remote sensing

MDPI

Article

An Anchor-Free Detection Algorithm for SAR Ship Targets with Deep Saliency Representation

Jianming Lv [1,2,3], Jie Chen [1,2,*], Zhixiang Huang [1,2], Huiyao Wan [1,2,3], Chunyan Zhou [1], Daoyuan Wang [4], Bocai Wu [3] and Long Sun [3]

[1] Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui University, Hefei 230093, China
[2] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Electronics and Information Engineering, Anhui University, Hefei 230093, China
[3] The 38th Research Institute of China Electronics Technology Group Corporation, Hefei 210039, China
[4] State Grid Anhui Electric Power Co., Ltd., Hefei 230061, China
* Correspondence: jiechen@ustc.edu

Abstract: Target detection in synthetic aperture radar (SAR) images has a wide range of applications in military and civilian fields. However, for engineering applications involving edge deployment, it is difficult to find a suitable balance of accuracy and speed for anchor-based SAR image target detection algorithms. Thus, an anchor-free detection algorithm for SAR ship targets with deep saliency representation, called SRDet, is proposed in this paper to improve SAR ship detection performance against complex backgrounds. First, we design a data enhancement method considering semantic relationships. Second, the state-of-the-art anchor-free target detection framework CenterNet2 is used as a benchmark, and a new feature-enhancing lightweight backbone, called LWBackbone, is designed to reduce the number of model parameters while effectively extracting the salient features of SAR targets. Additionally, a new mixed-domain attention mechanism, called CNAM, is proposed to effectively suppress interference from complex land backgrounds and highlight the target area. Finally, we construct a receptive-field-enhanced detection head module, called RFEHead, to improve the multiscale perception performance of the detection head. Experimental results based on three large-scale SAR target detection datasets, SSDD, HRSID and SAR-ship-dataset, show that our algorithm achieves a better balance between ship target detection accuracy and speed and exhibits excellent generalization performance.

Keywords: anchor-free; synthetic aperture radar (SAR); ship detection; deep saliency representation

Citation: Lv, J.; Chen, J.; Huang, Z.; Wan, H.; Zhou, C.; Wang, D.; Wu, B.; Sun, L. An Anchor-Free Detection Algorithm for SAR Ship Targets with Deep Saliency Representation. *Remote Sens.* **2023**, *15*, 103. https://doi.org/10.3390/rs15010103

Academic Editors: Xinghua Li, Fan Zhang, Bo Tang, Wei Yao, Zhongling Huang and Zongxu Pan

Received: 27 October 2022
Revised: 17 December 2022
Accepted: 20 December 2022
Published: 24 December 2022

1. Introduction

Synthetic aperture radar (SAR) has the advantage of all-day and all-weather detection capabilities. Due to its unique imaging principle, SAR also has the advantages of a strong penetration ability and a strong anti-interference ability. As a ground target observation technology, SAR can observe ships over a wide range and field of view. SAR imaging can be used to overcome the limitations of optical imaging under adverse weather and illumination conditions and can still observe ground object information in harsh environments; consequently, it is more suitable for ship monitoring [1–3]. As the technology has developed, SAR imaging has been widely adopted in many fields, such as military applications, marine traffic control, fishery management and trade activities [4].

However, for application to real scenes, SAR ship target detection still faces some challenges [5–7], such as the influence of complex surroundings, multiscale targets and target defocusing, all of which affect performance in detecting ships. In particular, the speckle noise in SAR images hinders the fine interpretation of ground objects. This noise leads to complex backgrounds and prevents SAR images from correctly reflecting the scattering characteristics of ground objects. Due to the use of multiresolution imaging

modes and the existence of a variety of ship shapes, the sizes of ship targets can also vary greatly; small ship targets are especially difficult to accurately detect and some false detection results are possible, thus degrading detection performance. At the same time, the generalization ability of existing algorithms is weak, meaning that their performance on other similar datasets is unsatisfactory.

To improve the sophistication of the interpretation of ship targets in SAR images, researchers have developed a series of algorithms for ship target detection in SAR images, mainly including traditional machine learning methods and deep learning-based methods. Traditional machine learning methods mainly rely on expert knowledge and experience for the manual selection of representative features to achieve ship target detection. However, such methods have weak generalization performance and limited accuracy in complex and diverse remote sensing application scenarios.

In recent years, deep learning-based methods have attracted extensive attention. Due to its powerful automatic feature extraction capability, deep learning has been widely and maturely applied for object detection in optical images of natural scenes and has achieved high detection performance on representative large-scale datasets of such images, such as COCO and Pascal VOC. In this context, many research teams have attempted to extend deep learning methods to the SAR ship target detection task and have achieved good research results. Depending on whether anchors are used, detection methods based on deep learning can be divided into anchor-based methods and anchor-free methods [8–10].

In an anchor-based detection method, it is necessary to set an a priori anchor size, then filter the anchors in accordance with the actual target characteristics to perform classification and regression. However, due to the multiscale characteristics of ship targets on the sea, any a priori anchor size set in an anchor-based method will have difficulty covering all ship sizes. Therefore, anchor-based target detection methods usually produce a large number of false positives, especially for small-scale ship targets, and this shortcoming greatly affects the detection performance.

In an anchor-free detection method, the size of each target is directly predicted without being limited by anchors and such methods have many application prospects in SAR target detection. Anchor-free detection algorithms avoid the need for complex parameter settings, produce a markedly reduced number of false candidates, require fewer model parameters and are more suitable for real-time inference and embedded edge deployment. Nevertheless, in view of the characteristic properties of SAR ship targets, anchor-free detection methods for SAR ship target detection are still in the preliminary exploration stage and have considerable room for improvement. Thus, this paper combines the advantages of anchor-free detection algorithm and two-stage detection algorithm. A novel detection algorithm for SAR ship targets with deep saliency representation called SRDet, which improves the performance of SAR target detection against complex backgrounds in terms of both speed and accuracy, is innovatively proposed in this paper. The primary contributions of this paper are as follows:

(1) To address the problems of a small number of SAR ship target samples and a large distribution of small and weak targets, a copy–paste data enhancement method that considers number of samples of SAR targets to support effective training of deep models and reduce overfitting.
(2) A lightweight anchor-free target detection network is constructed. We first introduce the state-of-the-art (SOTA) anchor-free detection framework CenterNet2 as the benchmark network and we then design a new lightweight backbone called LWBackbone, which can effectively increase the detection accuracy with fewer parameters and an improved inference speed.
(3) To suppress the influence of complex land background interference, unclear target edges, and multiscale effects, we propose a new mixed-domain attention mechanism called CNAM to suppress the interference from complex land backgrounds and focus on the ship area. In addition, considering the multiscale characteristics of SAR ship targets, we construct a receptive-field-enhanced detection head module

named RFEHead, in which the receptive field range is improved through the design of convolutions with different dilation rates to endow the detection head with better multiscale perception performance.

2. Related Work

2.1. Traditional SAR Target Detection Algorithm

Traditional SAR ship detection algorithms can be further divided into two categories: algorithms based on scattering [11] and algorithms based on multitype feature extraction [12]. These algorithms rely on the differences in the scattering properties of ships on the sea surface. Specifically, different scattering mechanisms serve as the basis for ship target detection in SAR images. Sugimoto et al. [13] proposed two different ship target detection algorithms, "optimized Pd" and "$P_T - P_S$", considering the different scattering mechanisms of ships and the sea surface. Algorithms based on multitype feature extraction distinguish ship targets from the background sea surface on the basis of their different features. These algorithms can be further divided into ship detection methods based on structural features, grayscale features and texture features.

Target detection methods based on structural features highlight the structure or shape information of the target to achieve improved accuracy. Good stability can be achieved when using such a method. However, prior information is needed and background clutter can easily cause disturbances. A typical target detection method based on grayscale features is the constant false alarm rate (CFAR) method [14–17]. In the CFAR method, the detection of target pixels is achieved by comparing the grayscale value of each single pixel against a detection threshold. The detection performance in complex scenes is typically poor. Target detection methods based on texture features consider features that reflect the properties of the image itself and can also express some characteristics of the target structure. One example of this type of method is extended fractal (EF) analysis, which relies on the grayscale information of the target image. The spatial distribution information of the gray levels is used to detect the target using the spatial difference between the energy reflected by the target and clutter. High accuracy is achieved when using this algorithm. However, it is difficult to extract the local texture features of the target. In general, traditional detection methods for ship targets in SAR images are easily interpretable, offer real-time performance and can achieve a certain detection accuracy. However, these methods rely on expert experience. Representative features are extracted manually in accordance with the characteristics of image data samples from specific scenes. In the face of complex and diverse remote sensing scenarios, it is difficult to ensure the applicability of specific manually extracted features, resulting in weak generalization and poor universality [18–21].

2.2. SAR Ship Detection Methods Based on Deep Learning

The detection effect achieved by traditional ship detection methods is often not sufficient to meet the needs of current real-time tasks. In recent years, with the continuing development of convolutional neural networks (CNNs), it has become possible to apply deep learning to realize effective target detection without the need for time-consuming and labor-intensive manual feature design. As a result, many researchers have begun to use deep learning methods for target detection. Many target detection algorithms based on CNNs have been proposed, which can be divided into two categories: (1) Anchor-based methods. The main idea is to generate multiple anchor boxes of different sizes and proportions based on the same pixel, usually by means of a region proposal network (RPN) or clustering, filter them and finally performing classification and regression. The advantage of this type of method is that prior knowledge of the target is introduced through the anchor boxes, thereby enhancing the accuracy of classification and localization. The disadvantage is that using a large number of anchors increases the computational burden. Classic anchor-based target detection networks include Faster R-CNN [22], Cascade R-CNN [23] and RetinaNet [24]. Faster R-CNN uses an RPN to generate a series of anchors, using two fully connected layers as the region-of-interest (ROI) head. Cascade R-CNN uses

three cascaded Fast R-CNN stages, each with a different positive threshold, to make the final stage more focused on localization accuracy. RetinaNet is used to classify a set of predefined sliding anchor boxes and adjust the output loss by adjusting the size to balance the foreground and background. (2) Anchor-free methods. Objects are predicted based on multiple key points or center points and corresponding boundary information, and target detection is performed directly on the image without establishing anchor boxes in advance. The network structure of an anchor-free method is more concise, and the detection speed is faster. Classical anchor-free target detection networks include CornerNet [25], FCOS [26] and ExtremeNet [27]. ExtremeNet predicts four heatmaps and center heatmaps for each category separately and predicts targets by enumerating all possible combinations of extreme points. CornerNet completely abandons the anchor concept and relies on a point detection method to identify targets for the first time. FCOS detects targets based on key points and incorporates the concept of segmentation.

With the extensive and successful application of deep learning technology in the field of natural image recognition, an increasing number of research teams have begun to apply deep learning technology for remote sensing image recognition and have achieved a series of excellent research results, superior to those of traditional ship target detection methods. Kang et al. [28] were the first to use the Faster R-CNN algorithm for object detection in SAR images. They modified the classification confidence and score and sent any detection frame with a score lower than 0.2 through CFAR training again to prevent missed detections. Fu et al. [29] proposed FBR-Net, which uses an anchor-free strategy to eliminate the influence of anchors and added an attention mechanism and an enhanced detection head to improve detection accuracy. Wang et al. [30] added a Spatial Group-wise Enhance (SGE) attention module based on CenterNet to reduce the amount of computation when faced with dense ship targets, yielding markedly improved ship detection performance. Sun et al. [31] proposed a novel few-shot learning framework named the scattering characteristics analysis network (SCAN), in which a scattering extraction module (SEM) was designed to combine the target imaging mechanism with the network. This module learns the number and distribution of the scattering points for each target type via explicit supervision. Sun et al. [32] proposed a category–position (CP) module to optimize the position regression branch features in FCOS networks. This module can improve target positioning performance in complex scenes by generating a guidance vector from the classification branch features. Yang et al. [33] proposed a one-stage ship detector with strong robustness against scale changes and various types of interference. First, a coordinate attention module (CoAM) was introduced to obtain more representative semantic features. Second, a receptive field increased module (RFIM) was designed to capture multiscale context information. Li et al. [34] proposed a new multidimensional-domain deep learning network for SAR ship detection that utilizes complementary features from the spatial domain and frequency domain. By the means of the polar Fourier transform, the rotation-invariant characteristics of a ship target are obtained in the frequency domain.

Most of the existing algorithms are oriented toward specific application requirements and higher detection accuracy; however, the computational complexity of these models is high, resulting in a slow inference speed. For military applications involving weapons targeting, such as applications based on airborne, spaceborne and missile-borne SAR imaging, there are high requirements on both the accuracy and real-time performance of target detection algorithms. If one of the existing large models is adopted, it will be difficult to suitably balance the demands for precise and real-time performance in practical engineering applications. Therefore, this paper innovatively proposes a novel SAR image ship target detection algorithm with deep saliency representation, called SRDet, which is better able to balance accuracy and speed.

3. Materials and Methods

A novel detection algorithm for SAR ship targets with deep saliency representation, called SRDet, is proposed in this paper to balance improved accuracy with the speed of SAR

target detection against complex backgrounds. SRDet consists of the following modules: the anchor-free target detection benchmark framework CenterNet2 [35], the feature-enhancing lightweight backbone LWBackbone, the mixed-domain attention mechanism CNAM, the receptive-field-enhanced detection head RFEHead, and a module for data enhancement considering semantic relationships.

3.1. Network Architecture

The network architecture of the proposed SRDet algorithm is illustrated in Figure 1. The model primarily includes three important components: the feature extraction network LWBackbone is the backbone module, the bidirectional feature pyramid network (BiFPN) [36] feature fusion layers form the neck module and the final enhanced detection head RFEHead is the head module. A novel SAR target detection algorithm with deep saliency representation is proposed in this paper. This improved variant of the anchor-free target detection algorithm CenterNet2, which is called SRDet, can balance the accuracy and speed of SAR target detection against complex backgrounds. First, to compensate for the typically small sample size and small target size of the SAR targets, we designed a copy–paste method that considers semantic relationships for data enhancement. Second, we adopted the SOTA anchor-free target detection framework CenterNet2 as a benchmark and designed a new feature-enhancing lightweight backbone called LWBackbone, which requires fewer model parameters to effectively extract the salient features of SAR targets. Additionally, a new mixed-domain attention mechanism, called CNAM, is proposed to effectively suppress interference from complex land backgrounds and highlight the target area. Finally, we designed a receptive field enhanced detection head module called RFEHead, in which convolutions with different dilation rates are used to improve the receptive field and multiscale perception performance. The overall process is as follows:

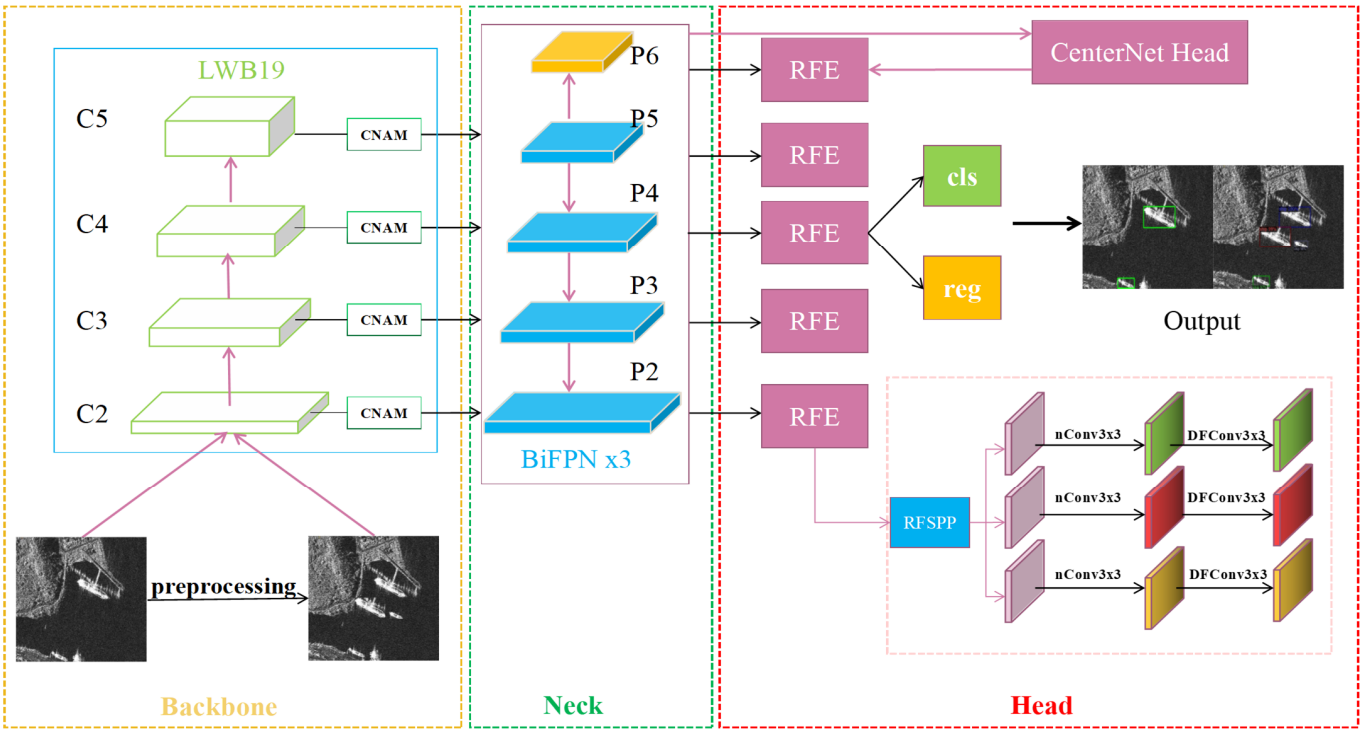

Figure 1. Overall architecture of SRDet. Preprocessing refers to performing data augmentation considering semantic relationships on the original image; LWB19 represents our designed lightweight backbone LWBackbone, which includes only 19 convolutional layers; CNAM represents our proposed attention module; and RFE represents our proposed augmented detection head module.

The input image is first passed through the backbone network LWBackbone for the extraction of target features and the extracted features are then sent to the BiFPN layers for feature fusion at different scales. The BiFPN architecture is used to introduce different weights in order to balance the feature information at different scales more effectively. After passing through three BiFPN layers in a row, the output features of the final BiFPN layer are passed to the detection head, before which a spatial pyramid pooling (SPP) module is added. To achieve a larger receptive field, the final features are obtained through CenterNetHead. Finally, a Fast R-CNN layer is used to calculate the final total loss and output the detection results.

3.2. Benchmark Target Detection Network

CenterNet2 is a target detection network developed as an improved two-stage variant of CenterNet by its authors [37]. The general idea of CenterNet is that to obtain the prediction results, the input image is divided into different areas, and each area is associated with a feature point network. The prediction results then indicate whether each feature point corresponds to an object and the type and confidence level of that object. Concurrently, the feature point is adjusted to obtain the center coordinates of the object and the width and height of the object are obtained through regression prediction. In this work, we adopt the two-stage concept for our detection algorithm but replace the RPN in the two-stage detection framework with a single-stage CenterNet and transfer the prediction results from the first stage to the second stage in a probabilistic way. In each stage of detection, the CenterNet2 model is used to extract regional features and perform classification, and Cascade R-CNN is used for classification and detection in the second stage. These models are trained together to maximize the accuracy of the predicted probabilities. The emergence of CenterNet2 has provided inspiration for the subsequent combination of excellent single-stage algorithms and two-stage algorithms. In the neck, the information from each layer in the CNN is utilized in the FPN to generate the final combination of expressive features. Due to the characteristics of SAR images, different feature layers have different resolutions. In a traditional FPN, the feature sharing in the fused output is not equal. Therefore, the feature fusion method of BiFPN is adopted instead in this paper to learn the different levels of importance of different features by means of learnable weights. The BiFPN is a weighted bidirectional feature pyramid network. Based on the PANet, the BiFPN deletes nodes with only one input edge to simplify the network.

The BiFPN module is used to integrate the features extracted by the backbone network so as to maintain all useful information. Low-level features contain more detailed spatial information and accurate location information, which is beneficial for small ship detection. Conversely, high-level features capture more semantic information but poorly reflect location information and are thus more suitable for detecting large ships.

3.3. Feature-Enhancing Lightweight Backbone: LWBackbone

In some real application scenarios, such as airborne SAR and spaceborne SAR, large and complex models are difficult to apply; thus, it is critical to study small and efficient networks for use in such scenarios. The DenseNet [38] network has a strong ability to extract features and requires fewer parameters and computations than ResNet [39]; thus, it is widely used. However, due to the dense connections in DenseNet, the detection speed is slow. Therefore, real-time detection requirements cannot be met when using DenseNet. Inspired by the recent VoVNetV2 network, our lightweight backbone LWBackbone is proposed to achieve real-time detection. LWBackbone consists of one-shot aggregation (OSA) modules. The first part of the backbone network is a stem block composed of a 3×3 deformable convolutional layer followed by a four-stage OSA module.

The OSA module consists of three 3×3 depthwise separable convolutions in series, the results of which are finally aggregated to one channel for output. We directly add the input to the output through residual connections, and we add an attention module (CNAM) to the final feature layer to further enhance the features. At the end of each stage, a 3 × 3 max

pooling layer with a stride of two is used for downsampling. The final output stride of the model is 32. The structure of the OSA module is shown in Figure 2. In summary, based on VoVNetV2, the residual connections of ResNet, the mixed-domain attention module CNAM and depthwise separable convolution, are introduced to form LWBackbone. The residual connections are added to enable the training of a deeper network and the attention mechanism is added to allow the model to better learn features. The network structure of LWBackbone is shown in Table 1. The abbreviation LWB19 indicates that the backbone network contains only 19 convolutional layers. In this table, the Type column lists each stage of the backbone network; the Output Stride column gives the output stride of each layer of the network; the Layers column describes each layer of the backbone network, where $\times 3$ denotes the presence of three depthwise separable convolutional layers in a row; and the Channels column gives the number of input and output channels of each layer. Due to a dataset of ship targets in SAR images is typically small and has multiscale characteristics, we select the lightweight LWBackbone (LWB19) as our benchmark backbone network. First, the three ordinary 3×3 convolutions in the first stage are replaced with 3×3 deformable convolutions (DFconv). The shape of deformable convolutions can be adjusted in accordance with the real situation to better extract the features of the input. Figure 3 shows the learning process for a deformable convolution. First, the bias is obtained through a convolutional layer, where the convolution kernel of this convolutional layer is the same as an ordinary convolution kernel. The output deviation size is the same as the input feature map size. The number of generated channel dimensions is 2N, corresponding to both the original output features and the offset features.

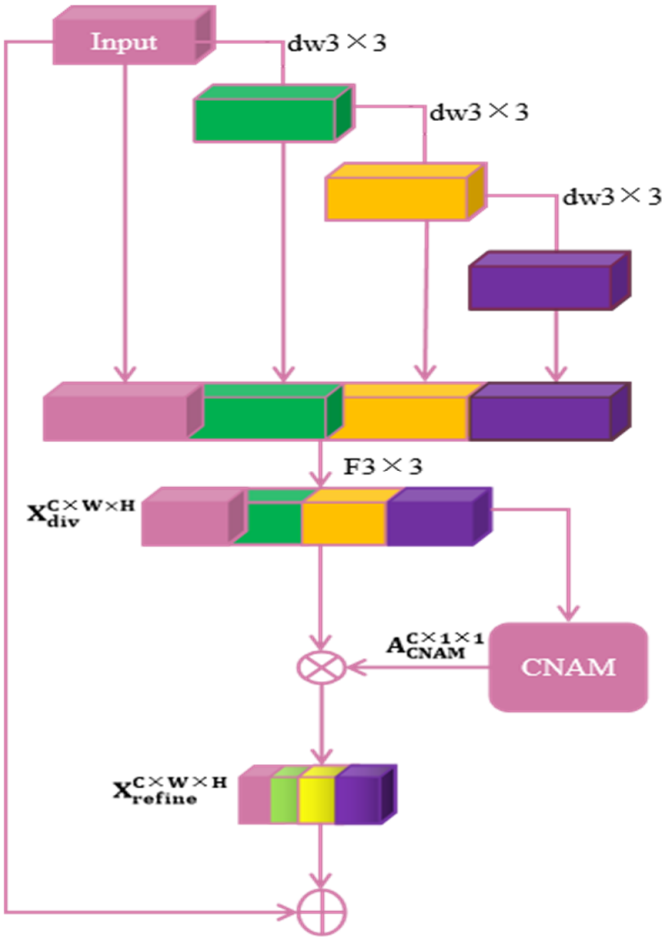

Figure 2. Structure of the OSA module.

Table 1. Network Structure of LWB19.

Type	Output Stride	Layer	Channels
Stage 1 Stem block	2 2 2	3 ×3 DFconv, stride = 2 3 ×3 DFconv, stride = 1 3 ×3 DFconv, stride = 1	64 64 64
Stage 2OSA module	4	(3 ×3 DWconv) ×3 Concat and 1 ×1 conv	64 112
Stage 3 OSA module	8	(3 ×3 DWconv) ×3 Concat and 1 ×1 conv	80 256
Stage 4 OSA module	16	(3 ×3 DWconv) ×3 Concat and 1 ×1 conv	96 384
Stage 5 OSA module	32	(3 ×3 DWconv) ×3 Concat and 1 ×1 conv	112 512

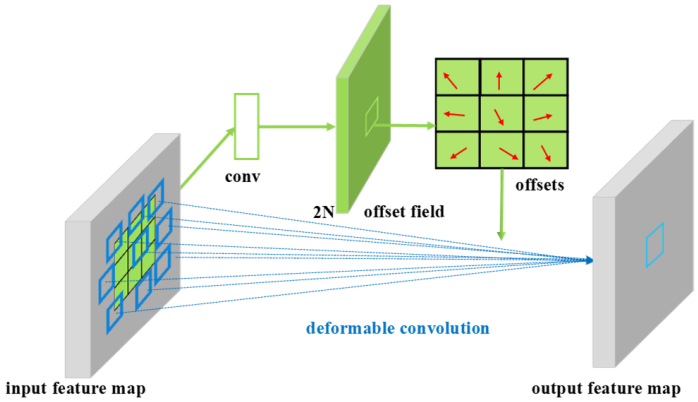

Figure 3. Deformable convolution. The number of generated channel dimensions is 2N, corresponding to both the original output features and the offset features.

In deformable convolution, an offset is applied to the convolution kernel at each sampling point of the input feature map to focus on a given ROI or target. Accordingly, depthwise separable convolution is used in the OSA module to marginally improve the detection accuracy of the model while reducing the number of model parameters. Moreover, we integrate the two attention mechanisms of a convolutional block attention module (CBAM) [40] and a normalization-based attention module (NAM) [41] to innovatively propose the CNAM attention mechanism, allowing the model to focus on ship target characteristics more effectively.

3.4. Mixed-Domain Attention Mechanism: CNAM

Due to the unique imaging principle of SAR imaging, densely distributed ships in a port will exhibit overlapping effects and the SAR land backgrounds are complex; consequently, background clutter can easily interfere with ship targets. In this paper, we propose a fused channel and spatial attention mechanism (CNAM) to pay more attention to ship features, thereby focusing the network's attention on the ship region. The SENet [42] attention mechanism is used in VoVNet. In SENet, only attention to different channels is considered with no regard for the spatial factor; consequently, this attention mechanism is not suitable for application to complex SAR images and its detection effect for small ships is not ideal.

A. Normalized channel attention

Previous attention mechanisms have focused only on salient features and ignoring non-salient features. Due to the different scales of the ship targets in SAR images, different channels can detect different ships; thus, we apply a sparse weight penalty factor to the

channel attention module to further suppress unimportant channels or pixels. The scale factor measures the variance of the channels and highlights their importance, as shown in Equation (1):

$$B_{out} = BN(B_{in}) = \frac{\alpha(B_{in} - \mu_B)}{\sqrt{\sigma_B^2 + \varepsilon}} + \beta \quad (1)$$

where μ_B is the mean, σ_B is the standard deviation and α and β are a trainable scale and shift, respectively. Normalized channel attention can be used to focus on effective channels and suppress ineffective channels. This process can improve the efficiency of information flow in the network. Figure 4 shows a schematic diagram of the channel attention mechanism, where F_1 denotes the input features; M_c denotes the output features; the parameters α are the scale factors of each channel, that is, the batch normalization (BN) layers, The weight values ω are obtained from Equation (2):

$$\omega_1 = \frac{\alpha_i}{\sum_{j=0} \alpha_j} \quad (2)$$

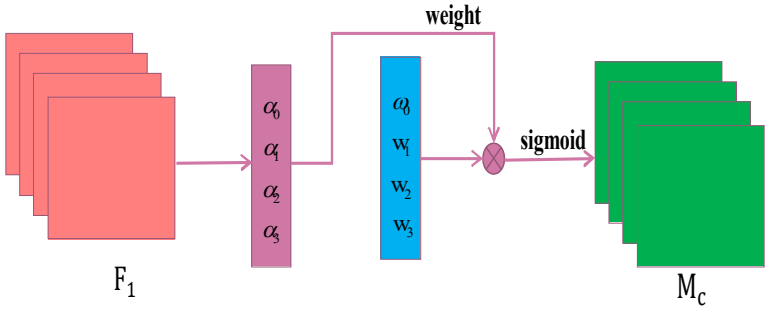

Figure 4. Normalized channel attention. α are the scale factors of each channel.

B. Spatial attention

In a SAR image, the pixel values of a ship target and the land background area may be very close, meaning that they are visually very similar; consequently, false detections or missed detections may easily occur. Therefore, we add a spatial attention mechanism to help the network learn which parts of the image to focus on. The feature map obtained from the channel attention module is used as input and global maximum pooling and global average pooling are then performed to obtain two feature maps with dimensions of H × W× 1. Subsequently, these two feature maps are spliced based on the channel dimension and a 7× 7 convolution is applied to reduce the number of channels to one. Finally, the sigmoid activation function is used to generate a spatial feature map, which is multiplied by the input features to obtain the final result. A flowchart of this process is shown in Figure 5.

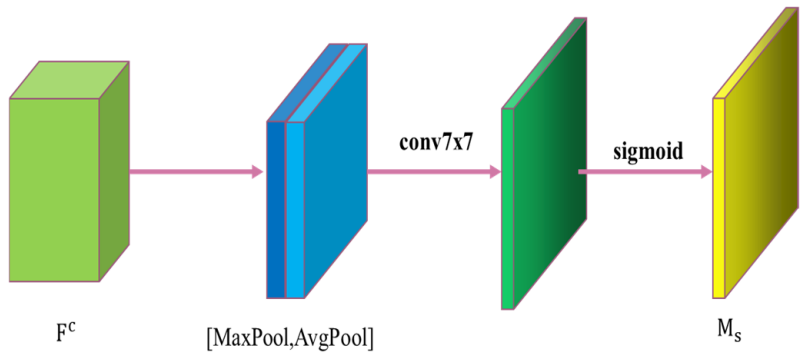

Figure 5. Spatial attention.

The input feature map F^c is obtained by compressing the feature map output by the channel attention module and the output feature map is denoted by M_s. The computation process is expressed as follows:

$$M_s = \sigma(f^{(7\times7)}([AvgPool(F^c); MaxPool(F^c)])) \tag{3}$$

C. CNAM

To more accurately capture ship feature information in SAR images, we fuse the normalized channel attention and spatial attention mechanisms. The input features are first passed through the normalized channel attention module; the input features are multiplied by the channel attention weights and the results are then sent to the spatial attention module; finally, the channel-weighted features are also multiplied by the spatial attention weights to obtain the adjusted features. A diagram of the overall structure of the CNAM mechanism is shown in Figure 6.

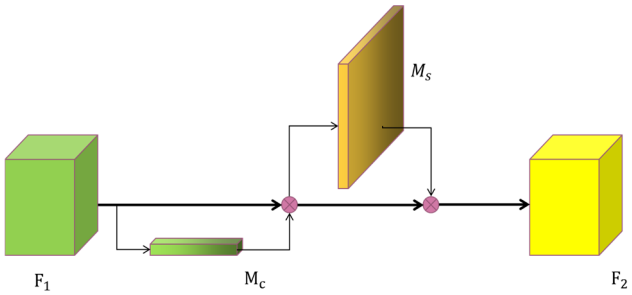

Figure 6. The mixed-domain attention mechanism CNAM.

3.5. Receptive-Field-Enhanced Detection Head: RFEHead

SAR ship targets generally have a large-scale range. To expand the receptive field, we add an SPP module with hollow convolution before the detection head to introduce multi-scale information. The receptive-field-enhancing SPP (RFSPP) module primarily consists of the following components: the input is passed through a 1×1 ordinary convolution and three convolutional layers with convolution kernels of different sizes and a dilated convolution layer is introduced. The input is also subjected to global average pooling to obtain image-level features, followed by a 1×1 convolution, then bilinear interpolation to the original size. Finally, the features from the five different scales are concatenated in the channel dimension and sent to a Conv1×1 layer for fusion before being output. A diagram of the structure of this module is shown in Figure 7.

For SAR ship targets, there is generally a marked imbalance between positive and negative samples; therefore, the proposed training loss function consists of two terms, with the CenterNet loss as the first-stage loss and the Cascade R-CNN loss as the second-stage loss:

$$\begin{aligned} \mathrm{L}_{loss} &= \mathrm{L}_{CenterNet} + \mathrm{L}_{Cascade\ R-CNN} \\ &= \mathrm{L}_{hm} + \mathrm{L}_{\mathrm{reg}} + \sum_{i=0}^{2} \left(L_{cls}^{i} + L_{reg}^{i} \right) \end{aligned} \tag{4}$$

For both the category loss L_{cls} and the category-independent confidence loss L_{hm}, we use an improved version of the focal loss function, which can well address the problem of imbalanced positive and negative samples. For the regression loss L_{reg}, we use the generalized intersection-over-union (*GIOU*) loss. As expressed in formula 5, we split all GT key points into a heatmap Y using a Gaussian kernel $Y_{xyc} = exp(-\frac{(x-\tilde{p}_x)^2+(y-\tilde{p}_y)^2}{2\sigma_{\tilde{p}}^2})$. When $Y_{xyc} = 1$, the point is a positive sample and the loss value of such an easily divided sample is very small. When Y_{xyc} takes any other value, the point is a negative sample and

the weight of the loss function (1-Y_{xyc}) is used to control the penalty. Additionally, α and β are both hyperparameters of the focal loss and are generally set to 2 and 4, respectively.

$$L_k = \frac{-1}{N}\sum_{xyc}\begin{cases}(1-\hat{Y}_{xyc})^{\alpha}\log(\hat{Y}_{xyc}) & Y_{xyc}=1\\ (1-Y_{xyc})^{\beta}(\hat{Y}_{xyc})^{\alpha}\log(1-\hat{Y}_{xyc}) & otherwise\end{cases} \tag{5}$$

We use the *GIOU* [43] loss function to calculate the regression loss, where the Intersection over Union (*IOU*) loss represents the difference in the intersection ratio between the predicted box and the real box. We denote the predicted box and the real box by *A* and *B*, respectively. C is the smallest box enclosing both *A* and B. We first calculate the ratio of the area of C that does not cover *A* or *B* to the total area of *C*, then subtract this ratio from the *IOU* of *A* and *B* to describe the detection effect of the predicted detection frame. Accordingly, the *GIOU* loss is defined as follows:

$$L_{giou} = 1 - GIOU = 1 - IOU + \frac{|C\backslash(A \cup B)|}{|C|} \tag{6}$$

where *IOU* represents the intersection ratio between the predicted box and the real box and $|C\backslash(A \cup B)|$ is the area in C that does not cover *A* or *B*.

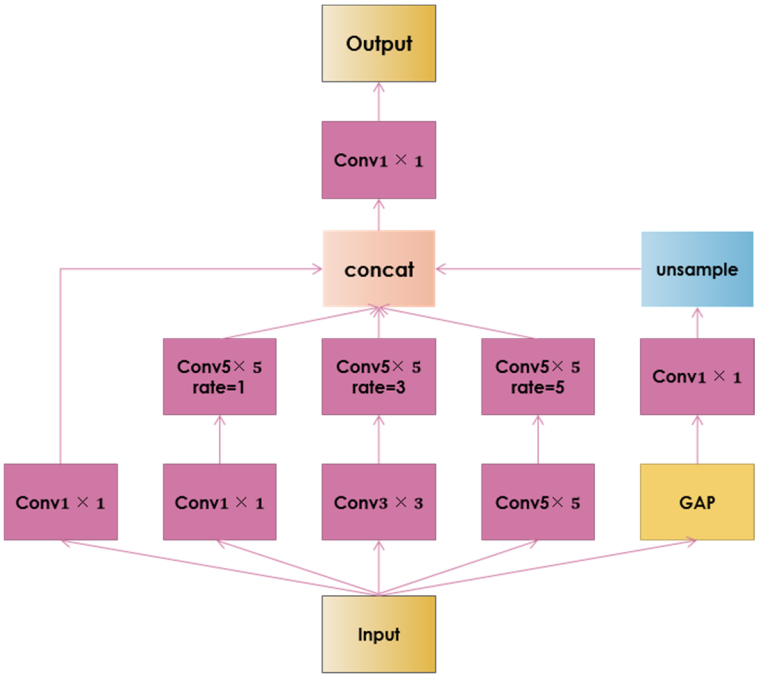

Figure 7. Structure of the RFSPP module. Rate represents the dilated rate and GAP represents the global average pooling.

3.6. Data Augmentation Considering Semantic Relationships

High-quality images (with rich object types and object scales) are the foundation for good processing results; thus, the image preprocessing operations remain important. Due to the remote sensing images exhibit complex spatial structures and capture diverse scenes, different images require different preprocessing operations, such as threshold segmentation [44], clustering [45] and data enhancement [46]. Some of the distinctive characteristics of SAR ship images and the detection difficulties they present are as follows: for application to real scenes, SAR ship target detection still faces some challenges, such as the influence of complex surroundings, multiscale targets and target defocusing, all of which affect performance in detecting ships.

Due to these problems, we believe it is beneficial to apply data enhancement methods. Commonly used data enhancement methods include flipping, rotation, scaling, mirror-

ing and jittering [47]. In this paper, rotation and horizontal flipping are used for data enhancement. In particular, the angular transformation of the images in the training set enhances the applicability of the trained model to images acquired at different angles, thus improving the generalizability of the model. Although these data enhancement methods increase the target sample size to a certain extent, they cannot increase the number of targets in an image and cannot solve the problems of multiscale targets in an image, the small proportion of small targets and the ease with which semantic information can be lost. Therefore, we introduce a data augmentation method that considers semantic relationships to solve this problem.

The cramming method is used to selectively copy a target object in an image in accordance with its label, perform a random transformation on it (e.g., a change in size by $\pm 20\%$ or rotation by $\pm 15^{\circ}$) and paste the copied target into a new position using the Poisson fusion method. By setting an appropriate threshold and reading the label file, we also ensure that the object pasted in this process does not overlap with any existing objects and is at least five pixels from the image boundary. To ensure that the enhanced dataset will contain strong semantic relationships, we also perform sea and land segmentation on the image before pasting to separate the land background from the sea background and only allow a ship target to be pasted onto the sea surface, preventing it from being pasted into a land region, thus the ship will be less likely to be confused with the land background.

To realize the semantic segmentation of sea and land, a classic segmentation threshold algorithm based on image binarization is adopted, namely, the Otsu algorithm. The Otsu method is simple to calculate and is not affected by image brightness or contrast; therefore, it is considered to be the best algorithm for threshold selection in image segmentation. The Otsu method can outline the area of the sea surface, which can help us determine whether the position of the pasted object meets the semantic requirements. Concurrently, the number of instances of pasting can be controlled. This method thus increases the number of targets and makes the positions of the target objects more diverse, enhancing their semantic information. As a result, the best data augmentation effect is achieved when using this method. Examples of the data augmentation results are shown in Figure 8.

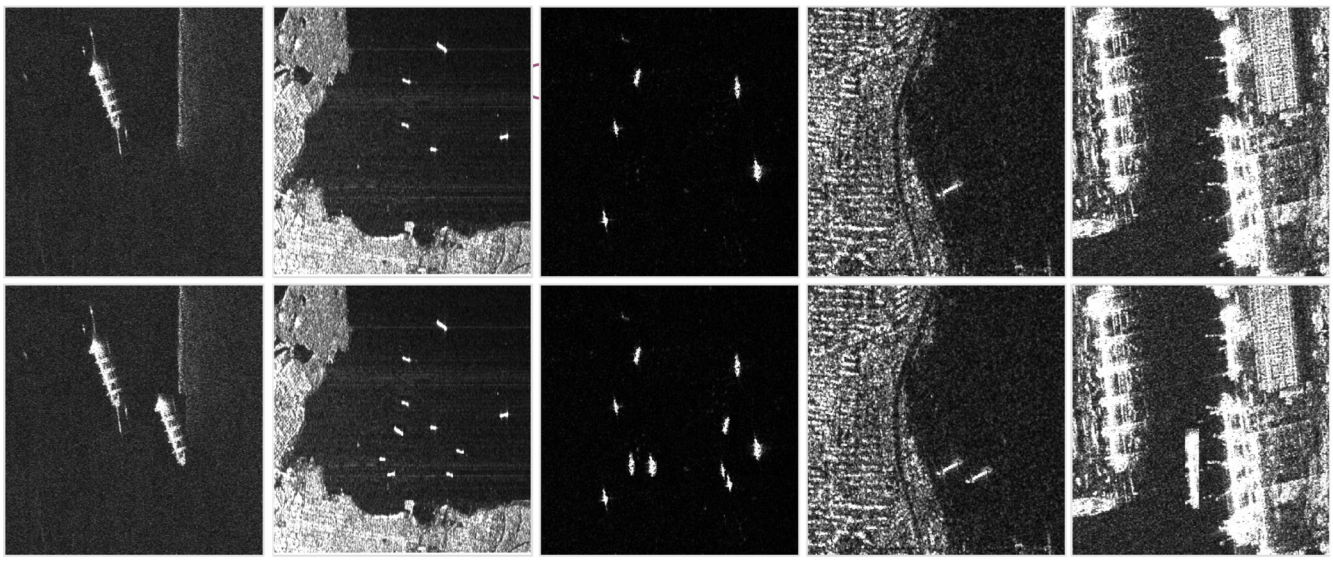

Figure 8. Enhanced renderings. The first row shows the original images and the second row shows the enhanced images.

4. Experiments

4.1. Dataset Introduction and Processing

To accurately evaluate the effectiveness of the proposed algorithm and preprocessing method, we conducted experiments using the commonly used ship dataset SSDD [48]. The SSDD dataset is the first widely used research dataset for ship detection based on deep

learning on SAR images. This dataset contains a total of 1160 images depicting a total of 2456 ships. The SSDD dataset contains multiscale SAR ships captured by different sensors in different polarization modes at different image resolutions from different scenes. For this study, the SSDD dataset was divided at a ratio of 8:2 by treating images with file numbers with a final digit of one or nine as the test set. Accordingly, there are 232 images in total in the test set and the remaining 928 images are regarded as the training set. High consistency of the network is conducive to the learning of network features and is also conducive to ensuring fairness in comparisons with other algorithms.

To more accurately evaluate the effectiveness of the algorithm and the preprocessing method, we applied the proposed copy-paste enhancement method considering semantic relationships to expand the SSDD dataset. The targets were copied from the SSDD dataset and randomly modified and pasted into the original image; we also ensured that the newly pasted targets did not overlap with the original targets in the image so that the target features would be more diverse. Then, we cleaned the newly obtained dataset to prevent the inclusion of individual images with poor results and named the new dataset ASSDD. The number of targets in the ASSDD dataset is increased from 2456 targets to 4449 targets, reflecting an increase in diversity. For ablation experiments, this dataset was used to verify the effectiveness of the preprocessing method.

To better verify our algorithm, we also conducted related experiments on two additional datasets, HRSID [49] and SAR-ship-dataset [50]. HRSID is a dataset for ship detection and segmentation in high-resolution SAR images that consists of 99 Sentinel-1B images, 36 TerraSAR-X images and 1 TanDEM-X image. These large scene images are cropped to 800 $\times$ 800, resulting in a total of 5604 high-resolution images that contain 16951 ship objects. For better comparisons with the official experimental results of other algorithms, we also scaled the SAR images to 1000 $\times$ 1000 pixels for experiments while leaving the other parameter settings essentially the same. SAR-ship-dataset is a high-resolution dataset constructed using 102 GF-3 images and 108 Sentinel-1 SAR images. The dataset consists of 43,819 images with an image size of 256 $\times$ 256 containing 59,535 ship targets. When using these two datasets, for better comparisons with other official algorithm results, we used the COCO evaluation index and divisions consistent with the official divisions of these datasets.

4.2. Experimental Setup

During model training, the momentum was set to 0.9, the optimizer was the stochastic gradient descent (SGD) optimizer, the decay rate was 0.0005, the batch size was two, the number of epochs was 300 and the learning rate was 0.001. The enhancement method used was EfficientDetResizeCrop, the training image size was 640, the number of BiFPN layers (NUM-BiFPN) was three and the number of output channels was 160. The model training environment used in this study was a system equipped with an Intel(R) Core(TM) i5-10600KF CPU @ 4.10 GHz with 32 GB of RAM, an NVIDIA GeForce RTX 2060 graphics card, Ubuntu 18.04, the Python programming language, PyTorch 1.7 as the deep learning framework and CUDA 10.1 and CUDNN 7.6.4 as the GPU acceleration libraries.

4.3. Evaluation Indices

In addition to the commonly used precision, recall, mean average precision (mAP) and F1 score, the evaluation metrics used in this study also included the number of parameters (parameter), the inference speed (FPS) and the maximum memory footprint (max-mem) to support a comprehensive analysis of model performance. First, we introduce the basic concepts: TP refers to the number of predicted positive examples that are actually positive, FP is the number of examples predicted to be positive that are actually negative, FN is the number of examples predicted to be negative that are actually positive and TN is the number of predicted negative examples that are actually negative.

Precision: Based on the prediction results, the proportion of correct predictions among the examples that are predicted to be positive is:

$$P = \frac{TP}{TP + FP} \tag{7}$$

Recall: Here, the positive examples are used as the judgment tool. Among the actually positive examples, the proportion of positive examples that are correctly predicted is as follows:

$$R = \frac{TP}{TP + FN} \tag{8}$$

For the case in which precision or recall alone is insufficient to evaluate the quality of a model, the $F1$ score combines the precision and recall metrics:

$$F1 = 2\frac{P \times R}{P + R} \tag{9}$$

The mAP is used to evaluate the detection performance of a model and represents the mean of the average precision (AP) values for each class; it is defined as follows:

$$mAP = \int_0^1 P(R)dR \tag{10}$$

Parameter: This metric is used to measure the model complexity. It includes the total number of weight parameters in all layers of the model and in the visual network components, primarily including convolutional layers, BN layers and fully connected layers.

FPS: The number of frames per second refers to the number of images for which a model can produce inference results per second, which is used to measure the real-time performance of the model.

4.4. Model Analysis

4.4.1. Ablation Experiments

To verify that each newly added module of the proposed algorithm functions as desired, we present a series of ablation experiments. The ablation experiments are primarily divided into four parts: (1) replacing the backbone network, (2) adding the new attention mechanism, (3) enhancing the detection head module and (4) using the preprocessed dataset.

(1) Replacing the backbone network. Due to the SSDD dataset contains only ship objects for detection, we replace the previous large backbone network with our new lightweight backbone network LWBackbone. Although the accuracy drops by approximately 0.005, the number of parameters drops by more than half and the inference time and maximum memory usage also decrease considerably, making the proposed model more lightweight and more suitable for subsequent embedded edge deployment. We also test the use of MobileNet in place of the backbone network and compare the performance under the same conditions. The performance of MobileNet is not superior to the performance of our backbone. Indeed, our proposed LWBackbone, which is specifically designed for object detection, is better than MobileNet in terms of both the mAP and parameter metrics. Our LWBackbone network has 10.5M fewer parameters than MobileNetv3 and 5.4M fewer than ShuffleNetv2 and achieves a higher mAP. Table 2 presents the quantitative comparison of the different backbones.

(2) Adding the new attention mechanism. The SENet attention mechanism is used in VoVNet. Although this mechanism does result in some enhancement, it has not been optimized. The purpose of this ablation experiment is to compare different attention modules and select the best. Table 3 shows the results of adding different attention mechanisms to the network and demonstrates why we select the new attention mechanism CNAM. The accuracy is improved by approximately 0.003, while the number of parameters remains basically unchanged, demonstrating that the proposed

attention mechanism is effective because it pays more attention to the distinctive features of SAR images from a mixed-domain perspective.

(3) Enhancing the detection head module by adding the RFSPP module before the detection head. The accuracy improves by approximately 0.002 when the proposed enhancement is added to the detection head. The receptive field of the proposed RFEHead is increased, allowing it to obtain multiscale spatial information of the targets and allowing the accuracy of the proposed algorithm to reach a SOTA level. Additionally, we test reducing the use of Cascade R-CNN in the detection head; here, the corresponding detection head configuration is denoted by CustomHead. It can be seen from the table that although the number of parameters of CustomHead is only 17.0M, the mAP of CustomHead is 0.016 lower than that of our proposed detection head, illustrating the effectiveness of our proposed module. Table 4 shows the quantitative comparison of the different detection heads.

Table 2. Comparison of different backbones.

ID	Backbone	mAP	Parameters	FPS	Max-Mem
1	Res2Net	0.9770	0.1 G	5	5149
2	ResNet50	0.9760	71.6 M	17	3660
3	MobileNetV2	0.9654	41.1 M	18	2535
4	MobileNetV3	0.9694	45.6 M	18	2900
5	ShuffleNetv2	0.9662	40.5 M	20	1927
6	LWBackbone	0.9721	35.1 M	25	1717

Table 3. Comparison of different attention mechanisms.

ID	Mechanism	mAP	Parameters	FPS	Max-Mem
1	SENet	0.9721	35.1 M	25	1717
2	CBAM	0.9742	35.2 M	20	1659
3	CNAM	0.9773	35.1 M	20	1700

Table 4. Comparison of different detection heads.

ID	Head	mAP	Parameters	FPS	Max-Mem
1	Head	0.9773	35.1 M	20	1659
2	CustomHead	0.9632	17.0 M	21	1457
3	RFEHead	0.9791	35.2 M	20	1764

To further demonstrate the superiority of our proposed algorithm, we visualize some of the results of the ablation experiments in Figure 9. The first column shows the ground truth for several images in the dataset and columns 1, 2 and 3 correspond to the first three groups of ablation experiments, in which different backbones, attention mechanisms and detection heads are used, respectively. When LWBackbone is used directly (in column 1), false and missed detections occur in the proposed model because of the replacement of the large backbone with a lightweight backbone, which makes the model's ability to extract features marginally weaker. As we add our other proposed improvements to the model, however, the model becomes more stable and its detection results become more accurate.

(4) Using the preprocessed dataset. We performed many ablation experiments on the new ASSDD dataset and the official version of the SSDD dataset and verified that the accuracy on the preprocessed dataset is markedly improved compared with that on the existing dataset. When ASSDD is used, the accuracy reaches a maximum of 98.55%. Table 5 compares the results for the different datasets.

Figure 10 shows a visualization of some of the results on the ASSDD dataset. The proposed algorithm can accurately detect the objects in these images after data augmentation. Thus, the generalization performance of the proposed model has been verified

through comparative experiments. Due to the increase in the number of targets at different positions, the model can extract more ship features from the augmented data, which is more conducive to model learning.

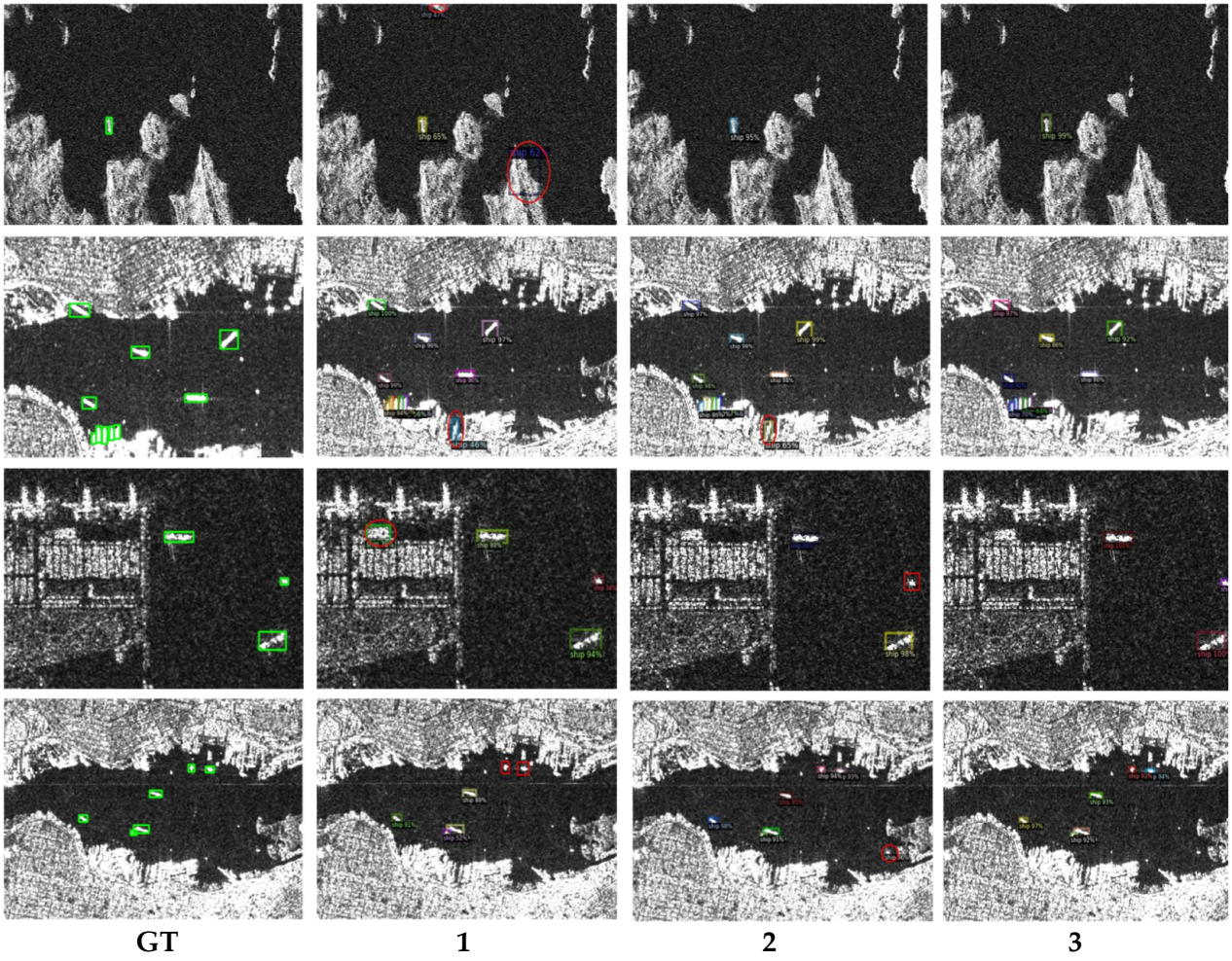

Figure 9. Ablation experiments. Note: A red circle indicates a target that is falsely detected by the algorithm and a red box indicates a target that is missed by the algorithm.

Table 5. Comparison of different datasets.

ID	Dataset	Method	mAP	Parameters (M)	FPS (img/s)	Max-mem
1	SSDD	CenterNet2	0.9760	71.6 M	17	3660
2	ASSDD	CenterNet2	0.9830	71.6 M	17	3660
3	SSDD	SRDet	0.9791	35.2 M	20	1764
4	ASSDD	SRDet	0.9855	35.2 M	20	1764

To better verify the effectiveness of the enhancement method proposed in this paper, it is also compared with other enhancement methods reported in the literature. From Table 6, we can see that the supposedly enhanced results of Mixup are actually worse, which may indicate that this data enhancement method is not suitable for the SAR target detection task. Cutout, Gridmask and Cutmix improved the mAP on the SSDD dataset by 0.26%, 0.29% and 0.44%, respectively, compared with the original dataset without enhancement. Compared with these mature data enhancement methods, the data enhancement method with semantic segmentation designed in this paper (SRDet) is more effective, improving the mAP by 0.64%.

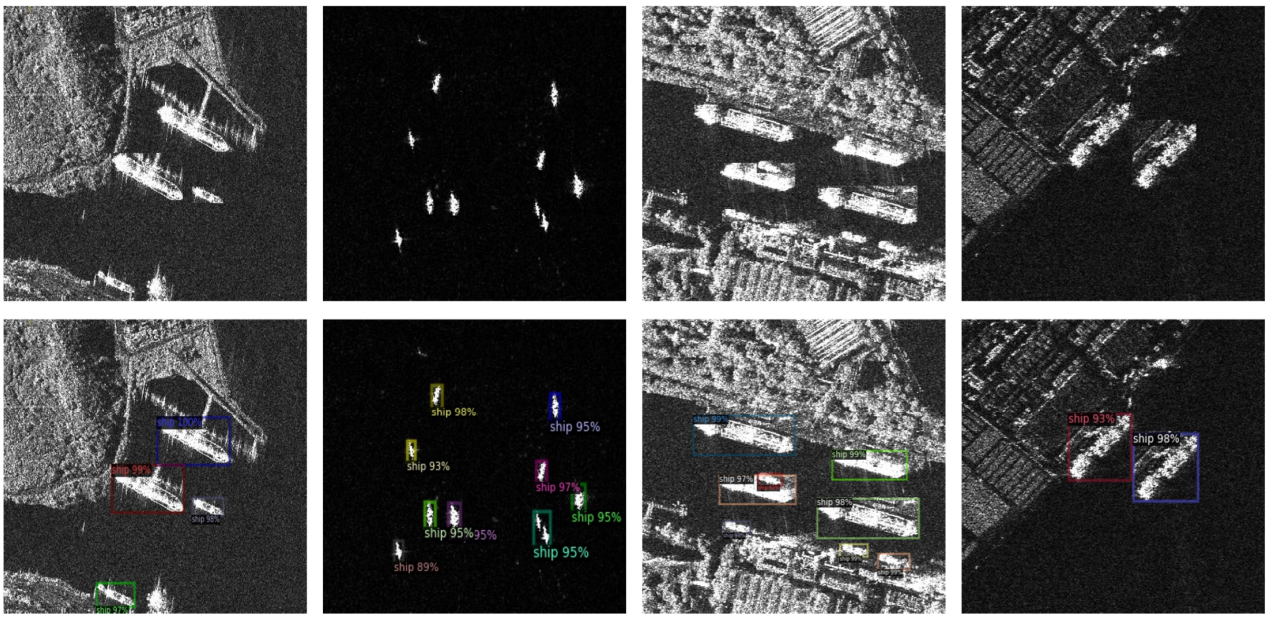

Figure 10. Visualization of results on the ASSDD dataset. The first row is the enhanced image and the second row is the corresponding detection result.

Table 6. Comparison of different enhancement strategies.

Method	Original	Mixup	Cutout	Gridmask	Cutmix	SRDet
mAP	0.9791	0.9690	0.9817	0.9820	0.9835	0.9855

Moreover, compared to some of the compared baseline methods, such as YOLOX and RetinaNet, our method in the worst case (i.e., our method under the condition of adding salt-and-pepper noise) still achieves a higher mAP. In addition, under the other two noise conditions, the mAP of our method is close to those of the other excellent baseline methods. Taken together, the above results fully verify that our method still shows strong robust performance under different noise conditions.

4.4.2. Comparison with Traditional CFAR Algorithms

To further verify the effectiveness of our proposed method, we compare it with the traditional CFAR detection algorithm and its improved variants. As shown in Table 7, this paper compares the performance of two traditional methods, CA-CFAR [16] and OS-CFAR [17], on the SSDD dataset. CA-CFAR is an algorithm that estimates the local environment and the time-dependent noise level within a reference window and then judges whether a pixel belongs to a target on the basis of a set threshold. OS-CFAR can achieve good results in multi-objective situations but requires high computing power. The experimental results show that our method is far superior to these traditional methods in terms of precision and $F1$ score and performs basically the same as the traditional methods in terms of recall rate. At the same time, the inconvenience of manually designing features and thresholds is eliminated and the generalization performance of the model is also better. In the future, a promising topic of research will be to investigate how to better combine traditional methods with deep learning methods for ship detection.

Table 7. Comparison of CFAR detection algorithms.

Method	*P*	*R*	*F1*
CA-CFAR [16]	0.859	0.981	0.916
OS_CFAR [17]	0.842	0.985	0.902
SRDet (ours)	0.951	0.983	0.967

4.4.3. Comparison with Two-Stage Detection Algorithms

To further validate the proposed method, we first compare it with several existing two-stage models (Faster R-CNN, Cascade R-CNN, FPN and the baseline CenterNet2). Under the same conditions, the proposed model has the fewest parameters and the highest accuracy among the two-stage detection algorithms. Table 8 presents the quantitative results of this series of comparative experiments.

Table 8. Comparison of the proposed model with existing two-stage detection algorithms.

ID	Method	*P*	*R*	*F1*	*mAP*	*Parameters (M)*	*FPS (img/s)*
1	Faster R-CNN [22]	0.923	0.982	0.952	0.964	41.1	3
2	Cascade R-CNN [23]	0.926	0.980	0.952	0.967	85.6	2
3	FPN [51]	0.930	0.975	0.952	0.965	63.56	14
4	CenterNet2 [31]	0.943	0.972	0.957	0.976	71.6	17
5	SRDet	0.951	0.983	0.967	0.979	35.1	20

As shown by the visualizations in Figure 11, Faster R-CNN has a large model volume and results in many missed and false detections. Marginal improvements are achieved using Cascade R-CNN; however, there are still many erroneous results. Thus, we conclude that the detection performance of SRDet for small and multiscale targets in complex scenes is markedly improved. For ships, more accurate detection boxes can be obtained with higher accuracy. This advantage can be primarily attributed to the proposed algorithm, which fuses the advantages of two-stage and one-stage algorithms while enhancing the features of SAR images from multiple perspectives.

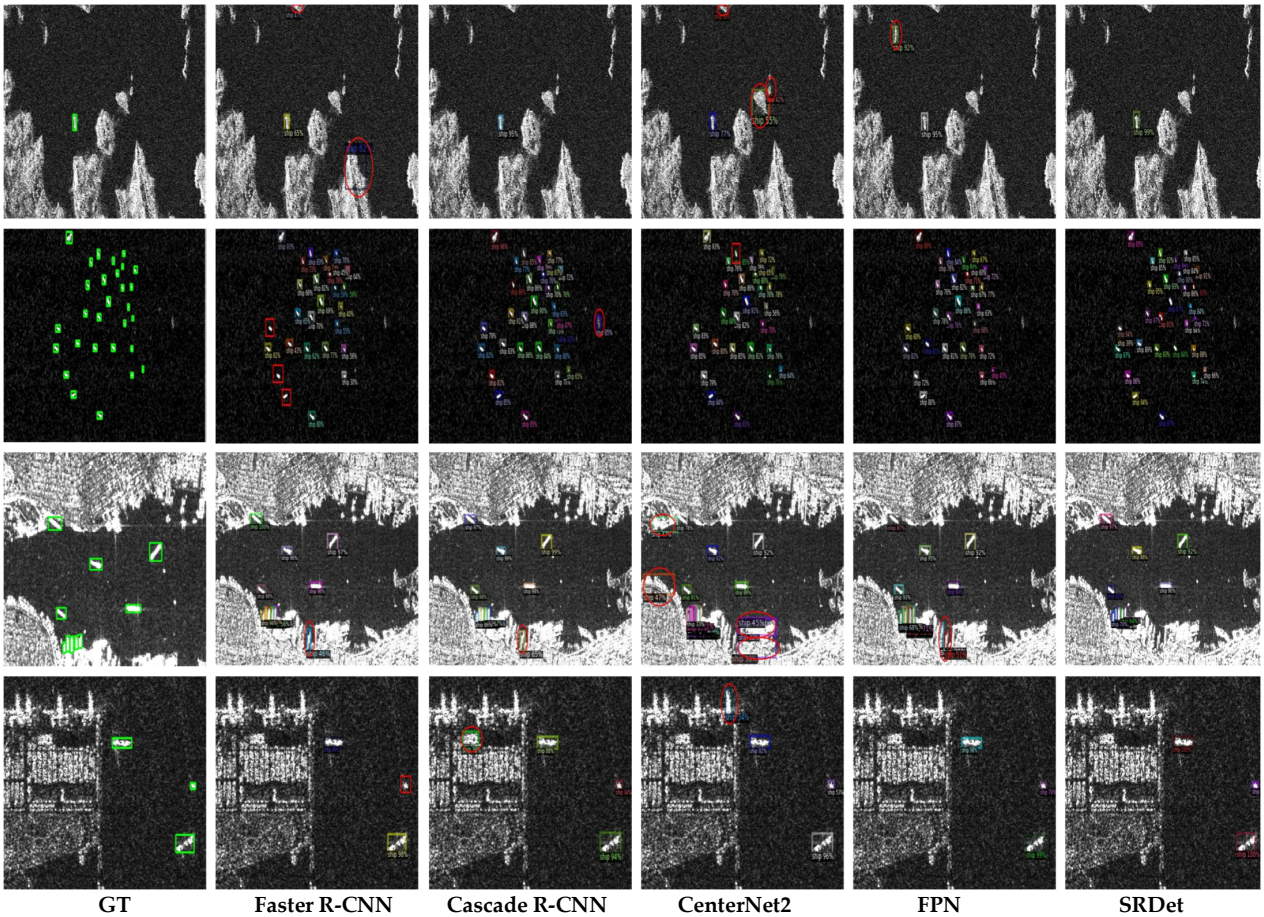

Figure 11. Visual comparison with two-stage object detection algorithms. Comparison results of five different methods on SSDD, GT stands for ground truth.

4.4.4. Comparison with One-Stage Detection Algorithms

The number of parameters and inference speed of the proposed model are not better than those of all one-stage models, but the accuracy of the proposed model is the highest among the models tested in this paper. As seen from the visualizations shown in Figure 12, several single-stage detection algorithms tend to miss some small ships and generate some offshore false positives. This may be because a one-stage model can use only feature maps with smaller resolutions, resulting in smaller targets from which it may not be possible to obtain many features, whereas the proposed model can effectively solve this problem. The computational speed of the proposed model is slower than that of the one-stage detection models, primarily because of the larger volume of the proposed model, which consists of two stages; however, it can solve problems, such as serious misdetection of ships against complex backgrounds and achieve better accuracy and recall, thus yielding the best map. In the proposed algorithm, the anchor-free approach is adopted and combined with the two-stage concept. The proposed algorithm is beneficial for locating the positions of targets and providing accurate predictions. Table 9 shows the comparison results of one-stage algorithms.

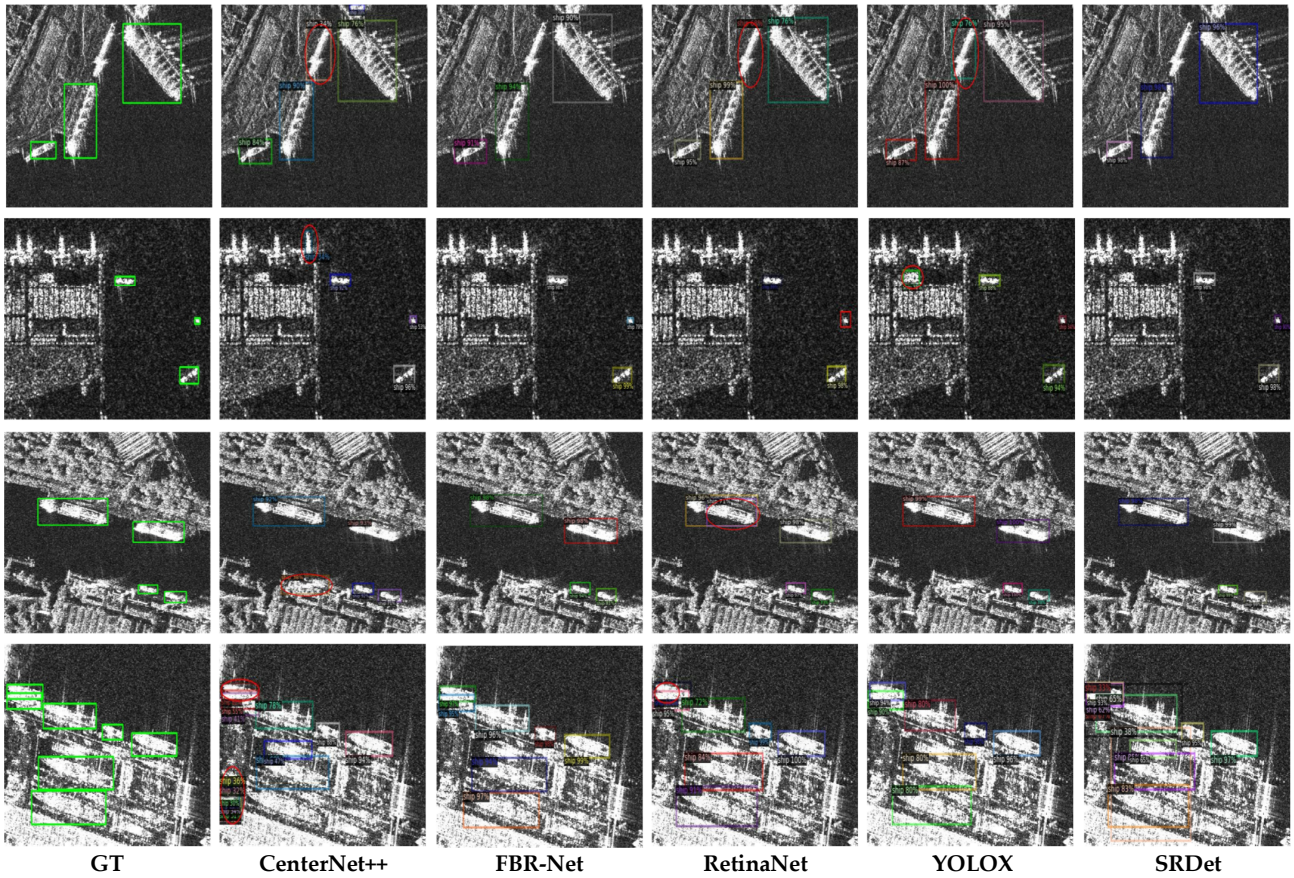

Figure 12. Visual comparison with one-stage object detection algorithms.

Table 9. Comparison of the proposed model with existing one-stage detection algorithms.

ID	Method	*P*	*R*	*F1*	*mAP*	*Parameters (M)*	*FPS (img/s)*
1	CenterNet++ [52]	0.833	0.952	0.889	0.951	57.83	25
2	FBR-Net [23]	0.914	0.940	0.934	0.941	32.5	25
3	RetinaNet [22]	0.877	0.948	0.911	0.901	37.74	25
4	YOLOX [53]	0.932	0.979	0.955	0.972	8.94	30
5	SRDet	0.951	0.983	0.967	0.979	35.1	20

4.4.5. Comparison with SOTA SAR Ship Detection Methods

To further verify the effectiveness of our proposed algorithm, we carried out a further performance comparison with existing advanced SAR target detection algorithms. Based on the comparison of the chosen evaluation indicators, the proposed algorithm is shown to achieve a high detection accuracy. However, because the codes of the existing SOTA SAR ship detection methods are not open source and we cannot reproduce some of the details of these methods, we can only cite the best results reported in the corresponding studies based on the chosen indicators. As seen from Table 10, the mAP of SRDet is 0.028 higher than that of CenterNet++ and 0.002 higher than that of AFSar. The proposed algorithm also has the highest recall of 0.983 and the highest $F1$ score of 0.967. Therefore, the above results in terms of multiple indicators show that our SRDet algorithm performs best.

Table 10. Comparison with SAR ship detection methods on SSDD.

Method	*P*	*R*	*F1*	*mAP*
CenterNet++ [52]	0.833	0.952	0.889	0.951
FBR-Net [29]	0.928	0.940	0.934	0.941
TWC-Net [54]	0.914	0.953	0.933	-
CRTransSar [55]	0.925	0.983	0.953	0.970
NMDNet [34]	0.946	0.932	0.939	0.962
AFSar [56]	0.941	0.982	0.961	0.977
SRDet	0.951	0.983	0.967	0.979

4.4.6. Experimental Results on Other Datasets

To verify the effectiveness and generalization ability of the proposed algorithm, we also conducted related experiments on the HRSID and SAR-ship-dataset datasets. Our algorithm is more accurate and faster than the baseline CenterNet2, indicating that our improved model works well. Additionally, we conducted a comparison with the algorithm launched when the official dataset was released. Under essentially the same parameters, our algorithm has certain advantages. As seen from Table 11, in terms of AP50, the result of our proposed algorithm on the HRSID dataset is 1.3% higher than the official HRSDNet result and 1.1% higher than that of the benchmark CenterNet2. On the SAR-ship-dataset, the accuracy of our proposed algorithm reaches 95.1%, which is much higher than that of some classical methods and 1.5% higher than that of the benchmark CenterNet2.

Table 11. Comparison with SAR ship detection methods on HRSID and SAR-ship-dataset.

Dataset	Model	*AP*	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
HRSID	RetinaNet [24]	59.8	84.8	67.2	60.4	62.7	26.5
HRSID	YOLOX [53]	61.4	87.2	68.9	63.0	57.0	21.8
HRSID	HRSDNet [49]	69.4	89.3	79.8	70.3	71.1	28.9
HRSID	CenterNet2 [35]	64.5	89.5	73.0	64.7	69.1	48.3
HRSID	SRDet	66.1	90.6	75.1	66.1	72.1	56.9
SAR-ship-dataset	RetinaNet [24]	58.9	92.3	67.7	52.3	66.4	69.9
SAR-ship-dataset	SLCANet [57]	54.2	88.3	53.4	48.5	62.1	42.3
SAR-ship-dataset	CenterNet2 [35]	60.1	93.6	69.8	53.5	67.6	72.4
SAR-ship-dataset	SRDet	65.9	95.1	78.8	59.3	72.8	78.7

Our team additionally cooperated with the 38th Research Institute of China Electronics Technology Group to accumulate some large-scale image data. We screened out two large slices of nonconfidential image data for detection, with an image size of 4000 × 4000. From Figure 13, we can see that most ship targets can be detected accurately; however, there are also false detections. This may be because the weights trained on the other datasets used in this study cannot perfectly generalize to detection in large images, meaning that false detections and missed detections will occur without de novo training.

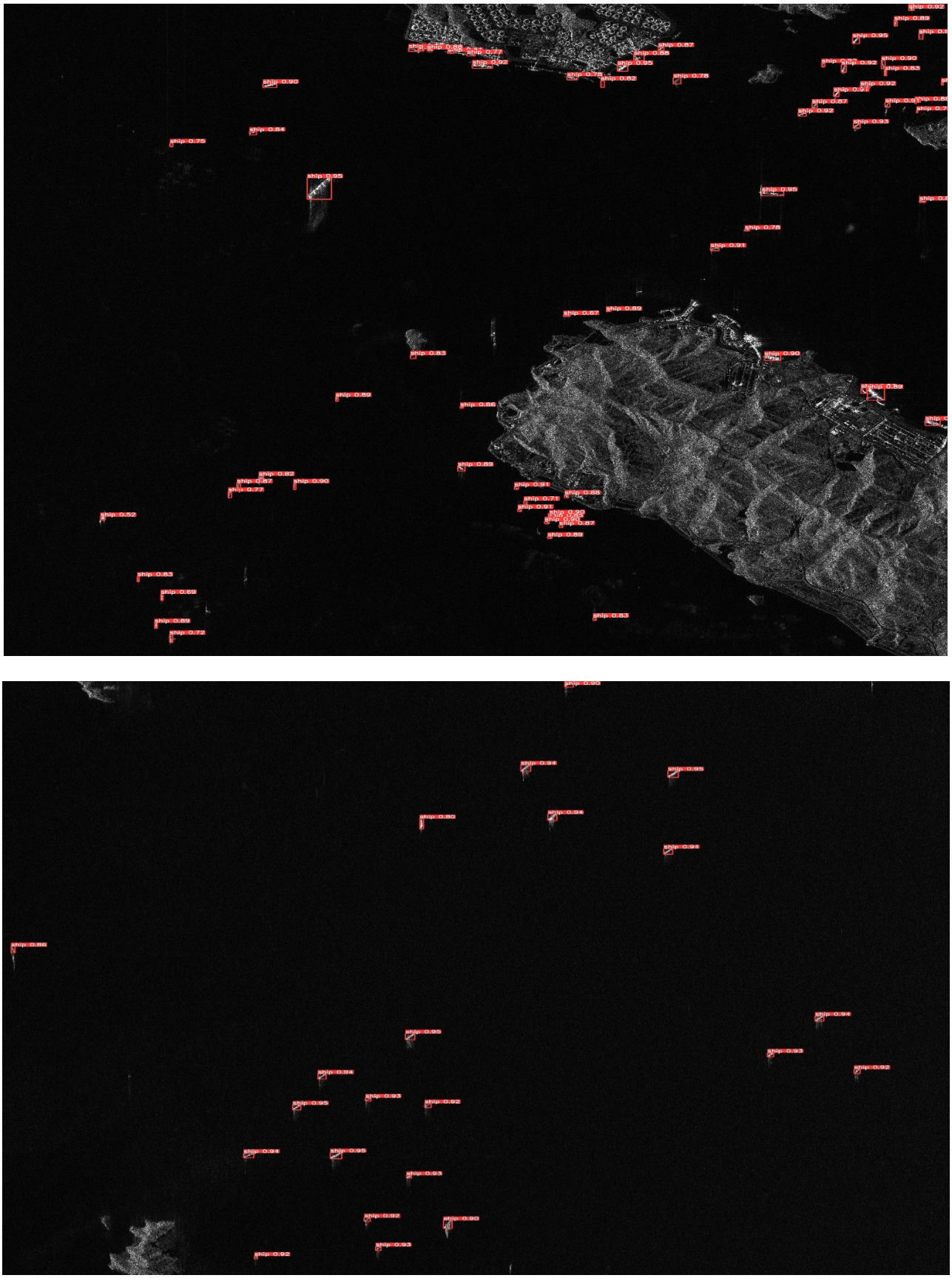

Figure 13. Detection results in large-scale SAR images. Both images are 4000 pixels × 4000 pixels.

In the future, we will directly use such large images to retrain the weights and update the detection model in order to obtain optimal detection results.

5. Conclusions

To mitigate issues related to unclear contour information, complex backgrounds and the sparse and multiscale nature of SAR image targets, we have proposed an anchor-free algorithm with deep saliency representation for the detection of SAR ship targets, called SRDet. First, due to the difficulty of SAR target acquisition and the typical small sample sizes and small targets, we first applied a copy-paste data augmentation method that considers semantic relationships to preprocess the data in order to reduce possible model overfitting during the training process. Second, the feature extraction backbone network was reconstructed; the CenterNet2 backbone network was replaced with a lightweight backbone network LWBackbone, reducing the number of model parameters and enabling the effective extraction of multiscale salient features of SAR targets. Additionally, a new mixed-domain attention mechanism called CNAM was proposed to effectively suppress interference from complex land backgrounds and highlight the target area. Finally, we designed a receptive-field-enhancement detection head module called RFEHead, in which convolutions at different dilation rates are used to enhance the receptive field and improve the multiscale perception performance. The proposed algorithm was verified to achieve superior performance in comparison with existing detection algorithms. The experimental results on the SSDD dataset showed that the mAP of the proposed method reached 97.9%. After the data were preprocessed, the proposed mAP reached 98.6%, the FPS reached 20 frames per second and the overall performance reached the SOTA level. Concurrently, we also validated our method on other SAR ship detection datasets and the experimental results showed that our method yields good results. In future research, the following topics should be explored to further improve the performance of target detection in SAR images:

(1) Domain knowledge relevant to SAR images can be further incorporated into SRDet. There is a large difference between the imaging mechanisms of SAR images and optical images. SAR target samples are more difficult to obtain and exhibit strong scattering. When the imaging angle and background change, the performance of a detection network will also decrease to a certain extent, and the generalizability tends to be poor. Considering the unique imaging mechanism and background scattering characteristics of SAR images, we plan to develop a network that is more suitable for target detection in SAR images.
(2) Due to the typically high density of ships in a port, the foreground frames can often be confused and not effectively distinguished. Therefore, to extract the features of ship targets, we plan to focus on ship detection in a rotated frame and on pixel segmentation to allow the model to obtain more accurate target features.
(3) In the experiments conducted in this study, we found that most SAR targets are small and unclear. Therefore, we plan to consider integrating a super resolution reconstruction network into the proposed model to make the contours of the targets clearer, which would be beneficial for feature extraction.

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

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62001003, in part by the Natural Science Foundation of Anhui Province under Grant 2008085QF284, and in part by the China Postdoctoral Science Foundation under Grant 2020M671851.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code and the ASSDD dataset are available at https://github.com/AHUCICG/SAR-detection/tree/master (accessed on 22 December 2022).

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

References

1. Li, Z.; Wu, J.; Huang, Y.; Sun, Z.; Yang, J. Ground-moving target imaging and velocity estimation based on mismatched compression for bistatic forward-looking SAR. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3277–3291. [CrossRef]
2. Zhou, Z.; Chen, J.; Huang, Z.; Wan, H.; Chang, P.; Li, Z.; Yao, B.; Wu, B.; Sun, L.; Xing, M. FSODS: A Lightweight Metalearning Method for Few-Shot Object Detection on SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5232217. [CrossRef]
3. Shao, Z.; Wu, W.; Wang, Z.; Du, W.; Li, C. SeaShips: A large-scale precisely annotated dataset for ship detection. *IEEE Trans. Multimed.* **2018**, *20*, 2593–2604. [CrossRef]
4. Gao, F.; He, Y.; Wang, J.; Hussain, A.; Zhou, H. Anchor-free convolutional network with dense attention feature aggregation for ship detection in SAR images. *Remote Sens.* **2020**, *12*, 2619. [CrossRef]
5. Yuan, S.; Yu, Z.; Li, C.; Wang, S. A novel SAR sidelobe suppression method based on CNN. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 132–136. [CrossRef]
6. Han, J.; Li, G.; Zhang, X. Refocusing of moving targets based on low-bit quantized SAR data via parametric quantized iterative hard thresh-olding. *IEEE Trans. Aerosp. Electron. Syst.* **2020**, *56*, 2198–2211. [CrossRef]
7. Chen, G.; Li, G.; Liu, Y.; Zhang, X.; Zhang, L. SAR image despeckling based on combination of fractional-order total variation and nonlocal low rank regularization. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2056–2070. [CrossRef]
8. Lee, Y.; Hwang, J.-w.; Lee, S.; Bae, Y.; Park, J. An Energy and GPU-Computation Efficient Backbone Network for Real-Time Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 752–760. [CrossRef]
9. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
10. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9308–9316.
11. Crisp, D.J. *The State-of-the-Art in Ship Detection in Synthetic Aperture Radar Imagery*; Defence Science and Technology Organisation Salisbury (Australia) Info Sciences Lab: Port Wakefield, SA, Australia, 2004; Research Report DSTO-RR-0272.
12. Paes, R.L.; Nunziata, F.; Migliaccio, M. On the capability of hybrid-polarity features to observe metallic targets at sea. *IEEE J. Oceanogr. Eng.* **2016**, *40*, 426–440. [CrossRef]
13. Sugimoto, M.; Ouchi, K.; Nakamura, Y. On the novel use of model-based decomposition in SAR polarimetry for target detection on the sea. *Remote Sens. Lett.* **2013**, *4*, 843–852. [CrossRef]
14. Chen, S.; Li, X. A new CFAR algorithm based on variable window for ship target detection in SAR images. *Signal Image Video Process.* **2019**, *13*, 779–786. [CrossRef]
15. Ai, J.; Qi, X.; Yu, W.; Deng, Y.; Liu, F.; Shi, L. A new CFAR ship detection algorithm based on 2-D joint log-normal distribution in SAR images. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 806–810. [CrossRef]
16. Kuang, C.; Wang, C.; Wen, B.; Hou, Y.; Lai, Y. An improved CA-CFAR method for ship target detection in strong clutter using UHF radar. *IEEE Signal Process. Lett.* **2020**, *27*, 1445–1449. [CrossRef]
17. Hyun, E.; Lee, J.-H. A new OS-CFAR detector design. In Proceedings of the 2011 First ACIS/JNU International Conference on Computers, Networks, Systems and Industrial Engineering, Jeju, Republic of Korea, 23–25 May 2011; pp. 133–136.
18. Ao, W.; Xu, F.; Li, Y.; Wang, H. Detection and discrimination of ship targets in complex background from spaceborne ALOS-2 SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 536–550. [CrossRef]
19. Dellinger, F.; Delon, J.; Gousseau, Y.; Michel, J.; Tupin, F. SAR-SIFT: A SIFT-like algorithm for SAR images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 453–466. [CrossRef]
20. Wang, S.; Wang, M.; Yang, S.; Jiao, L. New hierarchical saliency filtering for fast ship detection in high-resolution SAR images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 351–362. [CrossRef]
21. Lang, H.; Zhang, J.; Zhang, X.; Meng, J. Ship classification in SAR image by joint feature and classifier selection. *IEEE Geosci. Remote Sens. Lett.* **2015**, *13*, 212–216. [CrossRef]
22. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef]
23. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162. [CrossRef]
24. Lin, T.-Y.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
25. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European conference on computer vision (ECCV), Online, 8–14 September 2018; pp. 734–750.
26. 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 October–2 November 2019; pp. 9626–9635.

27. Zhou, X.; Zhuo, J.; Krahenbuhl, P. Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 850–859.
28. Kang, M.; Leng, X.; Lin, Z.; Ji, K. A modifified Faster R-CNN based on CFAR algorithm for SAR ship detection. In Proceedings of the International Workshop on Remote Sensing with Intelligent Processing (RSIP), Shanghai, China, 19–21 May 2017; pp. 1–4.
29. Fu, J.; Sun, X.; Wang, Z.; Fu, K. An anchor-free method based on feature balancing and refifinement network for multiscale ship detection in SAR images. *IEEE Trans. Geosci. Remote. Sens.* **2020**, *59*, 1331–1344. [CrossRef]
30. Wang, X.; Cui, Z.; Cao, Z.; Dang, S. Dense Docked Ship Detection via Spatial Group-Wise Enhance Attention in SAR Images. In Proceedings of the IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020.
31. Sun, X.; Lv, Y.; Wang, Z.; Fu, K. SCAN: Scattering Characteristics Analysis Network for Few-Shot Aircraft Classification in High-Resolution SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5226517. [CrossRef]
32. Sun, Z.; Dai, M.; Leng, X.; Lei, Y.; Xiong, B.; Ji, K.; Kuang, G. An Anchor-Free Detection Method for Ship Targets in High-Resolution SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7799–7816. [CrossRef]
33. Yang, X.; Zhang, X.; Wang, N.; Gao, X. A Robust One-Stage Detector for Multiscale Ship Detection With Complex Background in Massive SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5217712. [CrossRef]
34. Li, D.; Liang, Q.; Liu, H.; Liu, Q.; Liu, H.; Liao, G. A Novel Multidimensional Domain Deep Learning Network for SAR Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5203213. [CrossRef]
35. Zhou, X.; Vladlen, K.; Philipp, K. Probabilistic two-stage detection. *arXiv* **2021**, arXiv:2103.07461.
36. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
37. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6568–6577. [CrossRef]
38. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
40. 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), Online, 8–14 September 2018.
41. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-based Attention Module. *arXiv* **2021**, arXiv:2111.12419.
42. 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]
43. 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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666. [CrossRef]
44. Audebert, N.; Le Saux, B.; Lefèvre, S. How useful is region-based classification of remote sensing images in a deep learning framework. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 5091–5094.
45. Zhang, X.; Wang, G.; Zhu, P.; Zhang, T.; Li, C.; Jiao, L. GRS-Det: An Anchor-Free Rotation Ship Detector Based on Gaussian-Mask in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3518–3531. [CrossRef]
46. Harris, E.; Marcu, A.; Painter, M.; Niranjan, M.; Prügel-Bennett, A.; Hare, J. Fmix: Enhancing mixed sample data augmentation. *arXiv* **2020**, arXiv:2002.12047.
47. Ruiz, D.V.; Krinski, B.A.; Todt, E. IDA: Improved Data Augmentation Applied to Salient Object Detection. In Proceedings of the 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Porto de Galinhas, Brazil, 7–10 November 2020; pp. 210–217. [CrossRef]
48. Li, J.; Qu, C.; Shao, J. Ship detection in SAR images based on an improved faster R-CNN. In Proceedings of the In 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), Beijing, China, 13–14 November 2017; pp. 1–6.
49. Wei, S.; Zeng, X.; Qu, Q.; 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]
50. 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]
51. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
52. Guo, H.; Yang, X.; Wang, N.; Gao, X. A CenterNet++ model for ship detection in SAR images. *Pattern Recognit.* **2021**, *112*, 107787. [CrossRef]
53. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv* **2021**, arXiv:2107.08430.

54. Yu, L.; Wu, H.; Zhong, Z.; Zheng, L.; Deng, Q.; Hu, H. TWC-Net: A SAR ship detection using two-way convolution and multiscale feature mapping. *Remote Sens.* **2021**, *13*, 2558. [CrossRef]
55. 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]
56. Wan, H.; Chen, J.; Huang, Z.; Xia, R.; Wu, B.; Sun, L.; Yao, B.; Liu, X.; Xing, M. AFSar: An Anchor-Free SAR Target Detection Algorithm Based on Multiscale Enhancement Representation Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5219514. [CrossRef]
57. Hou, B.; Wu, Z.; Ren, B.; Li, Z.; Guo, X.; Wang, S.; Jiao, L. A Neural Network Based on Consistency Learning and Adversarial Learning for Semisupervised Synthetic Aperture Radar Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5220816. [CrossRef]