remote sensing

MDPI

Article

YOLO-SD: Small Ship Detection in SAR Images by Multi-Scale Convolution and Feature Transformer Module

Simin Wang [1,2], Song Gao [1,2,*], Lun Zhou [1,2], Ruochen Liu [1,2], Hengsheng Zhang [1,2], Jiaming Liu [1,2], Yong Jia [1,2] and Jiang Qian [3]

[1] Key Laboratory of Earth Exploration and Infomation Techniques (Chengdu University of Technology), Ministry of Education, Chengdu 610059, China
[2] The College of Mechanical and Electrical Engineering, Chengdu University of Technology, Chengdu 610059, China
[3] The School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, China
* Correspondence: gs@cdut.edu.cn

Abstract: As an outstanding method for ocean monitoring, synthetic aperture radar (SAR) has received much attention from scholars in recent years. With the rapid advances in the field of SAR technology and image processing, significant progress has also been made in ship detection in SAR images. When dealing with large-scale ships on a wide sea surface, most existing algorithms can achieve great detection results. However, small ships in SAR images contain little feature information. It is difficult to differentiate them from the background clutter, and there is the problem of a low detection rate and high false alarms. To improve the detection accuracy for small ships, we propose an efficient ship detection model based on YOLOX, named YOLO-Ship Detection (YOLO-SD). First, Multi-Scale Convolution (MSC) is proposed to fuse feature information at different scales so as to resolve the problem of unbalanced semantic information in the lower layer and improve the ability of feature extraction. Further, the Feature Transformer Module (FTM) is designed to capture global features and link them to the context for the purpose of optimizing high-layer semantic information and ultimately achieving excellent detection performance. A large number of experiments on the HRSID and LS-SSDD-v1.0 datasets show that YOLO-SD achieves a better detection performance than the baseline YOLOX. Compared with other excellent object detection models, YOLO-SD still has an edge in terms of overall performance.

Keywords: synthetic aperture radar (SAR); small ship detection; deep learning; YOLOX

Citation: 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. https://doi.org/10.3390/rs14205268

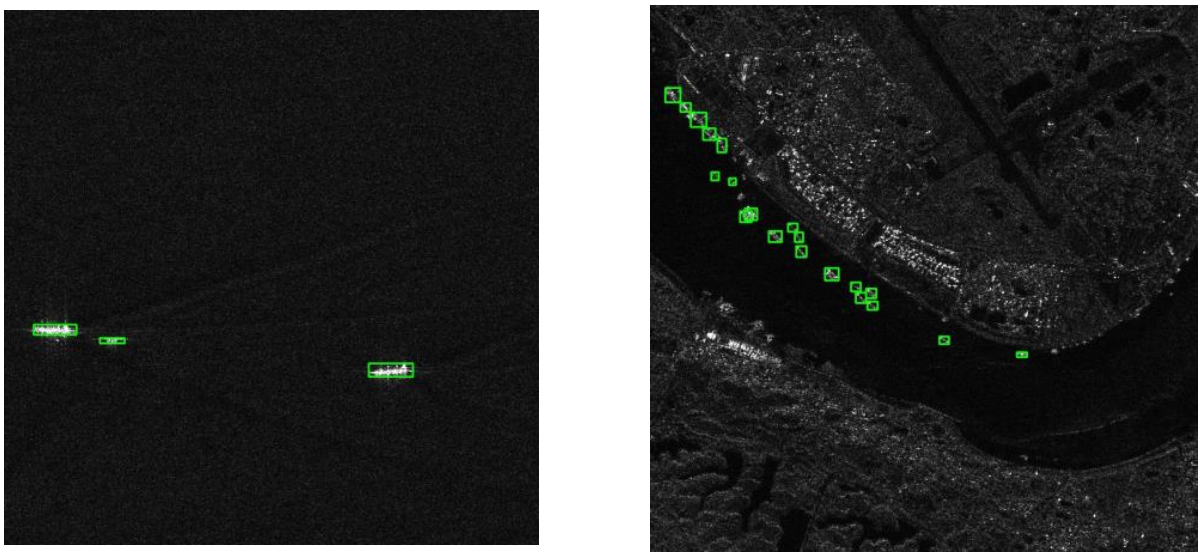
Academic Editor: Dusan Gleich

Received: 21 September 2022
Accepted: 18 October 2022
Published: 21 October 2022

1. Introduction

The imaging effect of traditional optical sensing is always affected by many factors, such as clouds and illumination, while Synthetic Aperture Radar (SAR) is not. SAR has the characteristics of strong penetration and a durable working period, so it is more suitable for ever-changing marine scenes [1,2]. With the increasingly serious challenge of maritime rights, SAR technology has become one of the important tools for marine monitoring.

When processing SAR images, the detection of ships is mainly achieved by comparing changes in pixel grey values and extracting feature information, such as the structure and shape of ship objects [3]. Poor visual effects lead to unsatisfactory detection in cluttered scenes, such as nearshore and harbors. Figure 1 shows two SAR image examples in the HRSID, a simple off-shore image (a) and a complicated in-shore image (b). In Figure 1a, the ships are large, sparsely distributed, and have a clear trail. The ships in Figure 1b have the characteristics of a small area and dense distribution, which makes ship detection difficult and challenging.



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

Figure 1. Two SAR image examples in different scenarios, with green boxes referring to ship objects labeled in the HRSID. (**a**) A simple off-shore image without interference; (**b**) a complicated image with inland interference and dense ships.

For ship detection in SAR images, traditional methods are mainly based on the feature differences between sea clutter and ship objects. The traditional methods can be divided into two types depending on the object of concern, based on auxiliary features and on statistical features. Analyzing the trail and leaking oil traces behind the ships, the algorithm based on auxiliary features achieves the indirect detection of ship targets and eliminates false detection in the results [4]. However, tail detection can only be performed if there are trails in the image, which makes the employment of this algorithm limited. The algorithm based on statistical features compares the marine background and ship objects in SAR images, analyzes the contrast information, and ultimately achieves ship detection. The Constant False Alarm Rate (CFAR) detection algorithm is the first and one of the most widely used [5]. By statistical inference and probabilistic modeling, the CFAR algorithm estimates the probability density function of ocean clutter and sets an appropriate threshold to separate the objects from clutter. However, the accuracy of pixel-based CFAR detection is poor as it is easily influenced by serious clutter and other factors [6]. According to different image characteristics, CFAR needs to select appropriate sea clutter distribution models. When the image resolution is low, a Gaussian distribution or negative exponential distribution is used to describe the sea clutter [7]. Dealing with high-resolution SAR images, Qin et al. [8] used the log-cumulants method to gain the parameters and proposed a CFAR detection algorithm based on a generalized gamma distribution, which showed better results. Linking the co-polarized channels to the burst time offset between the channels, Nunziata et al. [9] proposed an innovative dual-polarization model and a CFAR method to process the full-resolution CSK PingPong SAR data to observe ships and oil platforms. Based on the scattering characteristics of ships at sea, Ferrara et al. [10] proposed a physical model that processes full-resolution Single-Look Complex (SLC) SAR information combined with an efficient filtering technique designed to achieve the high-quality identification of targets and backgrounds at sea. The above traditional algorithms are suitable for processing single as well as simple SAR images, but are not quite effective when dealing with complex maritime situations.

Since the introduction of deep learning, it has continued to evolve and received widespread attention from scholars from all walks of life. By applying this technique to image processing, the detection accuracy and speed of tasks such as target detection and instance segmentation have been significantly improved [11]. Depending on their structure,

there are two types of deep learning-based detection algorithms: one-stage algorithms and two-stage algorithms [12]. The principle of the two-stage algorithm is to generate candidate frames first and then classify them on the basis of whether they contain objects or not. As a pioneer in object detection using deep learning, the R-CNN [13] algorithm has substantially improved detection accuracy compared with traditional detection algorithms. Based on R-CNN, scholars have made improvements, resulting in excellent algorithms such as Faster R-CNN [14], Mask R-CNN [15], Dynamic R-CNN [16], Sparse R-CNN [17], and Libra R-CNN [18]. Meanwhile, the one-stage model samples the image uniformly at all locations and transforms the detection mission into a regression classification task; examples are YOLO series [19,20], RetinaNet [21], YOLOF [22], etc. Generally, the one-stage algorithm adopts the end-to-end training mode, which is usually faster but has low accuracy. The two-stage algorithm can achieve great accuracy, but its computational overhead remains large.

To make further breakthroughs in the field of machine vision, Dosovitskiy et al. [23] redesigned the transformer, which was proposed by Vaswani et al. [24,25], to encode images as sequences and proposed the first visual transformer (ViT) for image classification. The best recognition results can be achieved by applying ViT in optically natural scenes. Carson et al. [26] introduced the transformer to object recognition and proposed the Detection Transformer (DETR). To reduce the computational consumption and the false drop rate involved in SAR complex backgrounds, Li et al. [27] added a transformer encoder after the backbone ResNet101 [28] and fused semantic information and location information. Srinivas et al. [29] combined the transformer with the backbone network and achieved the best results in areas of image generation and instance segmentation. These studies have all employed hybrid structures, combining CNN-based models with transformers to achieve excellent results on computer vision tasks. Inspired by this, this paper proposes the design of a kind of transformer structure for YOLOX [30] to achieve the high-efficiency detection of small ships in SAR images.

The accurate detection of small ships has always been a challenging research topic due to the characteristics of SAR images [31]. When the area of the ship is small (typically less than 48^2 pixels [32,33]), it is only shown as a bright spot. Lacking feature information during detection, these are easily confused with other interference, resulting in missed detection and affecting the final results. In order to improve the detection accuracy of small ships, scholars have performed a lot of research in this field.

For the poor detection of small objects caused by class imbalances, data augmentation is employed to expand the data of small ships, improve the model's attention to them, and thus enhance the contribution of small objects to the loss function calculation during training. For the partial or even full feature loss of small objects due to down-sampling, model optimization strategies such as feature fusion are often used to increase feature information. Specifically, on the basis of the traditional SSD detector [34], Juan et al. [35] introduced the data augmentation of rotation and expansion, added dilated convolution in the backbone, and finally improved the adaptability of the model to small objects. Chen et al. [36] inserted feature pyramids into the Region Proposal Network to compensate for the loss of small ships' location information at the bottom of the network. Yang et al. [37] designed a perceptual field enhancement module to integrate different convolutions and pooling, which enhanced the transfer of feature information and ultimately reduced the false alarms of small objects. To achieve small ship detection in PoISAR images, Jin et al. [38] replaced all normal convolutions in the network with extended convolutions when expanding the perceptual field. In the FBR-Net network proposed by Fu et al. [39], the designed ABP structure utilizes a layer-based attention method and spatial attention method to balance the semantic information of the features in each layer, which made the network more focused on small-scale ships. To improve the detection of small ships in SAR images with complex backgrounds, Guo et al. [40] combined feature refinement, feature fusion, and head enhancement approaches to design a highly accurate detector, called CenterNet++. Chang et al. [41] proposed a GPU-based deep learning detection method, called YOLOv2,

which offers superior detection speed and accuracy and greatly improves the efficiency of ship detection in SAR images. To further enhance the detection effect of small-scale ships, Su et al. [42] proposed a Spatial Information Integration Network (SII-Net). In SII-Net, a Channel-Location Attention Mechanism (CLAM) block and a multi-scale pooling layer were applied to obtain richer ship position information, and interpolations and poolings were employed after the PANet to enhance the model's attention to targets. While the network does achieve a high overall detection accuracy, it is less effective for densely distributed ship groups. Considering that contextual information is crucial for the detection of small and dense ships, Zhao et al. [43] proposed a novel CNN-based method. In this method, as many small ship proposals as possible are first made and then combined with contextual information to exclude spurious ships from the predictions.

Based on the above analysis, this paper uses the latest YOLOX algorithm as the baseline, focuses on enhancing the feature extraction capability, captures richer contextual information, and ultimately strengthens the detection ability of ships in high-resolution SAR image datasets, especially small ships. Our contributions can be summarized as follows:

1. To improve the sensitivity of minor object detection, we designed an MSC block. It combines several parallel convolutions with the residual network, which can obtain feature information of different sizes and perform multi-scale fusion, further enhancing the representation of semantic information;
2. In addition, we proposed the FTM block. It divides the high-layer feature information into two parts, processes them using the transformer encoder, and finally merges them via a cross-stage structure. With this FTM module, our model can capture global features effectively and achieve higher detection accuracy;
3. Taking YOLOX as the baseline, we incorporated MSC and FTM, and proposed an efficient detection model YOLO-SD for ship detection in high-resolution SAR images;
4. YOLO-SD was tested on the HRSID [32] and LS-SSDD-v1.0 datasets [33]. According to the experimental results, our detection accuracy was improved dramatically compared with YOLOX, which indicates the effectiveness of our model. Besides this, we compared our design with some existing excellent networks when applied to the same dataset, and the results showed that ours still excels in overall performance.

2. Material and Methods

Firstly, our method is derived by analyzing the shortcomings of the existing method, YOLOX. Next, each key point is described in detail, including the specific architecture and working principles, and the overall structure of YOLO-SD is shown. Last, we introduce the environment's setup, the datasets used, and the evaluation metrics.

2.1. Proposed Method Based on YOLOX

In the detection network, the bottom feature map has high resolution and rich detailed features, which makes it suitable for small ship detection. In the higher feature map, the image resolution seems low, but its semantic information becomes rich, which is appropriate for detecting large-scale ships [44]. Figure 2a shows the partial structure of YOLOX. YOLOX applies CSPDarkNet53 as the backbone and mainly uses C3, C4, and C5 for feature fusion and classification, but its actual detection of small ships is poor. Firstly, the C3 layer is located in the shallower region of the network, with rich detailed information and high resolution. YOLOX mainly utilizes it for feature extraction to achieve small ship detection. However, due to the small reception field of the C3 layer, the semantic information [45] obtained is weak. Thus, YOLOX will consider some real ships (especially small ships) as background, resulting in poor final detection. Next, the C5 layer has low resolution and is highly abstract, so lots of small ships have lost some or all of their detailed features, at which point it is no longer meaningful to process the C5 layer for small ship detection.

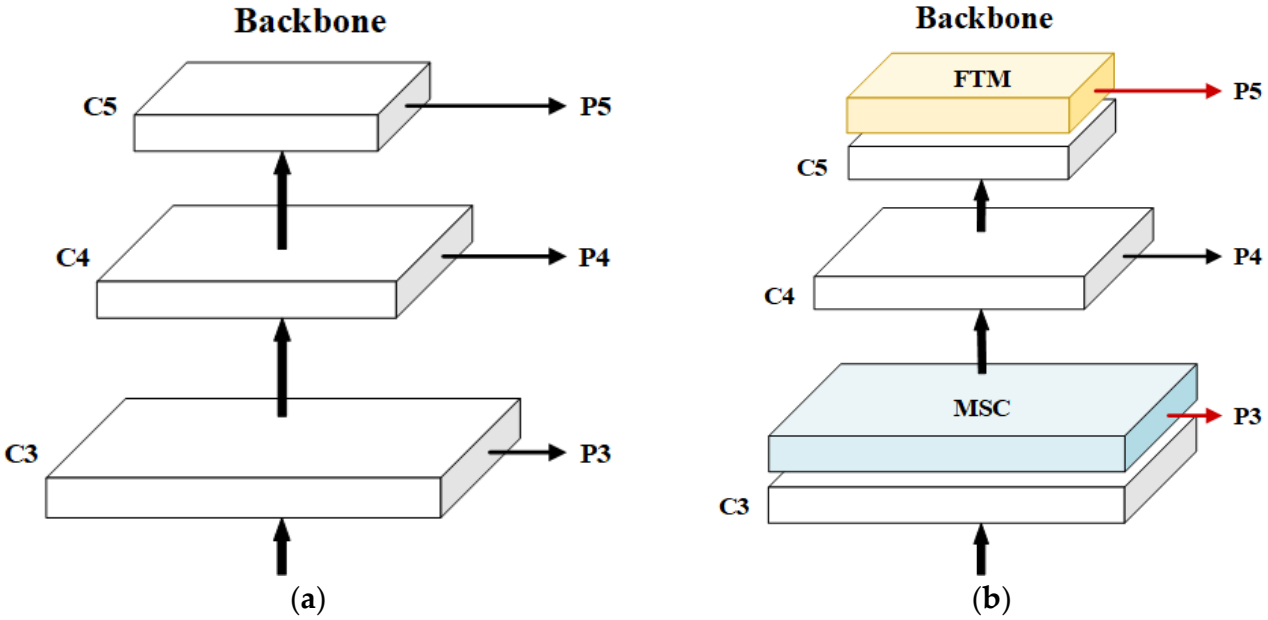


Figure 2. Comparison of partial model structure: (**a**) YOLOX. (**b**) YOLO-SD.

To solve these problems of YOLOX and improve the accuracy of small ship detection in SAR images, we have used YOLOX_L as our baseline to design a new detection model for small ships, called YOLO-SD. The partial structure of YOLO-SD is shown in Figure 2b. The MSC and FTM in the backbone CSPDarkNet53 are our designs, while the red lines indicate the new connection between the backbone and the neck. Firstly, several parallel multi-scale convolutions were inserted after the C3 layer. These convolutions help the network to obtain feature information from different reception fields and enhance semantic information. Secondly, after layer C5, in which most of the detailed features were lost, we added a newly designed FTM block, mainly consisting of a multi-headed attention layer and a fully connected layer, to optimize the feature information. Lastly, to improve the effect of feature fusion, we modified the connection between the backbone and the neck. In Figure 2b, P3, P4, and P5 are connected with MSC, C4, and FTM. In this method, we apply the newly improved feature maps to enhance the network's attention to small ships, which in turn improves the network's detection performance. The experiments demonstrate that YOLO-SD increases the computational overhead to a lesser extent than baseline YOLOX, but improves the accuracy significantly.

2.2. Specific Architecture of YOLO-SD

In order to optimize model performance, we considered deepening the backbone network as well as expanding its width. However, stacking structures directly not only increase the computational cost significantly, but also make the network prone to scattering [46]. Therefore, we applied the basic structure of CSPDarkNet53 [47] and improved on it by introducing the following design.

We propose the MSC module to improve the effect, as shown in the following Formulas (1) and (2):

$$y_{MSC} = Relu(x + Concat(c_0(x), c_1(x), c_2(x))) \tag{1}$$

$$\begin{cases} c_0(x) = w_{01} *_3 w_{00} *_1 x \\ c_1(x) = w_{11} *_5 w_{10} *_1 x \\ c_2(x) = w_2 *_1 x \end{cases} \tag{2}$$

where x is the input feature map and y is the output feature map. In Formula (2), c_i (i = 0, 1, 2) represents the i-th convolution branch, and $*_i$ (i = 1, 3, 5) represents the convolution with kernel size $i \times i$. W_{ij} (j = 0, 1) means the weight parameters of convolution, and the lower corner indicates the j-th convolution of the i-th branch.

The diagram of the MSC feature enhancement structure is shown in Figure 3. MSC is mainly a parallel filter structure, which connects the outputs of convolutions with different kernel sizes into a single output. These parallel convolutions are performed at different scales and can extract features from different receptive fields at the same time, which has two benefits. On the one hand, it improves the feature extraction effect for ships. On the other hand, deeper features can enhance the semantic information of the feature map and improve the model's ability to detect small ships. These parallel convolution operations occupy a lot of computer resources, so we add 1×1 convolution before processing to alleviate the problem. The 1×1 convolution can both further increase the network depth and reduce the dimension (changing the number of channels to 0.5, 0.25, and 0.25 times the number of input channels), as well as reduce the computational consumption. However, network widening and deepening induces training difficulties and gradient disappearance problems while improving performance. For this reason, we introduce the ResNet structure, which directly connects the input of MSC with the output of the concatenation operation. Through this design, we process and aggregate the information while limiting the amount of computation, deepen the network while enhancing the expressive ability, and improve the sensitivity of the model to small ships.

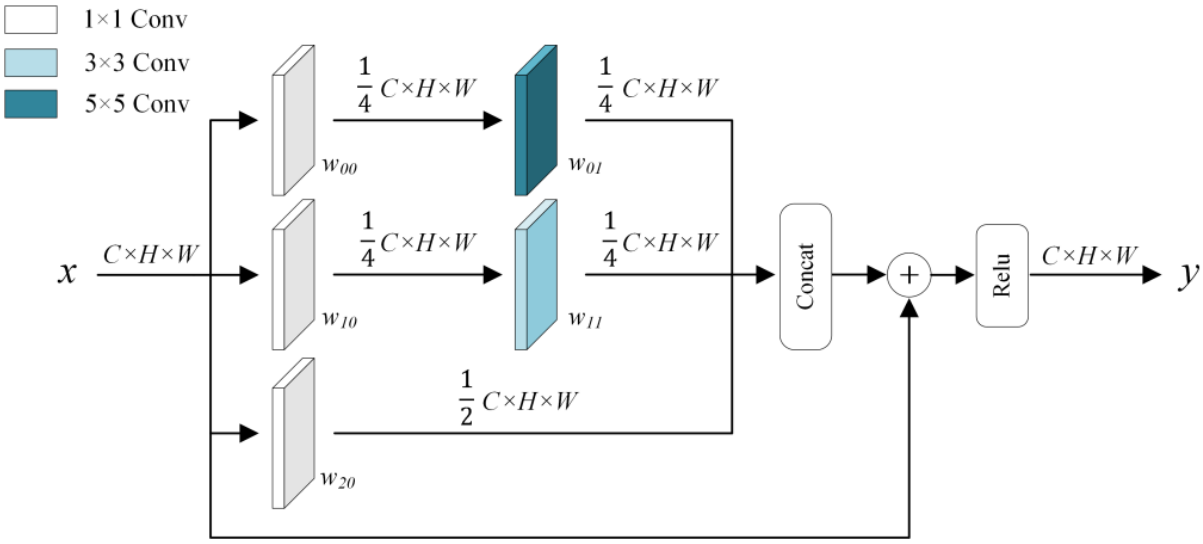


Figure 3. The structure of MSC, which adopts convolution operation with different kernel sizes and fuses after residual connection to extract and refine features in receptive fields of various sizes.

During detection, the model mainly relies on the backbone network to extract local feature information from SAR images. However, the large down-sampling factors involved in extraction may mean the model misses small-scale ships. In addition, the model is unable to capture sufficient global information due to the small actual receptive field of the convolutional neural network. To improve the capability of small ship detection in SAR images and minimize the leakage of small targets, we propose a Feature Transformer Module (FTM) that can capture rich global and contextual information, as shown in Figure 4.

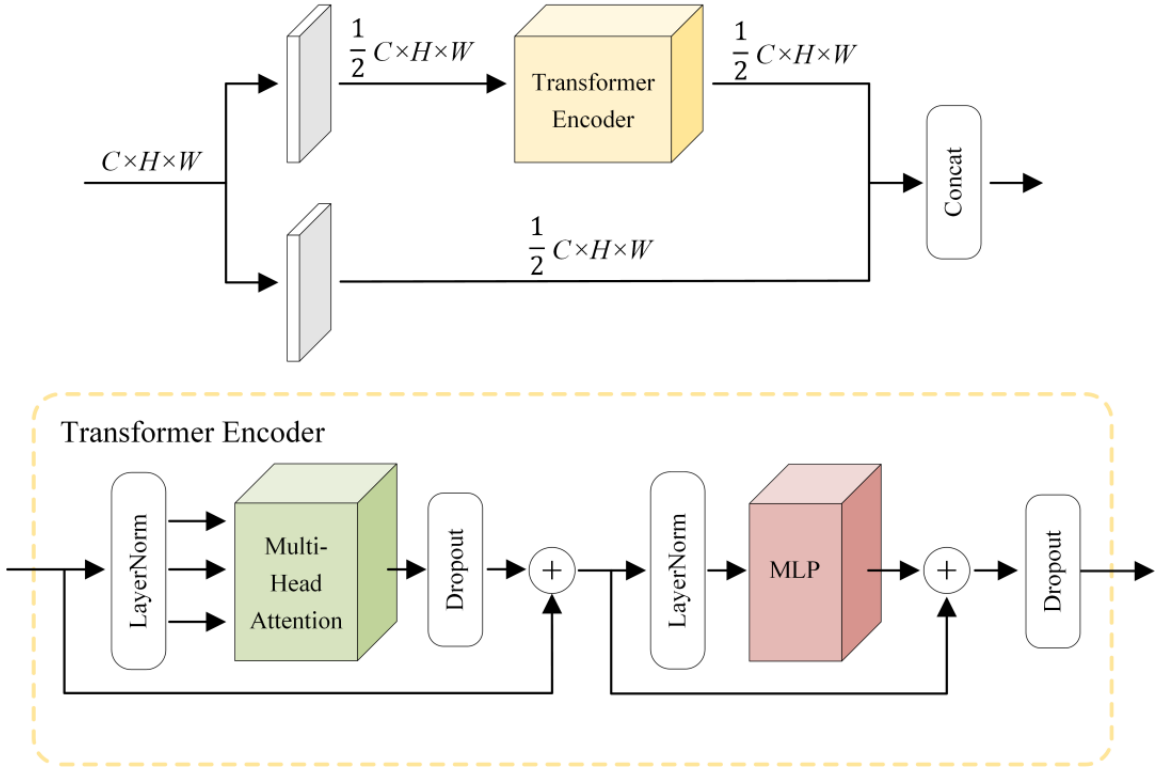


Figure 4. The structure of FTM mainly consists of a Multi-Head Attention block and a Multi-Layer Perception block in a Transformer Encoder.

The core of the FTM is a transformer encoder, consisting of a Multi-Headed Attention (MHA) block and a Multi-Layer Perception (MLP) block overlay. As the input to the encoder is a sequence with certain orders, we partition feature maps into sequences of specific length and width patches before the encoder. Inside the encoder, MHA enables the network to obtain the location information of surrounding ships by acquiring the relationships between ships under a global receptive field. Due to the higher learning capability of the nonlinear transform, it consists of two fully connected layers with a large number of intermediate hidden units to form an MLP block, which analyzes contextual information and enhances the characterization of ship features. In addition, a residual structure is added to keep the FTM well trained even when the layers are deepened. Layer normalization is employed to normalize the feature sequence so that the ReLu activation function can play a better role afterward. To cope with the structural gradient disappearance problem, the FTM transforms the input features by two 1×1 convolutions, one retaining the original features and the other using the transformer encoder. Compared to dividing the channels directly, such a division allows all the input features to be transformed, improves the reusability of features effectively, and keeps the overall computing effort lower. Using the FTM at the top layer of the backbone network before inputting to the neck, through continuous learning, contextual information is linked to enhance the correlation between ships, and thus reduce the omission of small-scale ships and improve the network's detection ability.

2.3. Overall Structure of YOLO-SD

The overall framework of YOLO-SD is shown in Figure 5. After the SAR images are input into our model, feature extraction is first implemented by the backbone network, a modified CSPDarknet. The MSC and FTM of the design are introduced to it, which means that the final feature map contains more valid small-scale ship features. In the neck network, the first fusion was performed from top to bottom to obtain P3, P4, and P5. To retain the shallow edge, shape, and other features, a bottom-up path enhancement structure was added later and achieved the second fusion to obtain N3, N4, and N5 feature maps.

Finally, decoupling heads separate the classification and regression tasks to obtain more accurate detection results for small ships in SAR images.

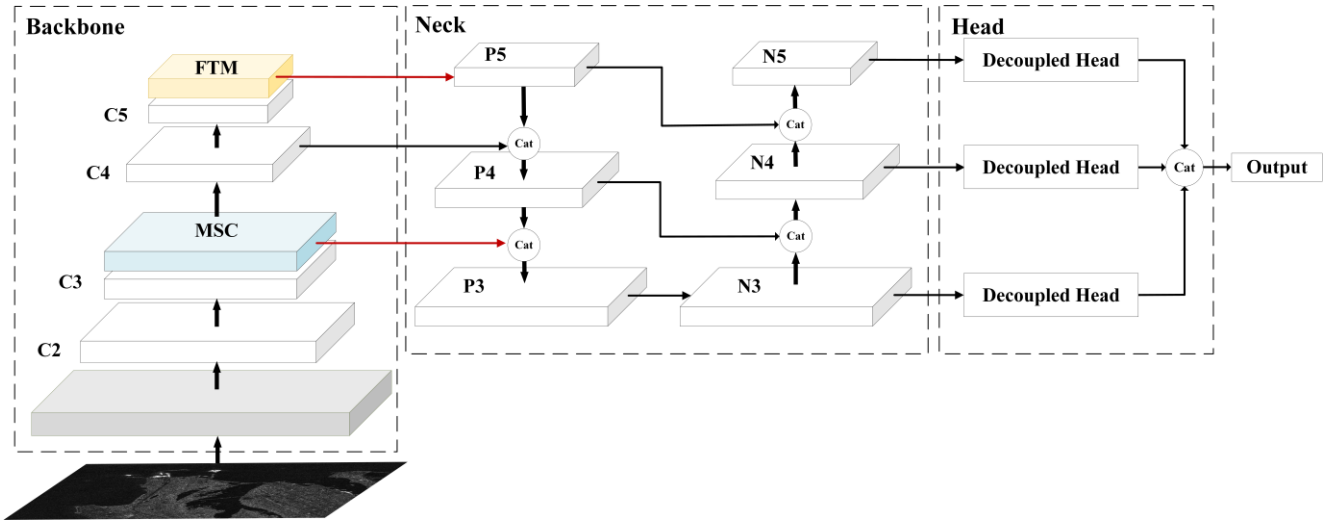


Figure 5. The overall structure of YOLO-SD, where MSC and FTM are added to the backbone, and the connections between the backbone and neck are modified.

2.4. Dataset

In order to test the practical effect of YOLO-SD, we employed the high-resolution SAR image dataset (HRSID) and LS-SSDD-v1.0 dataset. The specific information of these two datasets is shown in Table 1. The HRSID, including a total of 16,951 ship targets, cuts 136 panoramic SAR images into 5604 images with 800 × 800 pixels. In the LS-SSDD-v1.0 dataset, 15 images with 24,000 × 16,000 pixels are cut into 9000 sub-images, also with 800 × 800 pixels. The LS-SSDD-v1.0 dataset retains the pure background image, so the detection model can learn pure background features more effectively and reduce false alarms. The SAR images in two datasets were collected from Sentinel-1 and TerraSAR-X satellites with mixed HH, HV, VV and VH polarizations. With the help of Google Earth and the Automatic Identification System (AIS), all the ships in a SAR image can be completely labeled. When experimenting, the ratio of dataset division (training dataset:validation set:test set) was set to 13:7:7 and 2:1:1, respectively. Figure 6 shows the comparison of the number of ships of three sizes in the two datasets. When the datasets processed the ship targets, they were divided into three types according to the area size: small ships (area less than 48^2 pixels), medium ships (area between 48^2 and 145^2 pixels), and large ships (area greater than 145^2 pixels). According to Figure 6, the typical size of small ships in the two datasets is 48^2 pixels.

Table 1. Description of the two datasets used.

Parameter		HRSID	LS-SSDD-v1.0
Resolution (m)		0.5, 1, 3	5 × 20
Total Images		5604	9000
Total Ships		16,951	6015
Image Size		800 × 800 pixels	800 × 800 pixels
Dataset Division (Training Set:Validation Set:Test Set)		13:7:7	2:1:1
Size of Ships	Small	9242	6003
	Medium	7388	12
	Large	321	0

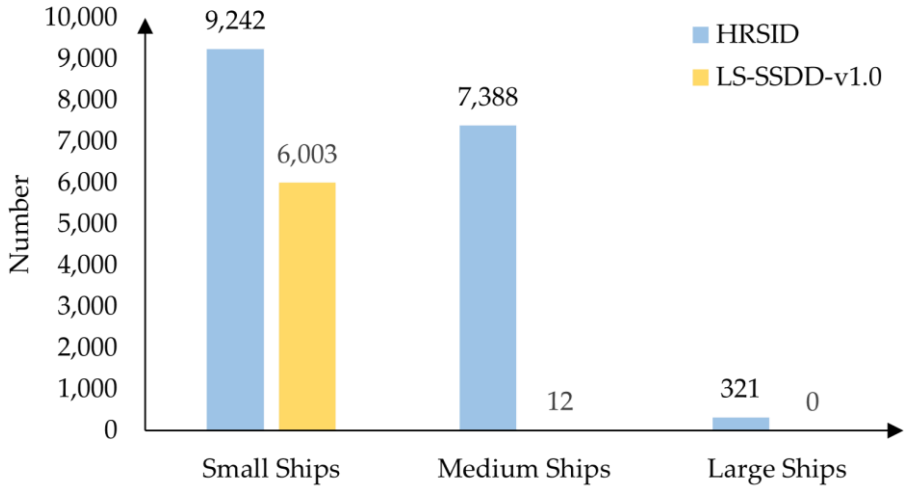


Figure 6. Comparison of the number of ships of three sizes in the two datasets.

2.5. Evaluation Metrics

For the accurate evaluation of the detection performance of each model, the indexes, including the MS COCO evaluation indexes [48], FPS, Parameters, and so on, were used in this work.

Intersection Over Union (*IoU*) is an important and standard index to measure the accuracy of object detection in the dataset. Its calculation is defined as follows, where A represents a real object box in the dataset, and B represents the corresponding prediction box obtained by detection models:

$$IoU = \frac{A \cap B}{A \cup B} \tag{3}$$

Recall refers to the proportion of correctly predicted samples in all real objects, while *Precision* means the proportion of correctly predicted samples in the objects targeted predicted by the model. Their calculation methods are shown in Formulas (2) and (3), where *TP* refers to True Positive and *FN* refers to False Negative.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

The widely employed Mean Average Precision (mAP) is the average of the accuracy of all categories. Since there is only one type in the HRSID and LS-SSDD-v1.0 (ship), the result obtained by calculating *AP* is the mAP. The derivation formula of AP is shown in Formula (6), where R represents Recall and P represents *Precision*. Generally speaking, the higher the *AP* corresponding to the model, the better the detection performance of this model. Since *AP* is obtained by integrating $P(R)$ with R, the *Precision-Recall* curve can display the overall performance of algorithms.

$$AP = \int_0^1 P(R)dR \tag{6}$$

This work used MS COCO evaluation indexes to reliably compare the experimental results between different detection models. According to different IoU thresholds and various object characteristics, indexes can be divided into six different types, including AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L. Once the mentioned IoU threshold is set to 0.5 and 0.75, the results obtained by Formula (6) are AP_{50} and AP_{75}. If IoU gradually increases between 0.5 and 0.95 (by 0.05), the average of the ten values obtained is AP. When only objects of a

specific size are calculated, such as small (the area of the detection object is less than 32^2 pixels), medium (32^2 < area < 96^2 pixels), and large (the area is greater than 96^2 pixels) objects, the averages obtained are AP_S, AP_M, and AP_L.

In addition to these indicators, we also introduced some other indexes, such as Frame Per Second (*FPS*) to evaluate detection speed and Parameters to describe model complexity. *FPS* represents the number of images that can be processed per second. The time required to detect each image can be obtained by taking the inverse of the *FPS*, as shown in the following Formula (7). In the CNN network, the parameter can describe the complexity of the model, and its calculation formula is shown in Formula (8). In Formula (8), K_h and K_w represent the size of the convolution kernel, C_{in} means the number of channels of the input feature map, and C_{out} means the number of channels of the output feature map. Therefore, the parameter of a convolutional layer can be obtained by Formula (8), and the parameter of the entire model can be obtained by adding the parameters of all layers.

$$t = \frac{1}{FPS} \tag{7}$$

$$Params = (K_h \cdot K_w \cdot C_{in}) \cdot C_{out} \tag{8}$$

3. Results

3.1. Experimental Environment

All experiments were completed on a server equipped with NVIDIA GeForce RTX 3090 and 24G video memory. Besides this, we used Python 3.7 compilation language, python 1.8.1 to realize our training, and CUDA 11.1 to speed up the calculation. Furthermore, because of the limitations of the hardware and network, the batch size was set to for all experiments, which means that our server for training needed to process four SAR images at a time.

3.2. Training and Testing

In all experiments, our model and all other models were based on the mmdetection platform [49] and applied the same settings. On the HRSID and LS-SSDD-v1.0 datasets, the learning rate was set to 0.001, the number of iterations in the training epochs was 24, the momentum was 0.9, and the weight attenuation decay was 0.0001. We employed the CSPDarkNet53 backbone network parameters pretrained on the ImageNet dataset, and set the input image resolution to 512 × 512, 640 × 640, and 800 × 800 pixels. Some image processing operations were included in the training pipeline, including Mosaic, RandomAffine, MixUp, RandomFlip, Resize, and MixUp processing. For a thorough assessment, MS COCO evaluation metrics were applied to compare the experimental results of each model. In the field of deep learning, the model is considered successful in detecting the target once the IoU between its predicted box and the real object box is higher than the threshold value of 0.5.

3.2.1. Ablation Experiments for YOLO-SD

First, to demonstrate the advance of YOLO-SD more objectively, we conducted ablation experiments on the HRSID dataset and compared it with YOLOX_L. In the experiments, the input image resolution was set to 800 × 800 pixels, and MSC and FTM were excluded from YOLO-SD separately and then trained; the results are shown in Table 2. Using MSC and FTM alone resulted in a 1.2% and 3.2% increase in final ground average accuracy, respectively, and combining them resulted in a 3.8% increase in AP. In addition, the learnable parameters for MSC and FTM are 0.20 M and 5.25 M, respectively. Compared to the baseline YOLOX_L, YOLO-SD has increased the parameters by 5.45 M, about 10%. From the experimental results, it is clear that both MSC alone and FTM can optimize the ship target information in the feature map and effectively improve the detection accuracy for small ships.

Table 2. Results of ablation experiments.

Baseline	MSC	FTM	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
√			55.7	79.8	63.5	58.5	47.3	1.2	**54.15**
√	√		56.9	80.5	65.1	59.7	49.7	0.02	54.35
√		√	58.9	82.8	67.5	61.5	**52.8**	1	59.4
√	√	√	**59.5**	**83.7**	**67.6**	**62.3**	51.9	**1.3**	59.6

Results were evaluated by MS COCO evaluation indexes, "√" represents the model structure used in this test, three "√" represent YOLO-SD we proposed, and bold data is the best result.

3.2.2. Comparison with YOLOX at Different Scales

Based on the concept of network splitting, width refers to the number of output channels of the network, while depth is the number of layers of the network. Depending on the depth and width, YOLOX can be divided into YOLOX_S, YOLOX_M, YOLOX_L, and YOLOX_X. Specifically, the ratio of the number of layers between them is 1:2:3:4 and the ratio of the number of output channels is 2:3:4:5. To obtain better experimental results, we tested these four complexities of YOLOX and trained our model based on the best YOLOX_L. Table 3 shows the results derived from YOLOX and YOLO-SD with different input scales. When the input was 512 × 512 pixels, the AP of YOLO-SD reached 51.9%, which was 2.3% higher than the baseline. When we used 640 × 640 and 800 × 800 pixel images, the increase was 1.2% and 3.8%, respectively, compared with YOLOX_L. When the resolution of the input SAR image increases, the AP obtained by YOLO-SD shows an upward trend. When dealing with low-resolution (512 × 512) SAR images, YOLO-SD has a higher AP than the baseline YOLOX_L, which means it can detect more ships and obtain more accurate results.

Table 3. Comparison of experimental results of YOLOX_L on HRSID.

Size	Mode	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params	Speed/s
512	YOLOX_L	49.3	73.5	55.3	51.0	50.1	3.9	**54.15**	**0.031**
	YOLO-SD	49.7	73.9	56.1	50.9	53.8	4.7	59.6	0.034
	YOLO-SD *	**51.9**	**77.1**	**58.3**	**53.4**	**55.0**	**8.1**	59.6	0.034
640	YOLOX_L	53.6	77.9	60.6	56.0	52.1	1.9	**54.15**	**0.034**
	YOLO-SD	54.7	78.3	61.6	56.7	**54.5**	1.6	59.6	0.038
	YOLO-SD *	**54.8**	**80.4**	**62.0**	**56.9**	54.0	**5.3**	59.6	0.038
800	YOLOX_L	55.7	79.8	63.5	58.5	47.3	1.2	**54.15**	**0.035**
	YOLO-SD	56.2	80.2	64.1	59.1	48.1	0.6	59.6	0.039
	YOLO-SD *	**59.5**	**83.7**	**67.6**	**62.3**	**51.9**	**1.3**	59.6	0.039

Results were evaluated by MS COCO evaluation indexes, * indicates it loaded the parameters pretrained on HRSID, and bold data is the best result.

According to the results, the Precision-Recall curves of each model have been drawn as shown in Figure 7. The area between the curve and the two coordinate axes is AP. From Figure 7, we can see that the AP of our YOLO-SD includes all the results of YOLOX. All these results suggest that the AP of YOLO-SD is higher, and the overall performance of YOLO-SD is significantly better than all kinds of YOLOX.

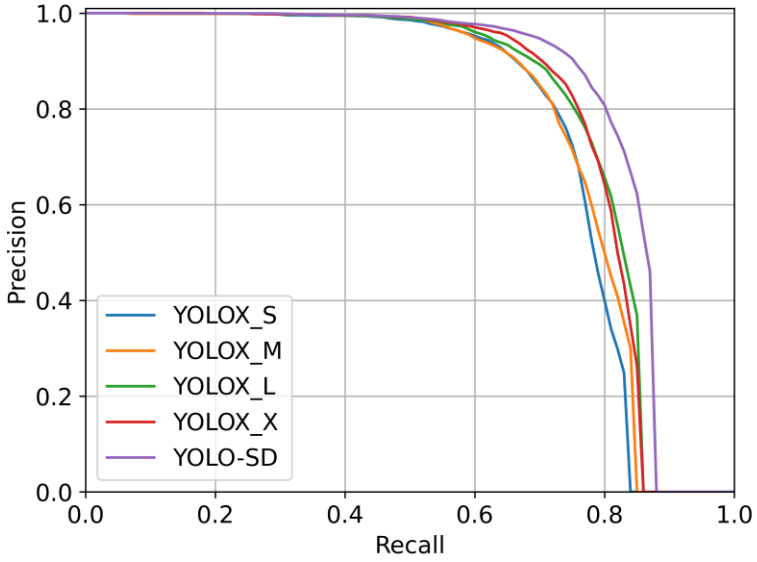


Figure 7. Comparison of the P-R curves of YOLOX with different complexity.

3.2.3. Comparison with Some Existing Models

In addition, we also tested some other excellent algorithms on the HRSID and LS-SSDD-v1.0 datasets to compare their performance with ours, as shown in Table 4. In contrast to Faster R-CNN, YOLO-SD achieved an AP improvement of 1.8% and took almost equal time. Compared to Libra Faster R-CNN, Mask R-CNN, Dynamic R-CNN, and Grid R-CNN, YOLO-SD was not only faster, but also had better accuracy. When facing one-stage algorithms, such as YOLOF, YOLOv3, RetinaNet, and YOLOX_L, although our computational speed was not superior, the accuracy improvement was larger, by 13.9%, 9%, 1.2%, and 3.8%, respectively. Furthermore, in terms of the average precision AP_S obtained for small-scale object detection, YOLO-SD achieved 62.3%, a 3.8% improvement compared to the baseline and 2.3% higher than Dynamic R-CNN, and performs best in this respect.

Table 4. Results of various models on HRSID.

Mode	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params	Speed/s
Faster R-CNN	57.7	79.4	65.9	57.8	64.2	**24.7**	41.12	0.037
Libra Faster R-CNN	58.3	78.8	66.6	58.8	**64.3**	23.0	41.39	0.040
Mask R-CNN	59.3	81.1	67.7	59.9	63.7	13.9	43.75	0.041
Dynamic R-CNN	59.1	80.3	**68.3**	60.0	63.8	22.7	41.12	0.034
Grid R-CNN	59.0	78.6	67.3	59.4	65.8	24.0	64.24	0.045
YOLOF	45.6	68.1	52.2	45.2	55.0	8.9	42.06	**0.023**
YOLOv3	50.5	80.2	54.6	53.7	45.8	0.017	61.52	0.029
RetinaNet	58.3	82.2	64.5	59.3	62.6	21.4	**36.33**	0.033
YOLOX_L	55.7	79.8	63.5	58.5	47.3	1.2	54.15	0.035
YOLO-SD *	**59.5**	**83.7**	67.6	**62.3**	51.9	1.3	59.6	0.039

Results were evaluated by MS COCO evaluation indexes, * indicates it loaded the parameters pretrained on HRSID, and bold data is the best result.

The results of the experiments on the LS-SSDD-v1.0 dataset are shown in Table 5. This dataset contains a high number of small ship annotations, accounting for 99.8% of all annotations. Therefore, it allows a more direct comparison of the ship detection of each model for small targets. As there is no target box larger than 96^2 pixels, the AP_L results are all indicated using "-". As can be seen from Table 5, our remaining five accuracy metrics increased by 1.4%, 2.4%, 2.8%, 1%, and 3.3%, respectively, compared to the baseline. YOLO-SD achieved the highest accuracy at a moderate computational speed compared

with the other models. The experimental results prove that YOLO-SD has the best detection capability for small ships.

Table 5. Results of various models on LS-SSDD-v1.0.

Mode	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params	Speed/s
Faster R-CNN	25.6	66.6	10.8	24.6	37.4	-	41.12	0.037
Libra Faster R-CNN	25.1	65.9	10.0	24.0	38.5	-	41.39	0.039
Mask R-CNN	27.1	70.1	13.4	26.0	40	-	43.75	0.041
Dynamic R-CNN	26.7	69.6	12.1	25.3	38.8	-	41.12	0.034
Grid R-CNN	27.2	70.3	11.3	25.6	40.1	-	64.24	0.045
YOLOF	16.6	50.7	3.9	15.0	31.8	-	42.06	**0.023**
YOLOv3	20.8	58.0	7.8	20.2	31.5	-	61.52	0.029
RetinaNet	21.9	60.7	7.2	20.2	36.7	-	**36.33**	0.033
YOLOX_L	28.3	72.0	12.2	27.1	39.9	-	54.15	0.035
YOLO-SD *	**29.7**	**74.4**	**15.0**	**28.1**	**43.2**	-	59.60	0.038

Results were evaluated by MS COCO evaluation indexes, * indicates it loaded the parameters pretrained on HRSID, and bold data is the best result.

Demonstrating the advantages of YOLO-SD more visually, we plotted the P-R curves of some of the single-stage models and two-stage models separately based on the experimental results on HRSID. Compared with two-stage models (Figure 8a), the Precision of YOLO-SD is significantly higher in the Recall region of 0.8~0.9. Compared with one-stage models (Figure 8b), our Precision is slightly high in the Recall region of 0.5~0.8. The above results demonstrate that the MSC block and the FTM block can effectively improve the accuracy of small ship detection in SAR images.

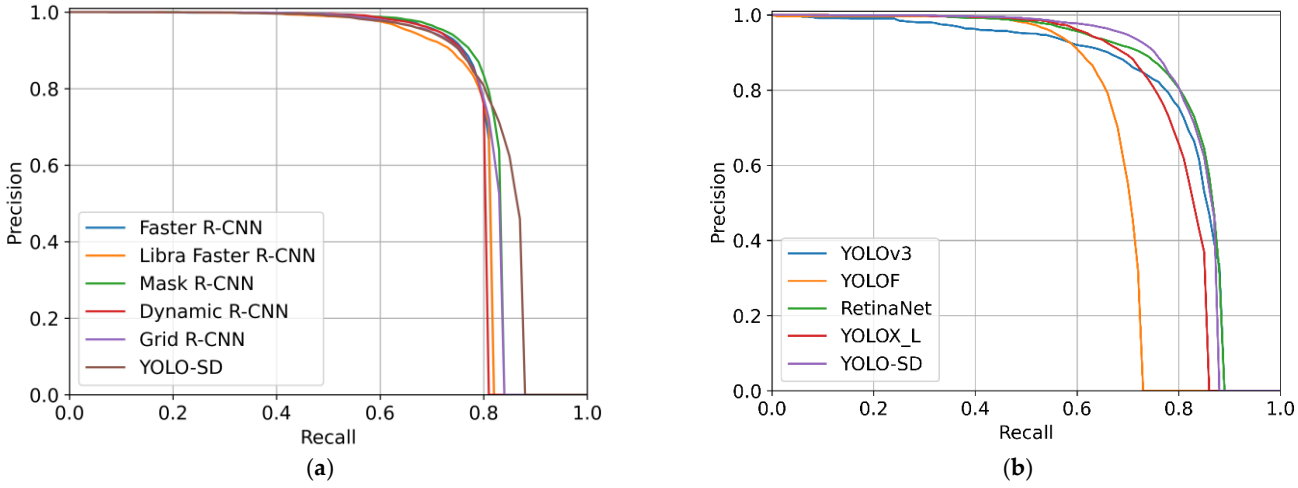


Figure 8. Comparison of Precision-Recall curves of different models. (**a**) Two-stage models. (**b**) One-stage models.

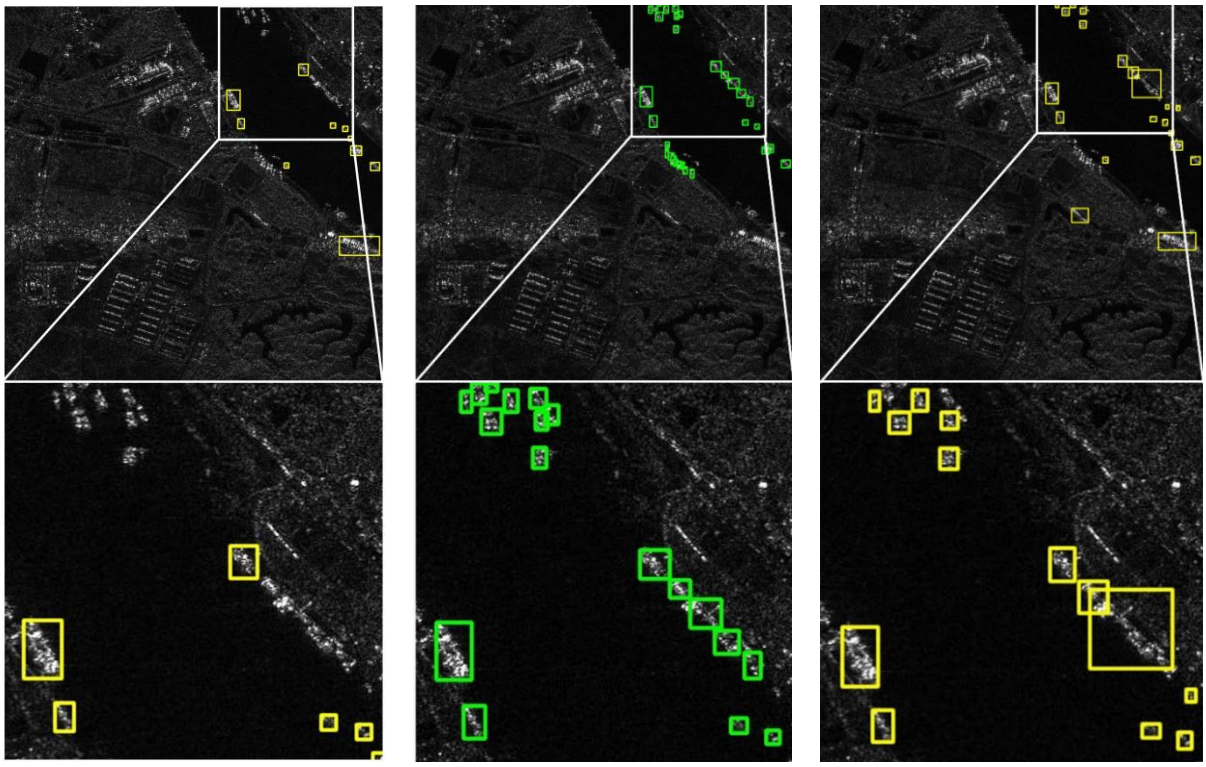
3.3. Detection Results and Analysis

3.3.1. Comparison with YOLOX_L

To demonstrate the superiority of YOLO-SD over the baseline YOLOX_L, some of the visual ship detection results on HRSID are displayed later. We applied two SAR images taken at a canal to detect and compare during the test. In addition, the most representative part in the original size image was selected to be magnified, which is thought to enhance the comparison between YOLO-SD, the baseline, and the true annotated boxes. In detail, the first column of the figures shows the results obtained using the baseline detection, the second column shows the real annotated boxes in the dataset, and the third column shows the results detected by our method. The first row displays all the original images, while

the second row shows the results after zooming in on the intercepted part. To illustrate the superiority of YOLO-SD, some of the visual ship detection results are displayed later.

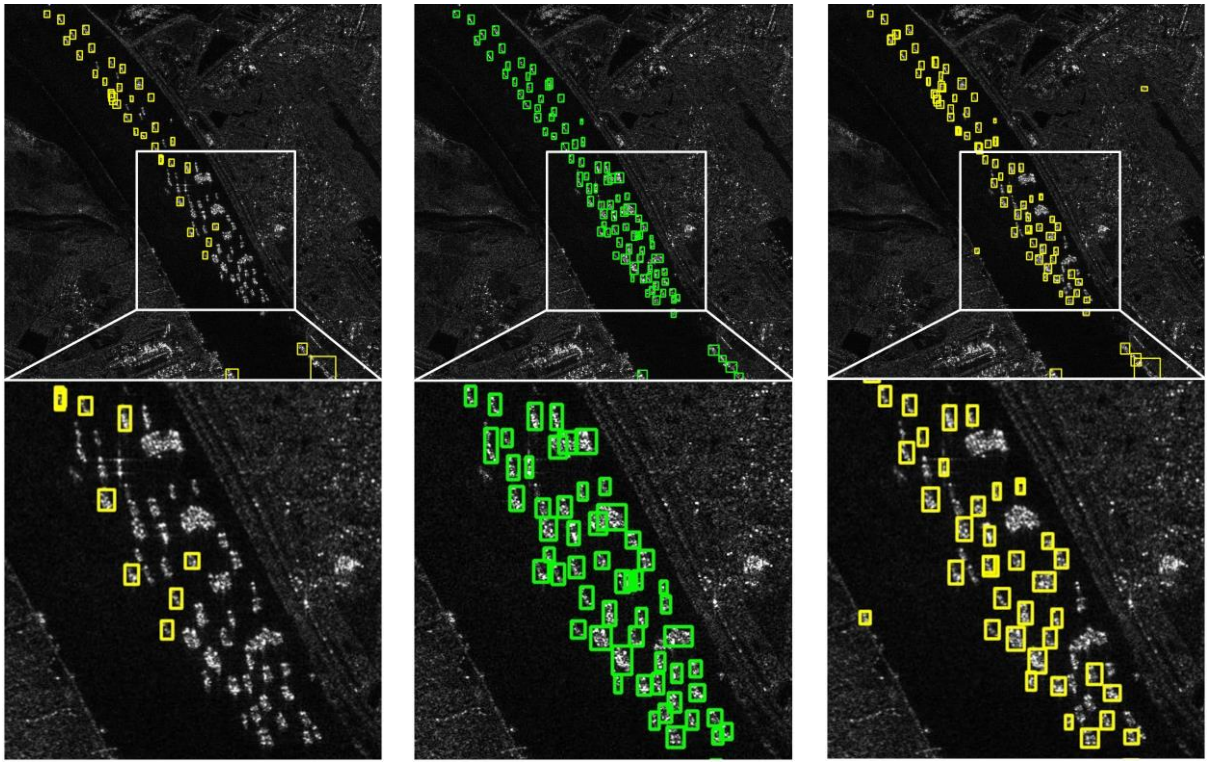
Figure 9 shows the first SAR image of the canal. Where the boats are small and densely distributed on both sides of this river (the enlarged part in Figure 9), our model (Figure 9c) detects more accurately. However, YOLO-SD produces partial false detections when dealing with some larger disturbances.



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

Figure 9. Results in the first river scenario, where the green boxes represent the ship positions marked in the dataset and the yellow boxes represent the detection results inferred by the detection model. (**a**) inference by YOLOX_L, (**b**) the real target frame marked in the dataset, and (**c**) inference by YOLO-SD.

In Figure 10, there is a large number of small ships sailing in the canal. They are distributed sparsely in the upper part of the river, and both YOLOX_L and YOLO-SD perform well. However, in the lower part of the river, YOLOX_L (Figure 10a) produced a huge number of missed detections due to the extremely narrow spacing between ships, and YOLO-SD (Figure 10c) maintained its excellent performance in detecting most of the ships. Table 6 records the detailed data of YOLOX_L and YOLO-SD detection in these two scenarios, where correct refers to the number of correct boxes and wrong refers to the number of boxes detected in error. The greater accuracy of YOLO-SD suggests that our model has an advantage over YOLOX_L in detecting densely distributed small ships.



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

Figure 10. Results in the second river scenario, where the green boxes represent the ship positions marked in the dataset and the yellow boxes represent the detection results inferred by the detection model. (**a**) inference by YOLOX_L, (**b**) the real target frame marked in the dataset, and (**c**) inference by YOLO-SD.

Table 6. Specific detection results of Figures 9 and 10.

Image	Model	Correct	Wrong
Figure 9	Ground Truth	31	-
	YOLOX_L	7	**2**
	YOLO-SD	**14**	5
Figure 10	Ground Truth	92	-
	YOLOX_L	34	**0**
	YOLO-SD	**65**	2

Correct indicates the number of ship targets successfully detected, Wrong means the number of false detections, and bold data is the best result.

3.3.2. Comparison with Other models

To compare the performances of each model, we then took one image from each dataset, and the visual results are shown in Figures 11 and 12. Figure 11 contains fewer distractions and a large number of clear small-scale ships. As the small islands and reefs occupy few pixels, many models incorrectly recognized them as ships (Figure 11d–i). Large and small-scale ships are both included in Figure 12, and there are also some land structures in the upper right section, which may interfere with detection. As the test results show, all the models found large ships, but missed some of the small ships and incorrectly identified land structures as ships.

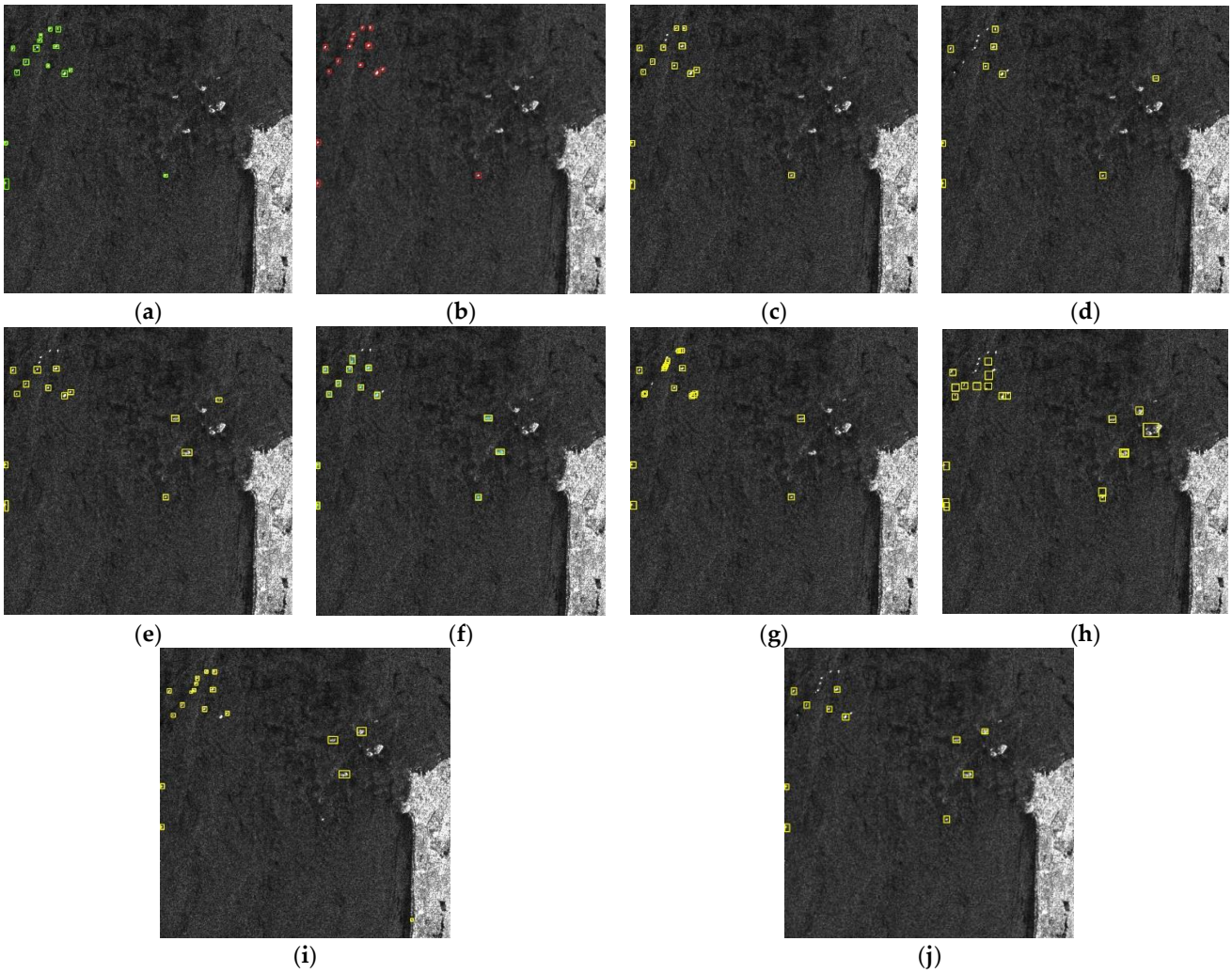


Figure 11. Partial detection results in LS-SSDD-v1.0, where the green boxes represent the ship positions marked in the dataset, the red boxes show the results detected by YOLO-SD and the yellow boxes represent the results inferred by other detection models. (**a**) ground truth; (**b**) YOLO-SD; (**c**) Dynamic R-CNN; (**d**) Faster R-CNN; (**e**) Grid R-CNN; (**f**) Mask R-CNN; (**g**) RetinaNet; (**h**) YOLOF; (**i**) YOLOv3; (**j**) Libra Faster R-CNN.

Table 7 records the detection results obtained for YOLO-SD with other models. Only our algorithm succeeded in identifying not only the larger objects but also the smaller ships, without any false detection. In all the results, our model performs the best in terms of detection, even when compared to Mask R-CNN and Grid R-CNN (two models with high AP in Tables 4 and 5).

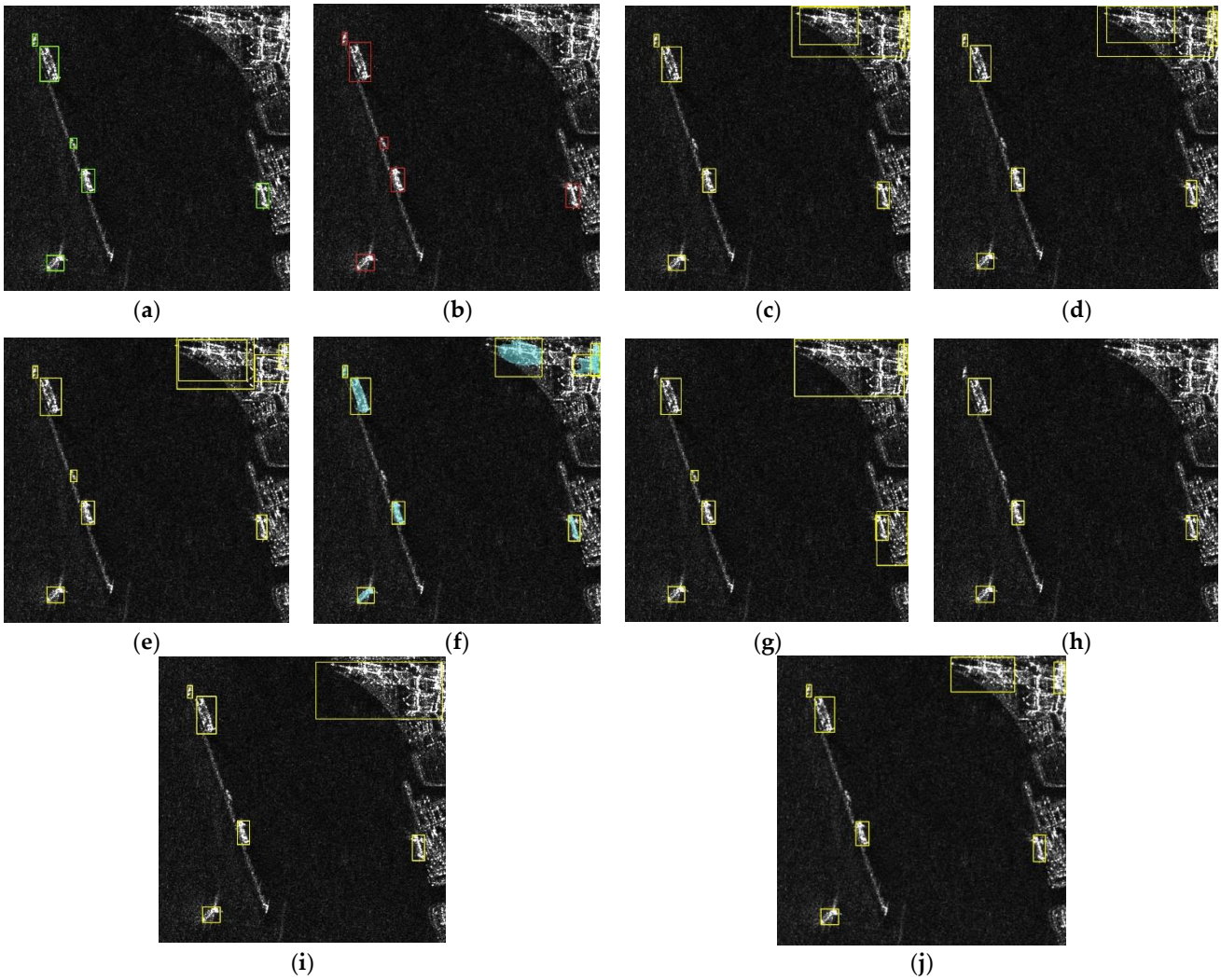


Figure 12. Partial detection results in HRSID, where the green boxes represent the ship positions marked in the dataset, the red boxes show the results detected by YOLO-SD and the yellow boxes represent the results inferred by other detection models. (**a**) ground truth; (**b**) YOLO-SD; (**c**) Dynamic R-CNN; (**d**) Faster R-CNN; (**e**) Grid R-CNN; (**f**) Mask R-CNN; (**g**) RetinaNet; (**h**) YOLOF; (**i**) YOLOv3; (**j**) Libra Faster R-CNN.

Table 7. Specific detection results of Figures 11 and 12.

Model	Figure 11		Figure 12	
	Correct	Wrong	Correct	Wrong
Ground Truth	15	-	6	-
YOLO-SD	**15**	**0**	**6**	**0**
Dynamic R-CNN	13	**0**	5	3
Faster R-CNN	8	1	5	3
Grid R-CNN	11	3	6	3
Mask R-CNN	10	3	5	3
RetinaNet	12	14	5	3
YOLOF	9	10	4	**0**
YOLOv3	13	3	5	1
Libra Faster R-CNN	8	3	5	2

Correct indicates the number of ship targets successfully detected, Wrong means the number of false detections, and bold data is the best result.

4. Discussion

As can be seen from the ablation experiments described in Table 3, both MSC and the FTM contribute to the improved accuracy of ship detection. The FTM increases the AP metric by 3.2%, which is 2% higher than MSC's 1.2%. The introduction of MSC improved the representation of feature maps in the backbone network, while also optimizing the fusion effect of the neck. With a large global field of perception, the FTM focuses on the feature information of the ship, while enhancing the correlation between ships, ultimately reducing the number of missed small objects. With a large global receptive field, the FTM focuses on the feature information of the ship while enhancing the correlation between ships, ultimately reducing the number of missed small objects. Due to the small area of ships in SAR images, feature loss becomes severe as the network deepens. Therefore, the FTM, which focuses on optimizing the depth of the network, can obtain higher detection accuracy. The two act at different network locations and there is no conflict between them, so YOLO-SD can achieve an accuracy improvement of up to 3.8%.

For YOLOX, different complexities have different advantages. YOLOX_S has the smallest number of parameters and the fastest calculation speed, while YOLOX_X has the highest detection accuracy. By comparing the PR curves, we objectively and equitably arrive at the best ground scale for YOLOX-L, and use it as the baseline for improvement. As the size of the input image was altered (from 800 to 640 and 512), the detection accuracy of all models decreased by varying degrees. In fact, when the SAR image was reduced, the ship object area became smaller and contained less feature information, making detection more difficult. Whereas MSC enables the model to detect targets from indistinguishable complex backgrounds by exploiting the rich semantic information, the FTM improves the correlation between all ships and enables the model to detect a larger number of ship objects. As a result, compared to the baseline, YOLO-SD is more capable of detecting ships and consistently obtains the highest AP.

In practical experiments on two different datasets, our model maintained accurate detection results. YOLO-SD also has the highest detection accuracy compared to other superior models, which proves its advantages.

5. Conclusions

Given the poor effect of existing models on small ship detection in complex SAR images, we propose an improved detection model based on YOLOX in this paper, named YOLO-SD. It is combined with MSC to extract different scale features and enrich semantic information, and the FTM block to optimize features. On the HRSID and LS-SSDD-v1.0, several sets of experiments were conducted to compare with the highly representative detection methods, including Faster R-CNN, Mask R-CNN, RetinaNet, etc. The experimental results show that our network performs better and achieves the highest accuracy when dealing with small ships. However, we still found some missed detection when dealing with unclear ships. In the future, we will conduct further research on small-scale ship detection methods with lower leakage rates. It is hoped that this article can help scholars find better ideas when analyzing and dealing with scenes containing dense, small targets. Both our code and the datasets used (HRSID and LS-SSDD-v1.0) are available at the link: https://doi.org/10.6084/m9.figshare.21316290.v3 (accessed on 17 October 2022).

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

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 41930112.

Data Availability Statement: Not applicable.

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

Abbreviations

All abbreviations mentioned in this article are listed below:

SAR	Synthetic Aperture Radar
MSC	Multi-Scale Convolution
FTM	Feature Transformer Module
HRSID	High-Resolution SAR Images Dataset
CFAR	Constant False Alarm Rate
CNN	Convolutional Neural Network
R-CNN	Region-CNN
YOLO	You Only Look Once
YOLOF	You Only Look One-level Feature
NLP	Natural Language Processing
RNN	Recurrent Neural Network
CV	Computer Vision
ViT	Vision Transformer
DETR	Detection Transformer
SETR	Segmentation Transformer
SSD	Single Shot Detector
CBAM	Convolutional Block Attention Module
RFB	Receptive Fields Block
FPN	Feature Pyramid Network
PAFPN	Path Aggregation Feature Pyramid Network
MLP	Multilayer Perceptron
FPS	Frames Per Second
IoU	Intersection over Union
TP	True Positive
TN	True Negative
FN	False Negative
FP	False Positive
MS COCO	Microsoft Common Objects in Context
P-R Curve	Precision-Recall Curve

References

1. Moreira, A.; Prats-Iraola, P.; Younis, M.; Krieger, G.; Hajnsek, I.; Papathanassiou, K.P. A tutorial on synthetic aperture radar. *IEEE Trans. Geosci. Remote Sens.* **2013**, *1*, 6–43. [CrossRef]
2. Guo, Q.; Wang, H.P.; Xu, F. Research progress on aircraft detection and recognition in SAR imagery. *J. Radars* **2020**, *9*, 497–513.
3. 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]
4. 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]
5. Novak, L.M.; Owirka, G.J.; Brower, W.S.; Weaver, A.L. The automatic target-recognition system in SAIP. *Linc. Lab. J.* **1997**, *10*. Available online: http://www.geo.uzh.ch/microsite/rsl-documents/research/SARlab/GMTILiterature/PDF/NOBW97.pdf (accessed on 17 October 2022).
6. Xu, P.; Li, Q.; Zhang, B.; Wu, F.; Zhao, K.; Du, X.; Yang, C.; Zhong, R. On-Board Real-Time Ship Detection in HISEA-1 SAR Images Based on CFAR and Lightweight Deep Learning. *Remote Sens.* **2021**, *13*, 1995. [CrossRef]
7. Novak, L.M.; Owirka, G.J.; Netishen, C.M. Performance of a high-resolution polarimetric SAR automatic target recognition system. *Linc. Lab. J.* **1993**, *6*, 11–24.
8. Qin, X.; Zhou, S.; Zou, H.; Gao, G. A CFAR detection algorithm for generalized gamma distributed background in high-resolution SAR images. *IEEE Geosci. Remote Sens. Lett.* **2012**, *10*, 806–810.
9. Nunziata, F.; Migliaccio, M. On the COSMO-SkyMed PingPong mode to observe metallic targets at sea. *IEEE J. Ocean. Eng.* **2012**, *38*, 71–79. [CrossRef]
10. Ferrara, G.; Migliaccio, M.; Nunziata, F.; Sorrentino, A. Generalized-K (GK)-based observation of metallic objects at sea in full-resolution synthetic aperture radar (SAR) data: A multipolarization study. *IEEE J. Ocean. Eng.* **2011**, *36*, 195–204. [CrossRef]
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
12. Wu, X.; Sahoo, D.; Hoi, S.C. Recent advances in deep learning for object detection. *Neurocomputing* **2020**, *396*, 39–64. [CrossRef]

13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *39*, 1137–1149. [CrossRef]
15. Nie, X.; Duan, M.; Ding, H.; Hu, B.; Wong, E.K. Attention mask R-CNN for ship detection and segmentation from remote sensing images. *IEEE Access* **2020**, *8*, 9325–9334. [CrossRef]
16. 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 European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 260–275.
17. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 14454–14463.
18. 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 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 821–830.
19. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
20. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
21. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
22. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 13039–13048.
23. 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 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed on 17 October 2022).
25. Yin, W.; Kann, K.; Yu, M.; Schütze, H. Comparative study of CNN and RNN for natural language processing. *arXiv* **2017**, arXiv:1702.01923.
26. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 213–229.
27. Li, J.; Qu, C.; Shao, J. Ship detection in SAR images based on an improved faster R-CNN. In Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), Beijing, China, 13–14 November 2017; pp. 1–6.
28. Xu, Z.; Sun, K.; Mao, J. Research on ResNet101 network chemical reagent label image classification based on transfer learning. In Proceedings of the 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Weihai, China, 14–16 October 2020; pp. 354–358.
29. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck transformers for visual recognition. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 16519–16529.
30. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
31. Wang, J.; Zheng, T.; Lei, P.; Bai, X. Ground target classification in noisy SAR images using convolutional neural networks. *IEEE J. Sel. Top Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4180–4192. [CrossRef]
32. 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]
33. 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]
34. 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 European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
35. Juan, S.; Long, Y.; Hua, H. Improved SSD algorithm for small-sized SAR ship detection. *J. Syst. Eng. Electron.* **2020**, *42*, 1026–1034.
36. 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]
37. 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.* **2021**, *60*, 5217712. [CrossRef]
38. Jin, K.; Chen, Y.; Xu, B.; Yin, J.; Wang, X.; Yang, J. A patch-topixel convolutional neural network for small ship detection with PolSAR images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 6623–6638. [CrossRef]
39. Fu, J.; Sun, X.; Wang, Z.; Fu, K. An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 1331–1344. [CrossRef]
40. Guo, H.; Yang, X.; Wang, N.; Gao, X. A CenterNet++ model for ship detection in SAR images. *Pattern Recognit.* **2021**, *112*, 107787. [CrossRef]

41. Chang, Y.L.; Anagaw, A.; Chang, L.; Wang, Y.C.; Hsiao, C.Y.; Lee, W.H. Ship detection based on YOLOv2 for SAR imagery. *Remote Sens.* **2019**, *11*, 786. [CrossRef]
42. 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]
43. Zhao, J.; Guo, W.; Zhang, Z.; Yu, W. A coupled convolutional neural network for small and densely clustered ship detection in SAR images. *Sci. China Technol. Sci.* **2019**, *62*, 42301. [CrossRef]
44. Cui, Z.; Li, Q.; Cao, Z.; Liu, N. Dense attention pyramid networks for multi-scale ship detection in SAR images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8983–8997. [CrossRef]
45. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 269–284.
46. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–15 June 2015; pp. 1–9.
47. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. Cspnet: A new backbone that can enhance learning capability of cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 390–391.
48. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
49. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Lin, D. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.