Precision, which can be calculated by the following formula:

$$AP = \int_0^1 P(R)dR \tag{14}$$

The P and R in the formula stand for Recall and Precision, respectively. AP uses different IOU thresholds as indicators forming a set of evaluation metrics, and the higher the AP values, the better the algorithm detection performance; this paper also refers to this evaluation criterion.

4.3. Experimental Details

All of the experiments were performed under the Pyorch framework, the CPU is the 12th Gen Intel (R) Core (TM) i7-12700F and the GPU is the NVIDIA GeForce RTX 3080 Ti. The initial learning rate was set to 0.01 with a learning rate decay factor of 0.01. The momentum was set to 0.937, and the weight decay was set to 0.0005. A cosine LR scheduler was used, starting from an initial epoch of 200. The batch size was set to 16, and the optimizer used was SGD. The input image size was resized to 640 × 640 for both training and validation. We selected YOLO V5 as the baseline model. For fairness, the parameters were kept consistent with YOLO V5. Additionally, the lowest data augmentation was used to facilitate a fair comparison with other methods.

4.4. Experimental Results

4.4.1. Comparison with the Existing Methods

We compare the DSF-Net proposed in this study with other object detection methods. These methods include state-of-the-art object detection methods, two-stage object detection methods, one-stage object detection methods, anchor-free methods, and the latest methods in the field of SAR small ship target detection. The results are given in Table 2 at length. The results of visualization are presented in Figures 10–13.

Table 2. Comparison with other detection Methods in LS-SSDD-v1 [52] of entire scenes. Where the superscript with * indicates that the results are from [27], and the superscript with † indicates that the results are from [53].

Method	P (%)	R (%)	mAP_{50} (%)	mAP_{50-95} (%)	F_1 (%)	GFLOPs
Faster R-CNN [14]	58	61.6	57.7	–	59.75	–
Cascade R-CNN [54]	54.1	66.2	59.0	–	59.54	–
YOLO V5 [39]	84	63.6	73.3	27.1	72	15.9
YOLO V8 [55]	82.4	67	74.4	29	74	28.4
Filtered Convolution [56]	–	–	73	–	–	–
Guided Anchoring * [57]	80.1	63.8	59.8	–	71.0	–
FoveaBox * [58]	77.5	59.9	52.2	–	67.6	–
FCOS * [59]	50.5	66.7	63.2	–	57.48	–
MTL-Det * [60]	–	–	71.7	–	–	–
ATSS * [61]	74.2	71.5	68.1	–	72.8	–
YOLO X † [19]	66.78	75.44	–	–	70.85	–
RefineDet † [62]	66.72	70.23	–	–	68.43	–
SII-Net [27]	68.2	**79.3**	76.1	–	73.3	–
DSF-Net(ours)	**86.4**	68.7	**76.9**	**29.4**	**77**	33.5

The bold font represents the best results among all methods.

We selected the typical scenes to demonstrate the detection effect. The scenes consist of densely arranged ship areas and inshore scenes. The detection results are shown in Figures 11 and 12 . At the same time, we also performed Grad-CAM visualization tests. In the visualization tests, we compared the results between the baseline and the proposed method, which are shown in Figure 13.

Combining the data in Table 2 with the results in Figures 11–13 for analysis, the detection results and the data results can corroborate each other in theory. In Table 2, YOLO X and Cascade R-CNN have a relatively high recall. According to Equation (11), we believe that higher Recall means that fewer missed detections occur, so there are fewer missed detections in the detection results of YOLO X. However, the Precision value is lower in comparison to the proposed method. According to Equation (12), the lower Precision is reflected in the detection results as more false detection. Figures 11 and 12 can well confirm this view: YOLO X is able to detect all the Ground truth but simultaneously generates a large number of false detections. In comparison with the state-of-the-art method, YOLO V8 detected the target correctly in some scenes as well as the proposed method, but the confidence score of the proposed method was much higher than that of YOLO V8. In Figure 11g,h, DSF-Net exhibits a confidence level approximately 2.7 times higher than YOLO V8 in the top right corner of the first image and the top left corner of the last image. This phenomenon is particularly notable in scenarios featuring densely distributed small SAR ship targets. Therefore, the proposed method in this paper effectively reduces the cases of false detection and missed detection and performs favorably in detecting small SAR ship targets.

4.4.2. Ablation Experiments

To demonstrate that each module of DSF-Net plays an active role in the detection of the small SAR ship targets, we carried out ablation experiments on the LS-SSDDv1.0 dataset. The experimental results are shown in Table 3.

Table 3. The ablation experiment on LS-SSDD-v1.0 [52] of the entire scenes.

YOLO v5 (Baseline)	PWSA&NLSA	TRF-SPP	R-Tradeoff Loss	P (%)	R (%)	mAP_{50} (%)	mAP_{50-95} (%)	F_1 (%)	GFLOPs
✓				84	63.6	73.3	27.1	72	15.9
✓	✓			85.1 (+1.1)	66.5 (+2.9)	75.5 (+2.2)	28.9 (+1.8)	75 (+3)	22.1
✓	✓	✓		85.5 (+1.5)	67.6 (+4.0)	76.1 (+2.8)	29.2 (+2.1)	76 (+4)	33.5
✓	✓	✓	✓	**86.4** (+2.4)	**68.7** (+5.1)	**76.9** (+3.6)	**29.4** (+2.3)	**77** (+5)	33.5
✓		✓		84.1 (+0.1)	66.4 (+2.8)	75.5 (+2.2)	29.0 (+1.9)	74 (+2)	27.2
✓			✓	84.3 (+0.3)	68.5 (+4.9)	76.2 (+2.9)	28.9 (+1.8)	76 (+4)	15.8

The bold font represents the best results among all methods.

By analyzing the experimental results in Table 3, we can mutually corroborate with the theory. The placement of PWSA in the Backbone and NLSA in the Bridge Node expands the extraction capability of the Backbone for small target features and increases the capture of long-range dependencies. Hence, there is a certain degree of improvement in both precision and recall. The role of the TRF-SPP module is to enhance the network's capability to capture contextual information and reduce the likelihood of missed detections, so the recall will be improved in theory. From the experimental data, the recall is improved by 6.3% compared to the baseline model after adding TRF-SPP to PWSA&NLSA. R-tradeoff Loss makes a tradeoff between CIOU loss and NWD loss without increasing the computational effort. The ability of CIOU loss to prevent false detection is retained as much as possible; concurrently, the R-tradeoff loss improves the anti-missed detection ability for small targets. From the experimental results, it can be seen that the precision is improved by 2.8%, the recall is improved by 8%, and the F1 score is improved by 6.9% compared to the baseline model. Additionally, there is no change in GFLOPs before and after replacing the loss with R-tradeoff Loss. To further verify that the proposed method can extract features more effectively, we represent the features captured in the inference as the heat maps. The results of the heat maps are shown in Figure 10. The color closer to red indicates that the network pays more attention to these features, and the color closer to blue indicates that the network pays less attention to these features.

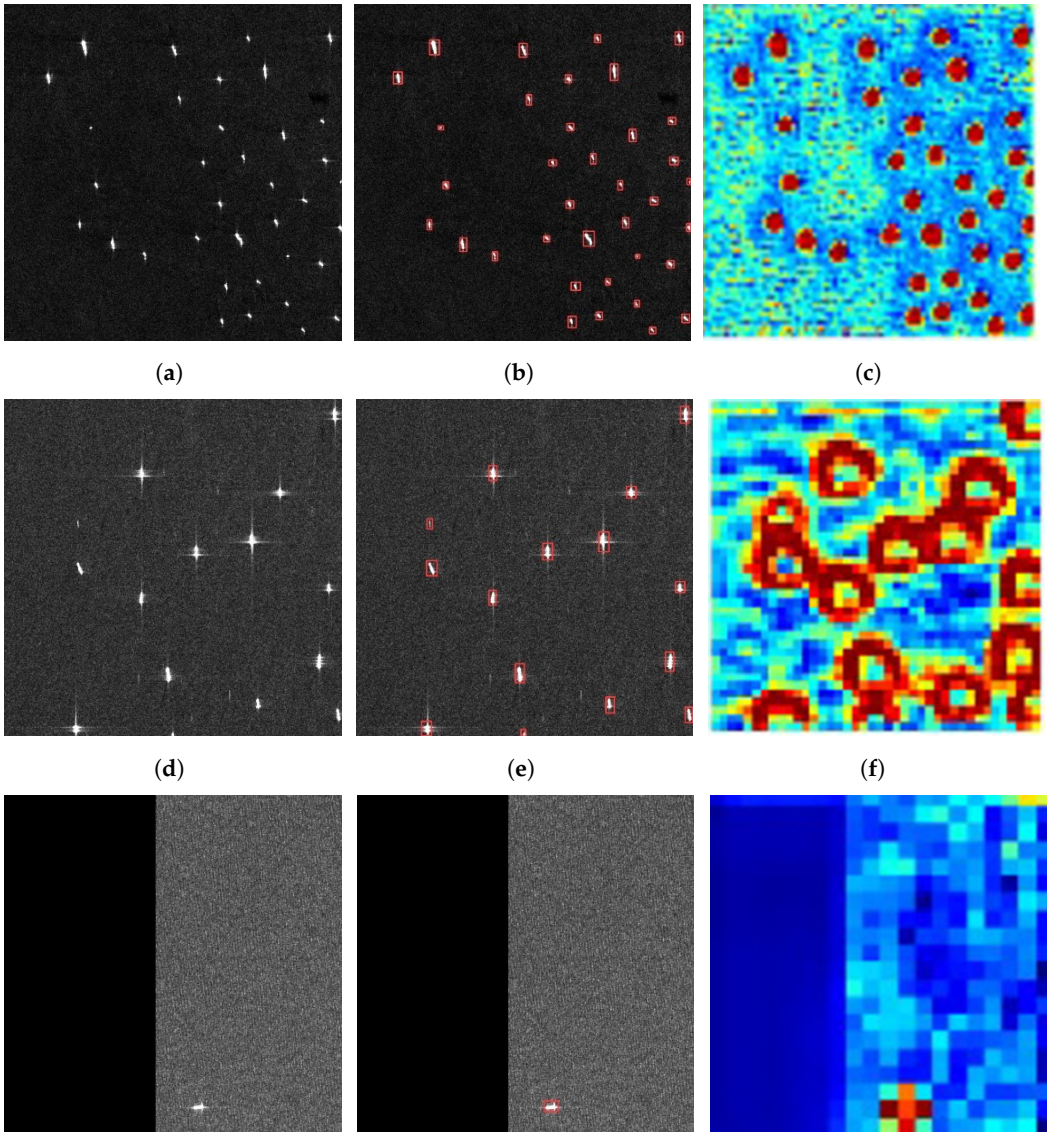
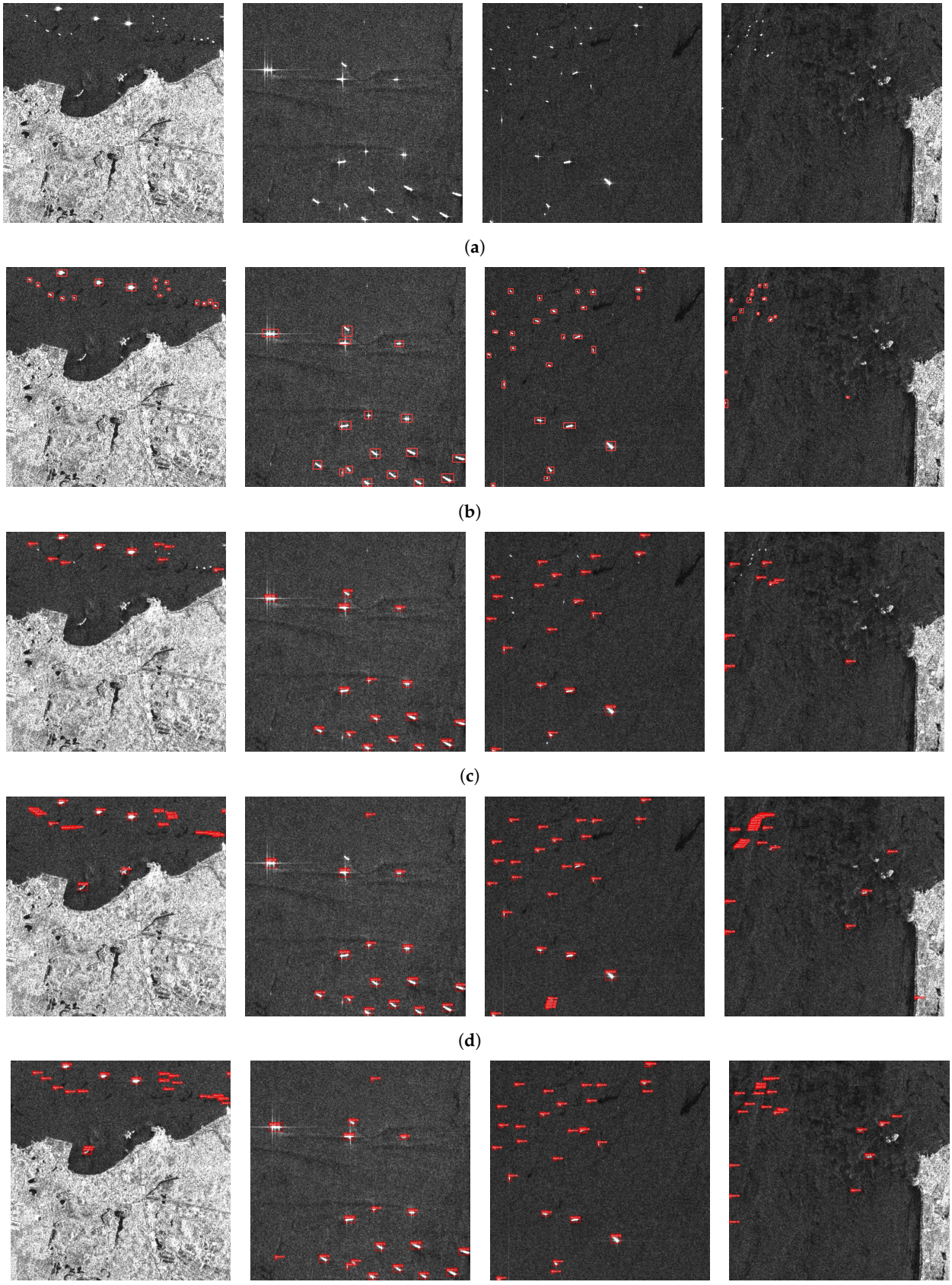


Figure 10. The results of feature map visualization. (**a**,**d**,**g**) are the original images; (**b**,**e**,**h**) are Ground Truth; (**c**,**f**,**i**) are the feature maps in the form of heat maps.

Some of the common scenes in small SAR ship target detection are shown in Figure 10. Among Figure 10, Figure 10a is the densely arranged ship scenes, Figure 10d is the small ship target scenes with coherent bright spots, and Figure 10g is the small ship target detection scenes with the imaging clutter and black edge. As can be seen from the figures, the proposed DSF-Net can better focus on the small ship targets for all the above scenes, and the features of different dimensions have excellent extraction effects.

To verify our assumptions about the network structure and our assertions about the Bridge Node and loss function, we conducted a series of contrast experiments. First, to demonstrate the importance of PWSA and NLSA modules, we replace PWSA and NLSA in the same position with mainstream attention modules and conduct comparative experiments. The results are shown in Table 4.



(**a**)

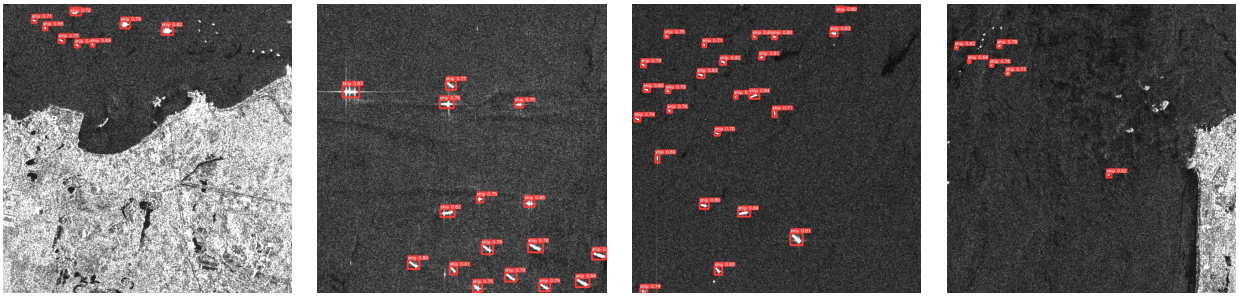
(**b**)

(**c**)

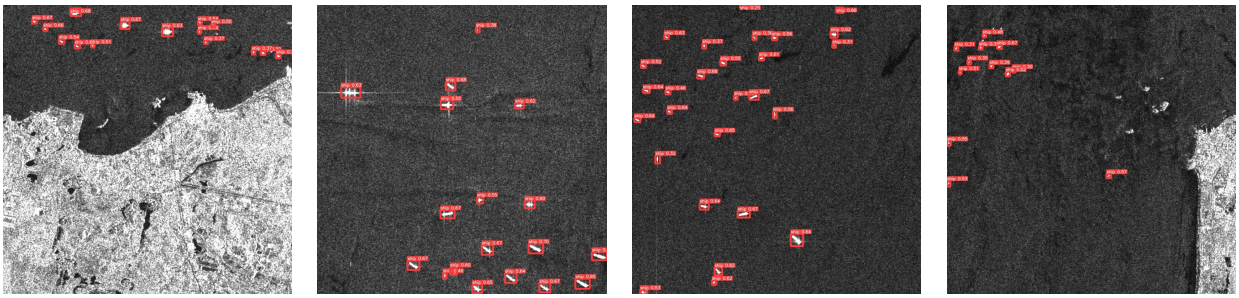
(**d**)

(**e**)

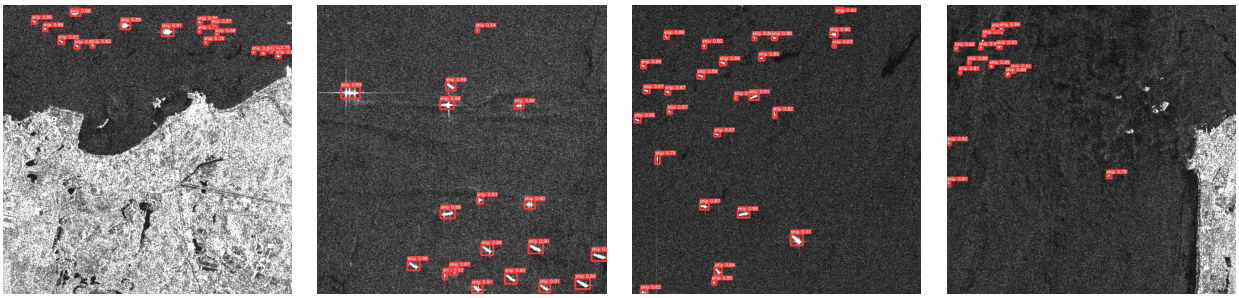
Figure 11. *Cont.*



(**f**)

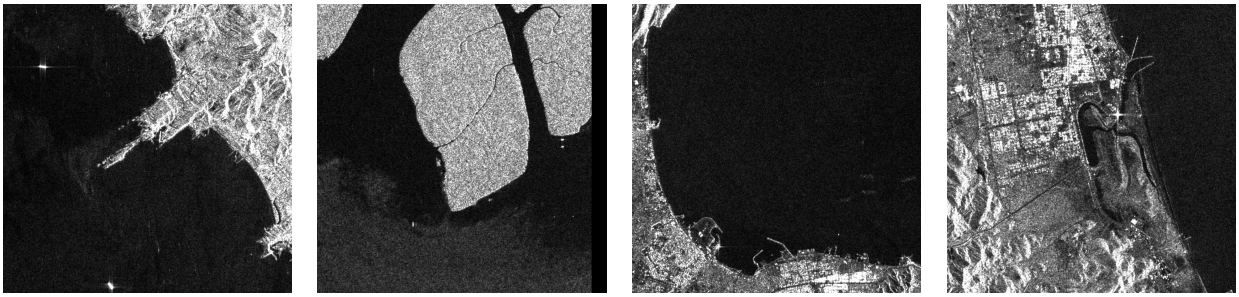


(**g**)



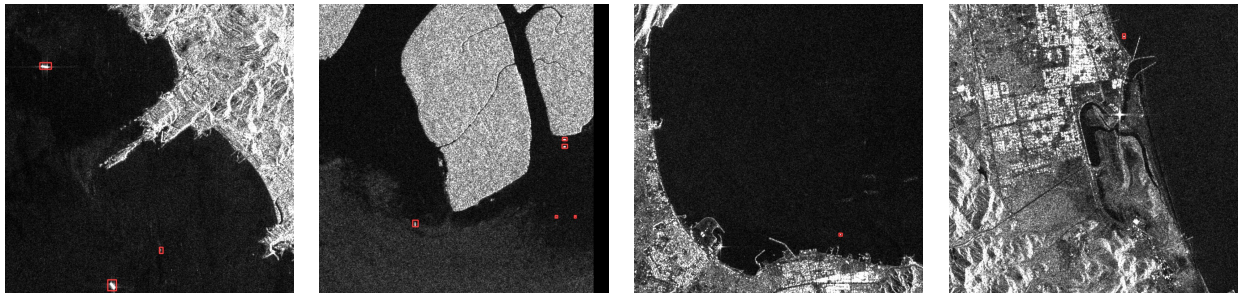
(**h**)

Figure 11. Dense scene of small SAR ship target detection results comparison, where (**a**) are the original images, (**b**) are the Ground truth, (**c**) are the results of Cascade R-CNN [54], (**d**) are the results of RetinaNet [63], (**e**) are the results of YOLO X [19], (**f**) are the results of YOLO V5 [39], (**g**) are the results of YOLO V8 [55] and (**h**) are the results of DSF-Net(ours).

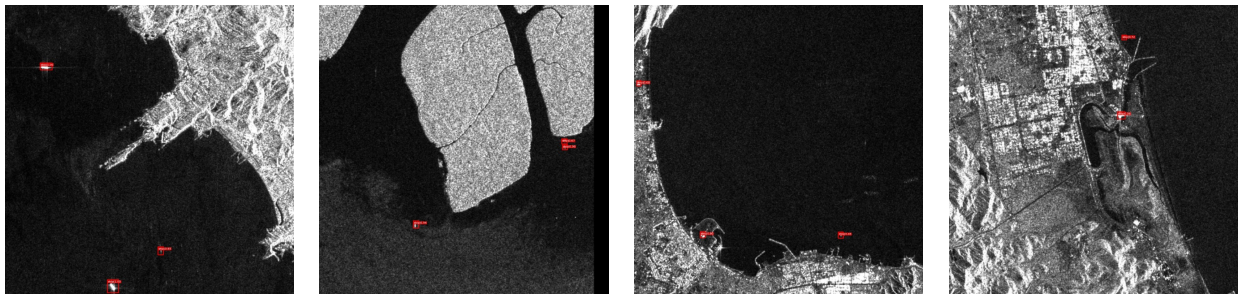


(**a**)

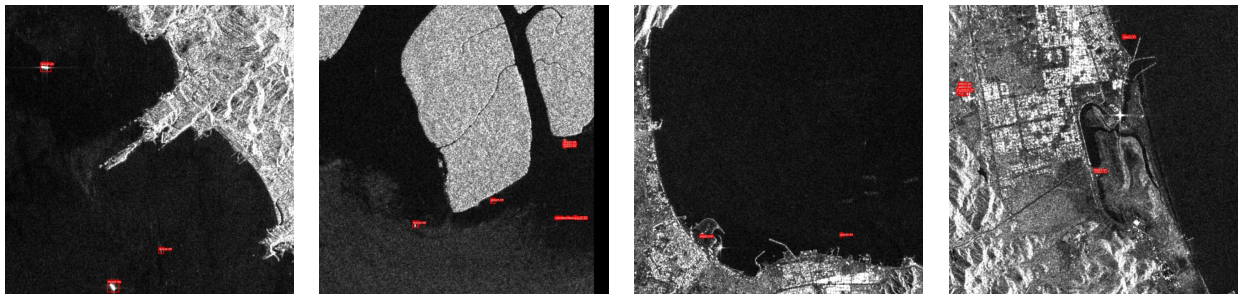
Figure 12. *Cont.*



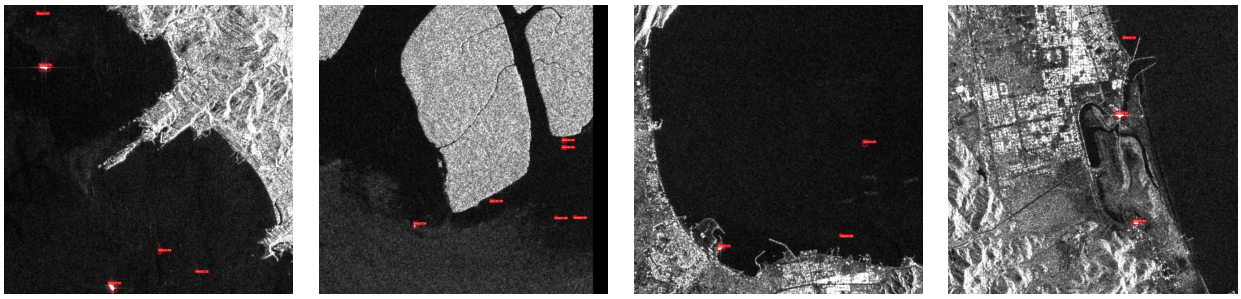
(**b**)



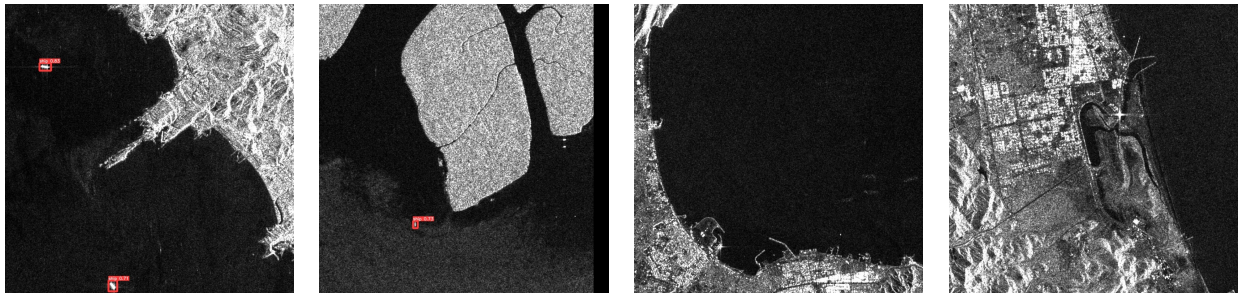
(**c**)



(**d**)

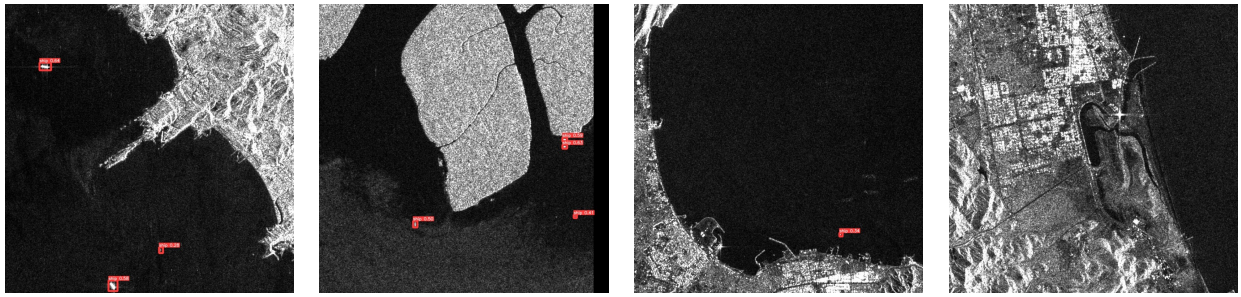


(**e**)

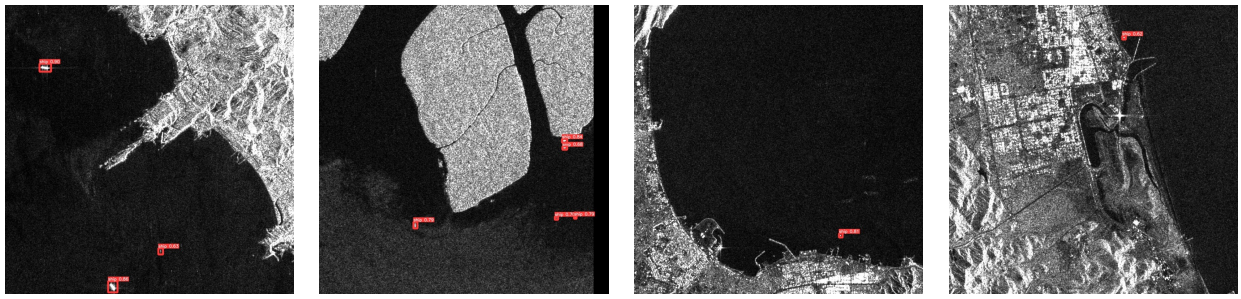


(**f**)

Figure 12. *Cont.*

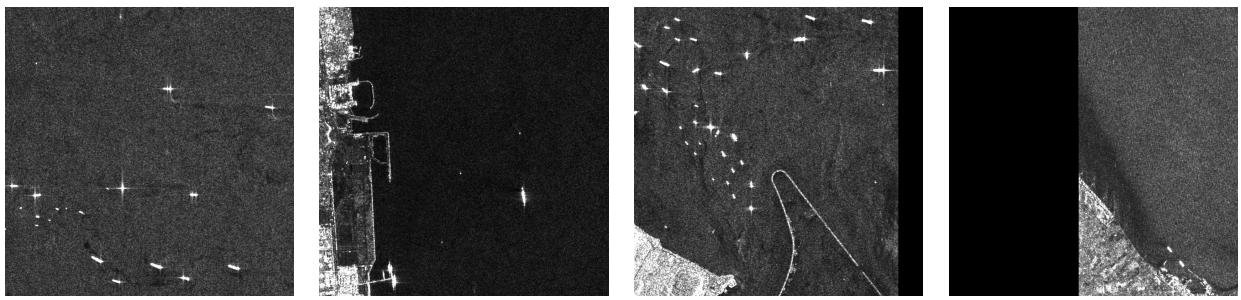


(**g**)

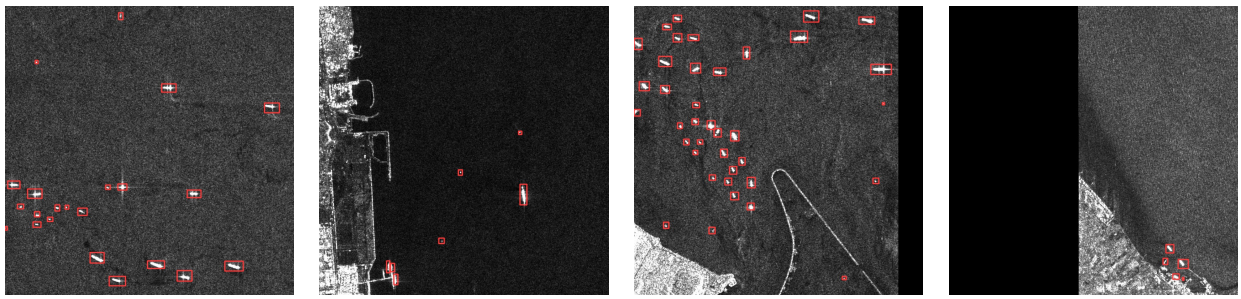


(**h**)

Figure 12. Comparison of small SAR ship target detection results for inshore scenes, where (**a**) are the original images, (**b**) are the Ground truth, (**c**) are the results of Cascade R-CNN [54], (**d**) are the results of RetinaNet [63], (**e**) are the results of YOLO X [19], (**f**) are the results of YOLO V5 [39], (**g**) are the results of YOLO V8 [55] and (**h**) are the results of DSF-Net(ours).

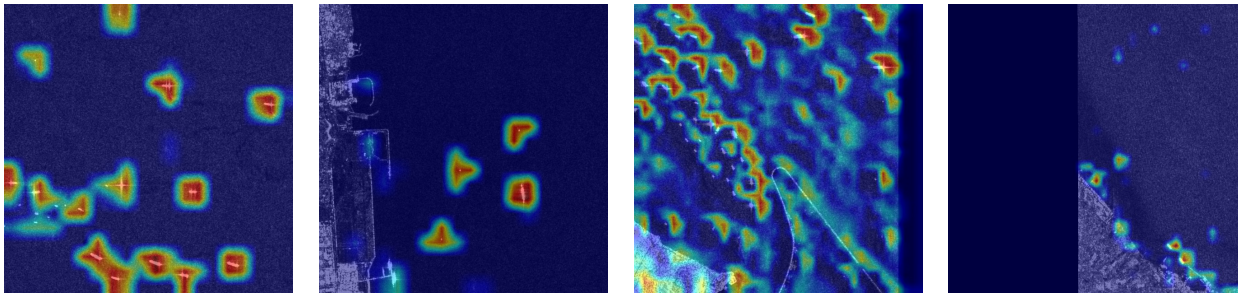


(**a**)

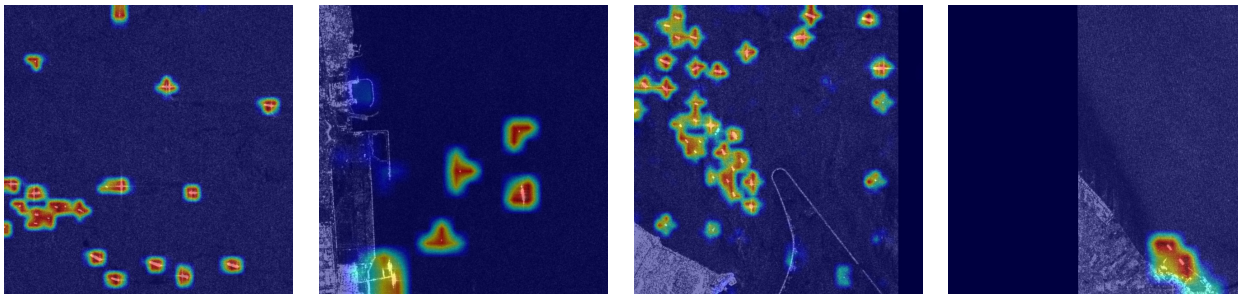


(**b**)

Figure 13. *Cont.*



(**c**)



(**d**)

Figure 13. The results of Grad-CAM between the baseline and our DSF-Net, where (**a**) are the original images, (**b**) are the Ground truth, (**c**) are the Grad-CAM of YOLO V5(baseline) [39], (**d**) are the Grad-CAM of DSF-Net(ours).

Based on the data presented in Table 4, it is evident that the networks incorporating the attention module exhibit a certain degree of enhancement in recall. The result indicates that attention is indeed effective in solving the missed detection of small targets and can enhance the Backbone's ability to extract features of small targets. However, in terms of precision, the networks with other attention modules decreased the precision to different degrees, which is consistent with our view on the stability of the Bridge Node. We believe that the instability in the feature fusion process is responsible for the loss of certain features, resulting in a decrease in precision. In addition, the F1 scores for SE Attention and Shuffle Attention have demonstrated improvements compared to the baseline model. Conversely, the F1 scores of other attention mechanisms have declined, further affirming the efficacy of the Squeeze operation and Channel shuffle operation in enhancing the extraction of features related to small targets.

Table 4. The results of the experiment comparing different attention modules.

Method	P (%)	R (%)	mAP_{50} (%)	mAP_{50-95} (%)	F_1 (%)	GFLOPs
YOLO v5	84	63.6	73.3	27.1	72	15.9
PWSA&NLSA	**85.1**	**66.5**	**75.5**	**28.9**	**75**	22.1
SE Attention [26]	81.9	65.6	72.5	26.3	73	15.8
Shuffle Attention [35]	83.7	64.5	72.1	26.9	73	15.8
CBAM Attention [31]	79.9	65.8	72.4	27	72	15.9
ECA Attention [64]	80.7	65.1	72	26.5	72	15.8
Coord Attention [65]	79.5	64.9	71.6	26.9	71	15.8

To further investigate the effectiveness of our designed R-tradeoff Loss, we conducted controlled experiments for different R values based on DSF-Net. According to the experimental results, We selected the most suitable R-value. The experimental results are shown in Table 5.

Table 5. The Result of each metric under different neutralization factor R.

R	P (%)	R (%)	mAP_{50} (%)	mAP_{50-95} (%)	F_1 (%)	GFLOPs
0.1	85.2	70.8	76.5	28.6	77	33.5
0.2	85.2	**71.3**	**76.9**	29.2	**78**	33.5
0.3	84.7	69.1	75.8	28.4	76	33.5
0.5	**86.4**	68.7	**76.9**	**29.4**	77	33.5
0.7	83.2	67	75.6	26.7	74	33.5
0.9	82.2	66.9	74.9	28.1	74	33.5
1.0	85.5	67.6	76.1	29.2	76	33.5

The bold font represents the best results among all methods.

From the data in Table 5, it can be concluded that when the value of R is small, the recall will be higher. The NWD loss dominates the loss function at this time, which is consistent with our view that NWD will decrease the missed detection of small ship targets. Additionally, it can be seen from Equation (11) that a decrease in FP will result in an improvement in Recall. With the increase of R, the advantage of CIOU gradually appears, FN starts to fall, and Precision starts to rise. Moreover, the relationship between R, Recall, and Precision is not linear. After considering multiple metrics, we chose R = 0.5 as the result after adding R-tradeoff Loss. Both ablation experiments and R-factor control experiments demonstrate the superior performance of the R-tradeoff loss for small-scale SAR ship target detection. The original intention behind the design of the R-tradeoff loss is to strike a balance between detection accuracy and the ability to resist missed detections in small-scale ship targets. These aspects are reflected in the metrics P and R, respectively. The F1 score, as a comprehensive metric combining both P and R, is highly suitable for evaluating the performance of the R-tradeoff loss. Based on the data presented in Table 5, it is evident that F1 scores exhibit varying degrees of improvement when R values are within the smaller range. The highest improvement, amounting to 6 percentage points compared to the baseline model, is consistent with our initial design intent.

To comprehensively evaluate the algorithm proposed in this paper, we conducted an analysis of its complexity from two perspectives: GFLOPs and Params, following the findings presented in reference [53,66].

Low complexity and high performance are often difficult to achieve simultaneously in algorithms. Significant performance improvements typically come with a substantial increase in algorithm complexity. However, DSF-Net, proposed in this paper, achieves a notable improvement in performance with only a slight increase in complexity. Compared to the baseline model, with an increase of 17.6 GFLOPs, the growth in the F1 score is 6.9 times the original. Compared to YOLO V8, various metrics show different degrees of improvement, with GFLOPs increasing by only 5.1. At the same time, we compared the change in single-layer GFLOPs after adding PWSA and NLSA. The shallowest PWSA module has a time of 0.5 ms and GFLOPs of 2.11; the shallowest NLSA module has a time of 0.5ms and GFLOPs of 0.02; while the corresponding layer's CSP module has a time of 0.9 ms and GFLOPs of 2.01. This indicates that the introduction of PWSA and NLSA in this paper adds complexity similar to that of a single CSP module yet delivers performance surpassing that of the CSP module.

The comparison of Parameters is shown in Table 6. It can be observed from this table that DSF-Net, proposed in this paper, achieves comparable or even superior results when compared to other methods with similar parameter counts. The F1 score is improved by up to 29.67% compared to previous methods, and the parameters are in a similar range.

Table 7 presents a comparison of model sizes between the proposed method and common two-stage and one-stage methods. In this comparison, when compared to the advanced YOLO-SD, our method exhibits a slightly lower mAP_{50-95} by 0.3%, a notably higher mAP_{50} by 2.5%, and requires 10.1 MB fewer parameters. It is worth noting that YOLO-SD has undergone pretraining. In summary, the DSF-Net proposed in this paper achieves a good balance between model performance and algorithm complexity.

Table 6. The comparison between our method and previous approaches in terms of parameters.

Method	P (%)	R (%)	F_1 (%)	Parameters
SSD [67]	43.05	52.57	47.33	2.37×10^7
RetinaNet [63]	50.75	57.87	54.51	3.57×10^7
RefineDet [62]	66.72	70.23	68.43	4.29×10^7
YOLOX [19]	66.78	**75.44**	70.85	2.53×10^7
DSF-Net(ours)	**86.4**	68.7	**77**	2.45×10^7

Table 7. The comparison between our method and previous approaches in terms of model size, where * indicates pre-trained.

Method	mAP_{50-95} (%)	mAP_{50} (%)	Params (MB)
Libra Faster R-CNN [68]	25.1	65.9	41.39
Mask R-CNN [69]	27.1	70.1	43.75
Dynamic R-CNN [70]	26.7	69.6	41.12
Grid R-CNN [71]	27.2	70.3	64.24
YOLOF [72]	16.6	50.7	42.06
YOLOX(L) [19]	28.3	72.0	54.15
YOLO-SD * [66]	**29.7**	74.4	59.60
DSF-Net(ours)	29.4	**76.9**	49.50

5. Conclusions

In this study, we present a novel network architecture that combines dual feature shuffle guidance for small SAR ship target feature extraction and multi-field feature fusion. Our approach aims to tackle the challenges of missed and false detection in the process of small SAR ship target detection.

From the perspectives of feature extraction capability and structure stability, we propose two kinds of shuffle-guided attention modules. At the same time, we propose the Bridge Node and feature dilution hypothesis and verify them by contrast experiments. Rethinking the multi-receptive field enhancement, we propose the TRF-SPP module, which improves the recall by 6.3% compared with the baseline model. We further investigate the sensitivity of the loss for different scale targets and present the R-tradeoff loss to achieve the tradeoff between different scales. The mAP_{50-95} increased by 8.5%, and the F1 score increased by 6.9% compared with the baseline model. Our future work will focus on the weakly supervised method for small SAR ship target detection.

Author Contributions: Conceptualization, J.Z.; Methodology, J.Z.; Software, J.Z.; Validation, J.Z.; Formal analysis, J.Z.; Investigation, Z.X. and J.Z.; Resources, Z.X. and J.Z.; Data curation, Z.X., J.Z. and K.L.; Writing—original draft, J.Z.; Writing—review & editing, Z.X., J.Z., K.H. and K.L.; Visualization, J.Z.; Supervision, Z.X. and K.H.; Project administration, J.Z.; Funding acquisition, Z.X. and K.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number No. 62271303; Pujiang Talents Plan, grant number No. 22PJD029.

Data Availability Statement: Not applicable.

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

References

1. Zhang, L.; Liu, Y.; Zhao, W.; Wang, X.; Li, G.; He, Y. Frequency-Adaptive Learning for SAR Ship Detection in Clutter Scenes. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5215514. [CrossRef]
2. Zhang, C.; Yang, C.; Cheng, K.; Guan, N.; Dong, H.; Deng, B. MSIF: Multisize Inference Fusion-Based False Alarm Elimination for Ship Detection in Large-Scale SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5224811. [CrossRef]
3. Li, J.; Xu, C.; Su, H.; Gao, L.; Wang, T. Deep learning for SAR ship detection: Past, present and future. *Remote Sens.* **2022**, *14*, 2712. [CrossRef]

4. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. *arXiv* **2023**, arXiv:2304.02643.
5. Singh, P.; Shree, R. A new homomorphic and method noise thresholding based despeckling of SAR image using anisotropic diffusion. *J. King Saud-Univ.-Comput. Inf. Sci.* **2020**, *32*, 137–148. [CrossRef]
6. Zhou, Y.; Liu, H.; Ma, F.; Pan, Z.; Zhang, F. A Sidelobe-Aware Small Ship Detection Network for Synthetic Aperture Radar Imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5205516. [CrossRef]
7. Ai, J.; Luo, Q.; Yang, X.; Yin, Z.; Xu, H. Outliers-robust CFAR detector of Gaussian clutter based on the truncated-maximum-likelihood-estimator in SAR imagery. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 2039–2049. [CrossRef]
8. Leng, X.; Ji, K.; Yang, K.; Zou, H. A bilateral CFAR algorithm for ship detection in SAR images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1536–1540. [CrossRef]
9. Hou, B.; Chen, X.; Jiao, L. Multilayer CFAR detection of ship targets in very high resolution SAR images. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 811–815.
10. Renga, A.; Graziano, M.D.; Moccia, A. Segmentation of marine SAR images by sublook analysis and application to sea traffic monitoring. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 1463–1477. [CrossRef]
11. Copeland, A.C.; Ravichandran, G.; Trivedi, M.M. Localized Radon transform-based detection of ship wakes in SAR images. *IEEE Trans. Geosci. Remote Sens.* **1995**, *33*, 35–45. [CrossRef]
12. Karakus, O.; Rizaev, I.; Achim, A. Ship Wake Detection in SAR Images via Sparse Regularization. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 1665–1677. [CrossRef]
13. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, 111, 257–276. [CrossRef]
14. Ren, S.; He, K.; 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] [PubMed]
15. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
16. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
17. 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. *arXiv* **2022**, arXiv:2207.02696.
18. Xu, S.; Wang, X.; Lv, W.; Chang, Q.; Cui, C.; Deng, K.; Wang, G.; Dang, Q.; Wei, S.; Du, Y.; et al. PP-YOLOE: An evolved version of YOLO. *arXiv* **2022**, arXiv:2203.16250.
19. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
21. Zhou, Y.; Fu, K.; Han, B.; Yang, J.; Pan, Z.; Hu, Y.; Yin, D. D-MFPN: A Doppler Feature Matrix Fused with a Multilayer Feature Pyramid Network for SAR Ship Detection. *Remote Sens.* **2023**, *15*, 626. [CrossRef]
22. Ma, X.; Hou, S.; Wang, Y.; Wang, J.; Wang, H. Multiscale and dense ship detection in SAR images based on key-point estimation and attention mechanism. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5221111. [CrossRef]
23. Xia, R.; Chen, J.; Huang, Z.; Wan, H.; Wu, B.; Sun, L.; Yao, B.; Xiang, H.; Xing, M. CRTransSar: A visual transformer based on contextual joint representation learning for SAR ship detection. *Remote Sens.* **2022**, *14*, 1488. [CrossRef]
24. Guo, Y.; Chen, S.; Zhan, R.; Wang, W.; Zhang, J. LMSD-YOLO: A Lightweight YOLO Algorithm for Multi-Scale SAR Ship Detection. *Remote Sens.* **2022**, *14*, 4801. [CrossRef]
25. Zhang, T.; Zhang, X. Squeeze-and-excitation Laplacian pyramid network with dual-polarization feature fusion for ship classification in sar images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 4019905. [CrossRef]
26. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
27. Su, N.; He, J.; Yan, Y.; Zhao, C.; Xing, X. SII-Net: Spatial information integration network for small target detection in SAR images. *Remote Sens.* **2022**, *14*, 442. [CrossRef]
28. Chen, P.; Li, Y.; Zhou, H.; Liu, B.; Liu, P. Detection of small ship objects using anchor boxes cluster and feature pyramid network model for SAR imagery. *J. Mar. Sci. Eng.* **2020**, *8*, 112. [CrossRef]
29. Zhang, L.; Liu, Y.; Qu, L.; Cai, J.; Fang, J. A Spatial Cross-Scale Attention Network and Global Average Accuracy Loss for SAR Ship Detection. *Remote Sens.* **2023**, *15*, 350. [CrossRef]
30. Xu, Z.; Gao, R.; Huang, K.; Xu, Q. Triangle Distance IoU Loss, Attention-Weighted Feature Pyramid Network, and Rotated-SARShip Dataset for Arbitrary-Oriented SAR Ship Detection. *Remote Sens.* **2022**, *14*, 4676. [CrossRef]
31. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
32. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
33. Cui, Z.; Wang, X.; Liu, N.; Cao, Z.; Yang, J. Ship detection in large-scale SAR images via spatial shuffle-group enhance attention. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 379–391. [CrossRef]

34. Gao, X.; Xu, L.; Wang, F.; Hu, X. Multi-branch aware module with channel shuffle pixel-wise attention for lightweight image super-resolution. *Multimed. Syst.* **2023**, *29*, 289–303. [CrossRef]
35. Zhang, Q.L.; Yang, Y.B. Sa-net: Shuffle attention for deep convolutional neural networks. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA; pp. 2235–2239.
36. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
37. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
39. Jocher, G.; Nishimura, K.; Mineeva, T.; Vilariño, R. YOLOv5 by Ultralytics. Code Repository. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 22 December 2022).
40. Liu, S.; Huang, D.; Wang, Y. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
41. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
42. Wang, J.; Lin, Y.; Guo, J.; Zhuang, L. SSS-YOLO: Towards more accurate detection for small ships in SAR image. *Remote Sens. Lett.* **2021**, *12*, 93–102. [CrossRef]
43. Yang, S.; An, W.; Li, S.; Zhang, S.; Zou, B. An Inshore SAR Ship Detection Method Based on Ghost Feature Extraction and Cross-Scale Interaction. *IEEE Geosci. Remote Sens. Lett.* **2023**, *19*, 4019905. [CrossRef]
44. Liu, S.; Kong, W.; Chen, X.; Xu, M.; Yasir, M.; Zhao, L.; Li, J. Multi-scale ship detection algorithm based on a lightweight neural network for spaceborne SAR images. *Remote Sens.* **2022**, *14*, 1149. [CrossRef]
45. Hong, Z.; Yang, T.; Tong, X.; Zhang, Y.; Jiang, S.; Zhou, R.; Han, Y.; Wang, J.; Yang, S.; Liu, S. Multi-scale ship detection from SAR and optical imagery via a more accurate YOLOv3. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6083–6101. [CrossRef]
46. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
47. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI conference on artificial intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
48. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
49. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740.
50. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized Gaussian Wasserstein distance for tiny object detection. *arXiv* **2021**, arXiv:2110.13389.
51. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
52. 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]
53. Du, Y.; Du, L.; Guo, Y.; Shi, Y. Semi-Supervised SAR Ship Detection Network via Scene Characteristic Learning. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5201517. [CrossRef]
54. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
55. Jocher, G.; Nishimura, K.; Mineeva, T.; Vilariño, R. YOLOv8 by Ultralytics. Code Repository. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 15 July 2023).
56. Zhang, L.; Wang, H.; Wang, L.; Pan, C.; Huo, C.; Liu, Q.; Wang, X. Filtered Convolution for Synthetic Aperture Radar Images Ship Detection. *Remote Sens.* **2022**, *14*, 5257. [CrossRef]
57. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region proposal by guided anchoring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2965–2974.
58. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; Shi, J. Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.* **2020**, *29*, 7389–7398. [CrossRef]
59. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 9627–9636.
60. Zhang, X.; Huo, C.; Xu, N.; Jiang, H.; Cao, Y.; Ni, L.; Pan, C. Multitask learning for ship detection from synthetic aperture radar images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8048–8062. [CrossRef]

61. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9759–9768.
62. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212.
63. 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.
64. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11534–11542.
65. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
66. Wang, S.; Gao, S.; Zhou, L.; Liu, R.; Zhang, H.; Liu, J.; Jia, Y.; Qian, J. YOLO-SD: Small Ship Detection in SAR Images by Multi-Scale Convolution and Feature Transformer Module. *Remote Sens.* **2022**, *14*, 5268. [CrossRef]
67. 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 Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
68. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra r-cnn: Towards balanced learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and PATTERN Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 821–830.
69. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
70. Zhang, H.; Chang, H.; Ma, B.; Wang, N.; Chen, X. Dynamic R-CNN: Towards high quality object detection via dynamic training. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 260–275.
71. Lu, X.; Li, B.; Yue, Y.; Li, Q.; Yan, J. Grid r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7363–7372.
72. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13039–13048.