A Novel CNN-Based Detector for Ship Detection Based on Rotatable Bounding Box in SAR Images

Rong Yang, Zhenru Pan, Xiaoxue Jia, Lei Zhang, and Yunkai Deng, *Member, IEEE*

***Abstract*—Thanks to the excellent feature representation capabilities of neural networks, deep learning-based methods perform far better than traditional methods on target detection tasks such as ship detection. Although various network models have been proposed for SAR ship detection such as DRBox-v1, DRBox-v2, and MSR2N, there are still some problems such as mismatch of feature scale, contradictions between different learning tasks, and unbalanced distribution of positive samples, which have not been mentioned in these studies. In this article, an improved one-stage object detection framework based on RetinaNet and rotatable bounding box (RBox), which is referred as R-RetinaNet, is proposed to solve the above problems. The main improvements of R-RetinaNet as well as the contributions of this article are threefold. First, a scale calibration method is proposed to align the scale distribution of the output backbone feature map with the scale distribution of the targets. Second, a feature fusion network based on task-wise attention feature pyramid network is designed to decouple the feature optimization process of different tasks, which alleviates the conflict between different learning goals. Finally, an adaptive intersection over union (IoU) threshold training method is proposed for RBox-based model to correct the unbalanced distribution of positive samples caused by the fixed IoU threshold on RBox. Experimental results show that our method obtains 13.26%, 9.49%, 8.92%, and 4.55% gains in average precision under an IoU threshold of 0.5 on the public SAR ship detection dataset compared with four state-of-the-art RBox-based methods, respectively.**

***Index Terms*—Neural network, rotatable bounding box (RBox), synthetic aperture radar, target detection.**

I. Introduction

SYNTHETIC aperture radar can work under all-weather and day-and-night conditions and make very high resolution images. Therefore, it plays an important role in remote sensing information extraction and is particularly suitable for remote monitoring. With the rapid development of spaceborne SAR system such as TerraSAR-X, RADARSAT-2, Sentinel-1, and Gaofen-3, SAR has been widely used in civil and military fields for target classification, reconnaissance, surveillance. However, because of the special imaging mechanism, SAR images may become unrecognizable for human in some cases, which makes searching for targets of interest in massive SAR images by eyes become time-consuming and often impractical. Consequently, how to detect and identify various targets quickly and accurately in SAR images has been the focus of research. Researches on ship detection are vital in many areas [1], such as marine monitoring, maritime management, and military intelligence acquisition. Many investigations that relate to ship detection in SAR imagery have been carried out recently [2], [3].

Traditional ship detection methods mainly rely on statistical analysis of image pixels, and most of them are threshold-based methods [4]–[6]. The threshold-based methods determine the threshold that distinguishes ship targets from the background by modeling sea clutter based on the theory of constant false alarm rate (CFAR) filtering [7], [8], which have become the classic methods for SAR image target detection, and have been widely used in practical ship target detection systems [9]. Sea clutter model with higher complexity usually has higher fitting accuracy, but it may cause difficulty in parameter estimation, so the researches on CFAR-based methods in recent years mainly focus on the tradeoff between the accuracy of sea clutter modeling and the computational complexity [10], [11]. In addition to the CFAR-based methods, Li and Zelnio [12] proposed a method based on the generalized-likelihood ratio test. This method requires statistical modeling of both ship targets and clutter. However, it is difficult to build a unified target statistical model because of the different shapes, sizes, and directions of ship targets, which limits the application of this method. Threshold-based methods work well in homogeneous areas such as offshore area, but they perform poorly in heterogeneous areas such as ports, and often require assistance from shoreline segmentation to obtain better results.

Some of these statistical analysis-based methods use different statistical properties of targets and backgrounds such as standard deviation and noncircularity in the images [13], [14]. In addition, principal component analysis [15] and Bayesian theory [16] have also been used to extract various statistical characteristics. These techniques can enhance the robustness of the detection algorithm, but they may require some prior knowledge. Besides, researchers have also conducted a lot of research on ship detection by combining SAR images under different polarization channels [17], [18].

Manuscript received October 31, 2020; revised December 11, 2020; accepted January 4, 2021. Date of publication January 8, 2021; date of current version January 29, 2021. This work was supported by the National Natural Science Foundation of China under Grant 61901446. *(Corresponding author: Xiaoxue Jia.)*

Rong Yang and Zhenru Pan are with the Department of Space Microwave Remote Sensing System, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: yangrong16@mails.ucas.ac.cn; panzhenru16@mails.ucas.ac.cn).

Xiaoxue Jia, Lei Zhang, and Yunkai Deng are with the Department of Space Microwave Remote Sensing System, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: xiaoxue_snowing@163.com; 314forever@163.com; ykdeng@mail.ie.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3049851

Compared with single-polarization SAR, polarimetric SAR (PolSAR) can provide information about the target structure from the scattering matrix [19], which helps to improve the performance of SAR object detection. Therefore, researchers have made a lot of attempts to use PolSAR data for ship detection [20]. Common ship detection algorithms for PolSAR images mainly include polarization target decomposition [21], [22], statistical theory methods [23], machine learning methods [24], [25], and subaperture-based methods [26], [27].

These traditional methods make full use of the statistical characteristics of SAR images, and the performance of the detector based on traditional methods can be further improved by prior information. However, there are many inherent disadvantages to traditional methods, which make them difficult to completely meet the practical application requirements of ship detection. On the one hand, hand-designed features often have poor robustness and cannot guarantee stable performance in complex scenarios; on the other hand, traditional methods often require multiple operating steps and are time-consuming, which is unacceptable in real-time ship detection.

Recently, with the development of deep learning, the performance of object detection algorithms based on deep convolutional neural networks (CNNs) are far superior to traditional algorithms [28], [29]. At the same time, with the in-depth study of CNNs and the availability of high-performance computing chips, detection algorithms based on CNNs with different characteristics and applicable scopes have been developed. From the perspective of the network structure, the CNN-based detection algorithm realizes the transition from two-stage models to one-stage models, from top-up structure to top-down structure, and from single-scale network to feature pyramid network (FPN) that implements multiscale detection [30]. Ren *et al.* proposed Faster R-CNN [31] in combination with region proposal network (RPN) to improve the repeated operation of candidate box feature extraction, which has become the current mainstream target detection algorithm. Subsequently, You Only Look Once [32] and single-shot multibox detector (SSD) [33] were proposed in order to improve the speed of the detection algorithm. Such algorithms are based on regression ideas and integrate classification and detection into the same network, which makes real-time detection possible. Furthermore, methods based on visual attention suppress noise in the network and highlight the effective features of targets. The introduction of these technologies further improves the detection accuracy [34].

The successful application of CNNs in the field of target detection has injected great vitality into SAR image ship detection. At the beginning, researchers tried to use CNNs to improve the detection performance of traditional methods. In [2], a full CNN was used to perform land and sea segmentation, and then a CFAR method was used for ship detection, which improved the detection performance of CFAR at the sea-land junction. In [35], Faster R-CNN was improved and combined with CFAR. The improved Faster R-CNN scanned the target area of potential ships and sent it to the protection window of CFAR to improve the detection rate of small targets.

Later, with the access of a large amount of available training data, scholars began to focus on SAR image ship detection methods based entirely on CNNs that utilize end-to-end mode, which improves the efficiency and accuracy of detection. Li *et al.* [36] proposed an improved faster R-CNN method and provided a dataset called SAR ship detection dataset (SSDD), which was widely used in subsequent ship detection researches. Jiao *et al.* [37] performed multiscale and multiscene ship detection using a densely connected network as backbone and introduced Focal Loss to faster R-CNN structure.

In SAR images, the detection difficulty of inshore ships is much greater than that of offshore ships. This is because the backscatter characteristics of many land-based facilities in nearshore scenes are very similar to ship targets. Without the help of the sea-land segmentation operation, the ship detector may misclassify some objects on land as ship targets, which increase the false alarm rate of the detector. To solve this problem, Cui *et al.* [38] proposed a detector based on dense attention pyramid network to improve ship detection performance in inshore areas. With the increase in the complexity and depth of the model, the performance of the model has also been greatly improved. However, the problem of poor real-time performance still exists while using the two-stage models. In response to this problem, Wang *et al.* [39] use the SSD model for ship target detection and trains the network through transfer learning to improve the detection speed, but the detection performance for small targets is poor because the receptive field of the top-level feature map is large, which do not match the scales of small ships. In addition, due to the lack of refinement of the detection results, the detection accuracy based on one-stage methods is often inferior to that based on two-stage detection [40].

Most deep learning based SAR ship detection algorithms use horizontally placed rectangular bounding box (BBox) to locate ship targets [25], [41]–[48]. However, the BBox shows poor performance when the ship targets are close to each other in the coast [49]. At the same time, the overlapped BBoxes may be suppressed by nonmaximum suppression (NMS) operation in the postprocessing steps if the ship targets are densely arranged, which results in missed detection [50]. Driven by this problem, the detection algorithm based on rotatable bounding box (RBox) has become a hot spot recently. Wang *et al.* [49] used an improved SSD model to detect ship targets and estimate the orientation angle of the ship simultaneously, realizing the ship detection based on RBox in SAR images. An *et al.* [51] used an RBox-based detection method to improve the performance of SAR image ship detection, which was originally used to perform ship detection in optical remote sensing images. Pan *et al.* [52] use a multistage network to optimize the localization results of the RBox-based model, which improves the detection accuracy.

Although the introduction of the RBox can effectively improve the detection accuracy of densely distributed ship targets, it also brings new problems to the training of network models, which have not been mentioned in previous studies for SAR ship detection.

First, the combination of multiscale feature maps was widely used for improving the model's detection performance for targets of different sizes in previous studies. Since the RBox-based model generates much more anchors on the feature map than the BBox-based model [51], the feature map with unreasonable

scale will introduce huge computational cost while unable to effectively improve model performance. Besides, the standard one-stage models remove the RPN in order to improve the detection efficiency, which leads to the misalignment between the region of the features used for prediction and the real region of target. This misalignment is an important reason for the poor accuracy of the one-stage model [53]. A feasible method to alleviate this misalignment is to ensure that the scale distribution on the original image corresponding to the small feature map input into the prediction network covers the targets' real scale distribution on the original image as much as possible, which also requires a reasonable feature map scale combination. However, how to quickly find a reasonable feature map scale combination is still a challenge since previous studies only used heuristic solutions for this problem.

Second, the feature map after multiscale fusion is completely shared by the classification branch and the localization branch in the current detection models [34], [37], [38], [40], [42], [49]–[52], which makes the optimization of the fusion feature map fully coupled by different learning tasks. However, classification tasks and localization tasks have completely opposite requirements for the spatial sensitivity of features, which leads to different learning tasks conflicting with each other during the training process [54], [55]. Therefore, trying to decouple the optimization process of different learning tasks on the fused feature maps will be a direction worth exploring.

Third, the anchor-based (BBox-based or RBox-based) detection methods need to generate massive anchors on the image, and at the same time, a fixed intersection over union (IoU) threshold is used to distinguish these anchors into positive samples and negative samples according to their IoU with the target boxes [31], [33], [51]. An IoU threshold that is close to 0 will cause the positive samples to contain many anchors that do not match the targets, whereas an IoU threshold that is close to 1 will cause the lack of positive samples. Both cases are not conducive to model learning, so the anchors that have an IoU over 0.5 with a true target are usually taken as a positive sample in most studies. However, the anchors based on RBox will cause some targets to generate too many or too few positive samples that meet the specific IoU threshold due to their different aspect ratio and various orientation angles, which makes the model focus too much on those targets with more positive samples and ignore targets with fewer positive samples. The above characteristics of the RBox-based detection method exacerbate the imbalance between positive and negative samples in the training stage, which eventually leads to model degradation.

In addition, our previous work [52] is devoted to designing more complex networks such as multistage models to improve the accuracy of ship detection, which are much slower than one-stage models. However, high detection speed is also critical in practical applications such as satellite military reconnaissance mission. How to improve model detection accuracy without sacrificing detection speed remains a challenge to be solved.

In this article, a ship detection method called as R-RetinaNet is proposed aiming for solving the above-mentioned issues for RBox-based ship detection. The main improvements of R-RetinaNet as well as the contributions of this article are mainly reflected from the following aspects:

First, in order to ensure the speed of the detection method, we choose one-stage model as our basic design framework. At the same time, a feature map scale calibration method is proposed to align the scale distribution of the output backbone feature maps with the scale distribution of the targets, which avoids the heuristic operation or exhaustive search during model design process.

Second, inspired by [56], a new pyramid network named task-wise attention feature pyramid network (TA-FPN) was proposed to decouple the optimization process of different learning tasks on the fused feature map, which produced better results than traditional FPN.

Third, an adaptive IoU threshold (AIT) training method is proposed for the training of the model in order to alleviate model degradation caused by the severe imbalance of positive samples on different targets, which significantly improves the model's performance.

The rest of this article is organized as follows. Section II introduces the proposed methods. The experimental results on two datasets and the comparison with several state-of-the-arts methods are explained in Section III. Finally, Section IV concludes the article.

II. Methods

In this part, the details of the proposed network structure as well as the training process of the proposed model and corresponding hyperparameters will be introduced. At first, the key points and the overall architecture of the proposed network are derived by analyzing the problems of some existing methods. Next, the AIT training method is described in detail to show how it works.

A. Overall Scheme of Network Structure

The current mainstream detection models can be divided into one-stage models and two-stage models. The one-stage models are widely used in various application scenarios that require low latency such as video detection due to their faster detection speed compared with two-stage models. In general, the size of images produced by remote sensing systems such as SAR is much larger (e.g., 3000×3000) than that of traditional sensors such as ordinary cameras, which creates a large computational burden for SAR ship detection. Therefore, this article chooses the one-stage model RetinaNet as the basic design framework in order to meet the high-speed requirements of applications such as ship monitoring.

The network structure proposed in this article is illustrated in Fig. 1, which can be divided into four parts: the feature extraction part, the feature fusion part, the prediction part, and the postprocessing part.

1) Feature Extraction: Each training image will be first randomly cut into a 320×320 slice containing at least one target, and then the slice is fed into a feature extraction network to obtain feature maps with different scales and different semantic information. If not specified, the feature extraction network used in our model is ResNet50, which shows a good compromise between calculation and performance in our experiments.

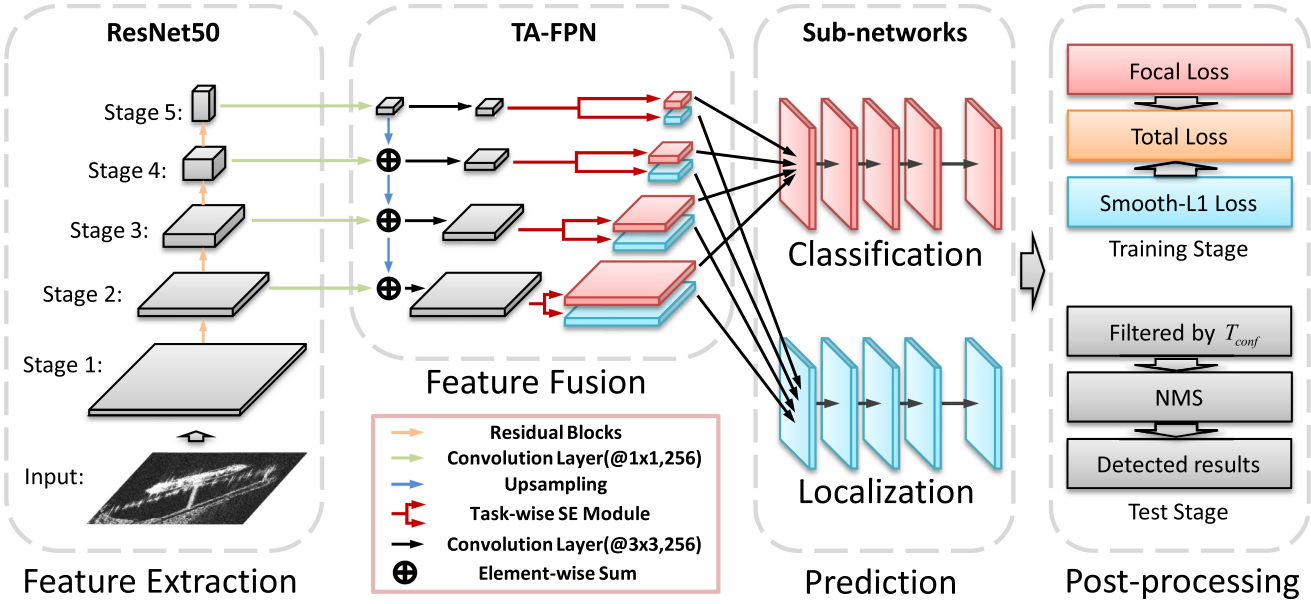


Fig. 1. Architecture of the proposed model R-RetinaNet.

2) Feature Fusion: Next, we will select feature maps that can well match the target scale distribution of dataset and feed them into the feature fusion network named TA-FPN to obtain multi-scale feature maps with strong semantic information. The feature optimization module named task-wise squeeze-and-excitation (SE) module in feature fusion network will attempt to decouple the optimization process of different learning tasks on the multiscale feature maps to a certain extent by performing two different channel optimizations on the feature maps according to different learning tasks.

3) Prediction: The output of the feature fusion network will be input to the classification branch network and the localization branch network, respectively. The classification branch network will predict the confidence of the prior anchors corresponding to each pixel on the feature maps of different scales. This confidence determines the probability that the anchor belongs to a target. Localization branch network outputs the normalized coordinate deviation between the anchor belonging to a target and the true bounding box of that target.

4) Postprocessing: In the training stage, the output of the classification branch and the localization branch will be used to calculate the classification loss and the localization loss, respectively. In the test stage, the anchors with confidence less than T_{conf} will be removed, and then the coordinates of the remaining anchors will be compensated by the output of the localization branch network to obtain the detection results. Finally, NMS operation with a threshold of T_{nms} is performed on the detection results to eliminate detection results with large overlap.

B. Scale Calibration of Output Feature Maps

The standard RetinaNet [57] uses the ResNet50 without the classification layer as the feature extraction network. In addition, standard RetinaNet added two convolutional layers on the last feature layer of ResNet50, which improves the detection performance of large targets. We divide the feature extraction network of RetinaNet into different blocks as shown in Table I for the following explanation. Stage_n indicates that the resolution of the output feature map at this stage is $1/2^n$ times that of the original image, therefore different stages correspond to feature maps of different scales.

Feature maps with different scales need to be output from the backbone network for subsequent processing. In order to detect

TABLE I
FEATURE EXTRACTION NETWORK OF RETINANET

Block Name	Conv Block Setting	Output Size
Input	-	$320 \times 320, 3$
Stage_1	$7 \times 7, 64$, stride 2	$160 \times 160, 64$
Max pool	3×3, stride 2	$80 \times 80, 64$
Stage_2	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$80 \times 80, 256$
Stage_3	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$40 \times 40, 512$
Stage_4	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$20 \times 20, 1024$
Stage_5	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$10 \times 10, 2048$
Stage_6	$3 \times 3, 256$, stride 2	$5 \times 5, 256$
Stage_7	$3 \times 3, 256$, stride 2	$2 \times 2, 256$

TABLE II
SCALE SETTING OF OUTPUT FEATURE MAPS IN DIFFERENT METHODS FOR DATASET SSDD

Method	Backbone Network	Stage of Output Feature Maps
DAPN [38]	ResNet101	{2,3,4,5}
DNN-AM [40]	Customized	{2,3,4}
LSSD [42]	VGG16	{4,5}
R-SSD [49]	VGG16	{3,4,5,6,7,8}
DRBox-v2 [51]	VGG16	{3,4}

ship targets of different scales, CNN-based methods usually generate prior anchors with multiscales on different feature maps. Therefore, it is necessary to determine which feature maps of the backbone network are required for prediction as well as the scales of prior anchors used for detection after the feature extraction of the backbone network is completed. Previous studies used different methods such as K-means clustering algorithm or fine-tuning to determine the scales of the prior anchors aiming to cover the target scale distribution in the dataset as much as possible [36], [37], [40], [51]. However, different heuristic settings for the same dataset SSDD [36] shown in Table II are used in previous works without a good explanation for the scale selection of output backbone feature maps, which may not guarantee the optimal performance. Feature maps with inappropriate scales may not only fail to boost the detection performance, but also bring additional computational burden for subsequent processing. At present, the method to find the optimal feature map stage setting is to test different combinations one by one without a fast and reasonable guidance scheme. How to quickly find the optimal scale configuration remains a challenge to be solved.

In this article, we propose a method and a new indicator called ideal target scale (ITS) to try to guide the calibration of the output feature map scale setting. The ITS for feature map of Stage_n is defined as

$$\mathrm{ITS}_n = m \times 2^n \tag{1}$$

where m is the size of the convolution kernel of the subnetworks. For the one-stage detection model proposed in this article, a 3×3 convolution kernel is used to detect objects at each position on the feature maps output from the backbone network, therefore $m = 3$. The physical meaning of ITS_n is the side length of the original image square area corresponding to the small piece of feature map covered by the convolution kernel when the subnetworks are detecting objects on the feature map of Stage_n at a specific position.

To analyze the impact of feature maps at different stages on the prediction of a specific target, we define the target scale s_i^B of target i in the original image as

$$s_i^B = \sqrt{h_i^B \times w_i^B} \tag{2}$$

where h_i^B and w_i^B are the height and width of the BBox of target i, respectively.

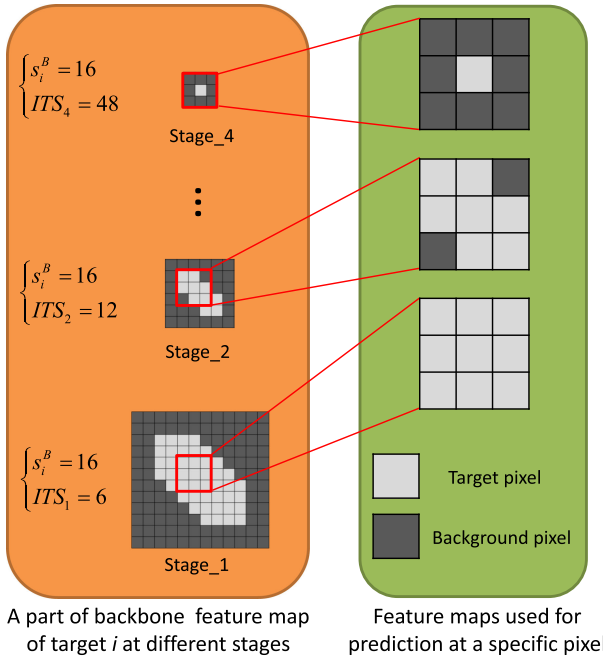


Fig. 2. Feature map used for the center pixel prediction at different stages with $m = 3$. Note that the target with $s_i^B = 16$ is taken as an example here because $s_i^B = 16$ a typical value in our dataset.

In the two-stage models, the RoI pooling layer is used to align the feature map region input into the prediction network with the target's real region [31]. However, in the one-stage model, the region of the feature map input into the prediction network at each anchor point is only determined by m and cannot be dynamically adjusted according to the region of the target, which is interpreted in Fig. 2. Taking a target with $h_i^B = w_i^B = 16$ as an example, as can be seen from Fig. 2, on the one hand, the small piece of feature map used for prediction at a specific position will contain too many background pixels when the ideal target scale ITS_n is too large than the scale of target s_i^B such as Stage_4 in Fig. 2, which will introduce too much noise for prediction, and the low resolution of the feature map may also lead to insufficient positive samples during the training stage [51]. On the other hand, although low-level feature maps can provide dense anchors on the original image, too much reliance on high-resolution feature maps may bring unnecessary computational burden, and the convolution kernel will focus on partial information if the scale the target s_i^B is too large than ITS_n such as Stage_1 in Fig. 2, which is not conducive to the detection of large targets. Therefore, the backbone feature maps output for prediction should contain a feature map whose ITS_n is closest to the target s_i^B in order to ensure the detection performance of the model, such as Stage_2 in Fig. 2. When considering targets with different s_i^B, this means that the range of ITS_n of the backbone feature maps should cover the distribution of the s_i^B in the dataset.

In addition, the feature maps should also be used as few as possible in order to balance the accuracy and speed of the model. Following the above design principles, we propose a feature map scale calibration method to align the scale distribution of

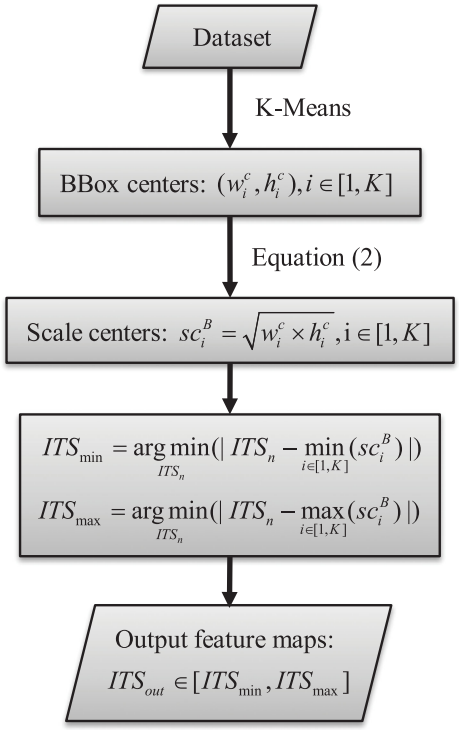


Fig. 3. Flowchart of feature scale calibration method.

the output feature maps of the backbone network with the scale distribution of the targets. The flowchart of our feature scale calibration method is summarized in Fig. 3. We first use the K-means algorithm to cluster the scales of different targets in the dataset according to the height and width of the BBoxes to obtain K BBox centers. Next, K scale centers are obtained by (2). Then, we find two feature maps whose ITS are closest to the smallest scale center and the largest scale center, respectively, and mark their ITS as $\mathrm{ITS}_{\min}$ and $\mathrm{ITS}_{\max}$. Finally, feature maps with ITS $\in [\mathrm{ITS}_{\min}, \mathrm{ITS}_{\max}]$ are selected as output feature maps to ensure that the targets' scale distribution is well covered without introducing too many redundant feature maps.

C. Feature Fusion Based on TA-FPN

In feature extraction networks, feature maps in deep layers often have strong semantic features and low resolution, which is very helpful for target classification but restrict their localization capacity in detection tasks. Feature maps in shallow layers have higher resolution but their low-level features harm their representation capacity for object recognition [58]. In order to obtain high-resolution feature maps with strong semantic features, FPN [58] is proposed, which uses a top-down pathway to enhance the semantic information of the low-level high-resolution feature maps. The introduction of FPN greatly improved the model's adaptability to multiscale targets. Since then, researchers have used various techniques such as attention mechanisms to enhance the feature fusion capabilities of FPN [38], [40], [49].

However, these studies mainly focus on the global optimization of features without taking into account the different learning goals of classification branch and localization branch. In general, classification tasks require features to maintain translation invariance and scale invariance, which means that the features used for classification should remain as constant as possible when the position and scale of the same target change. However, the localization task requires features to maintain translation and scale variability, which means that changes in target position and size should be expressed in the features. This feature space deviation of two different tasks seriously hurts the training process [54].

Driven by this problem, researchers have begun to make some explorations. Recent research shows that this internal contradiction can be greatly alleviated by decoupling the gradient flows of classification task and the localization task in the spatial dimension of the feature maps [54], which successfully improves the model performance. However, this spatial decoupling depends on the RPN integrated in the two-stage models, which cannot be applied in the one-stage models. Therefore, we designed a new feature optimization module based on task-wise attention mechanism for one-stage models to decouple the gradient flows of classification task and the localization task in the channel dimension of the feature maps and a new FPN named TA-FPN is proposed, which is shown in the middle of Fig. 1. Each fused feature map obtained by merging the lateral connection and high-level features will be recalibrated separately in the task-wise SE module, which is illustrated in Fig. 4. The encoder of the task-wise SE module consists of a fully connected layer and a ReLU activation function. The decoder consists of a fully connected layer and a sigmoid activation function. The global information of each channel in the fused feature map is first input to an encoder, which can be expressed as

$$\mathbf{e}_n = \mathrm{ReLU}\left(\mathbf{W}_n \mathbf{z}_n\right) \tag{3}$$

where $\mathbf{W}_n \in \mathbb{R}^{\frac{C}{r_{\mathrm{SE}}} \times C}$ is the weight of encoder at Stage_n and C is the number of channel of input feature map. r_{SE} is the reduction factor and it is used to reduce the model complexity. $\mathbf{z}_n \in \mathbb{R}^{C\times 1\times 1}$ is the output of global pooling layer at Stage_n. $\mathbf{e}_n \in \mathbb{R}^{\frac{C}{r_{\mathrm{SE}}} \times 1\times 1}$ is the output of encoder at Stage_n and is then fed into decoders, which can be expressed as

$$\mathbf{a}_{nk} = \mathrm{Sigmoid}(\mathbf{M}_{nk}\mathbf{e}_n + \mathbf{b}_{nk}) \tag{4}$$

where $\mathbf{a}_{nk} \in \mathbb{R}^{C\times 1\times 1}$ is the output vector of the kth decoder at Stage_n and $k \in \{1, 2\}$. $\mathbf{M}_{nk} \in C \times \mathbb{R}^{\frac{C}{r_{\mathrm{SE}}}}$ is the weight of kth decoder at Stage_n. $\mathbf{b}_{nk} \in \mathbb{R}^{C\times 1\times 1}$ is the bias of kth decoder at Stage_n and it is initialized to 2.19 to avoid the output of the decoders being too small at the beginning of training. Each channel of the fused feature map will be enhanced or suppressed by multiplying with the corresponding element in the output vector of a decoder. This structure allows the model to learn how to adaptively focus on features in different channels according to different learning tasks, which is why we named it TA-FPN. It is worth noting that the encoder in a task-wise SE module must be shared by the two branches, because task-wise SE module with two independent encoders did not provide any performance improvement according to our experiments. Detailed analysis of TA-FPN is explained in Section III-D.

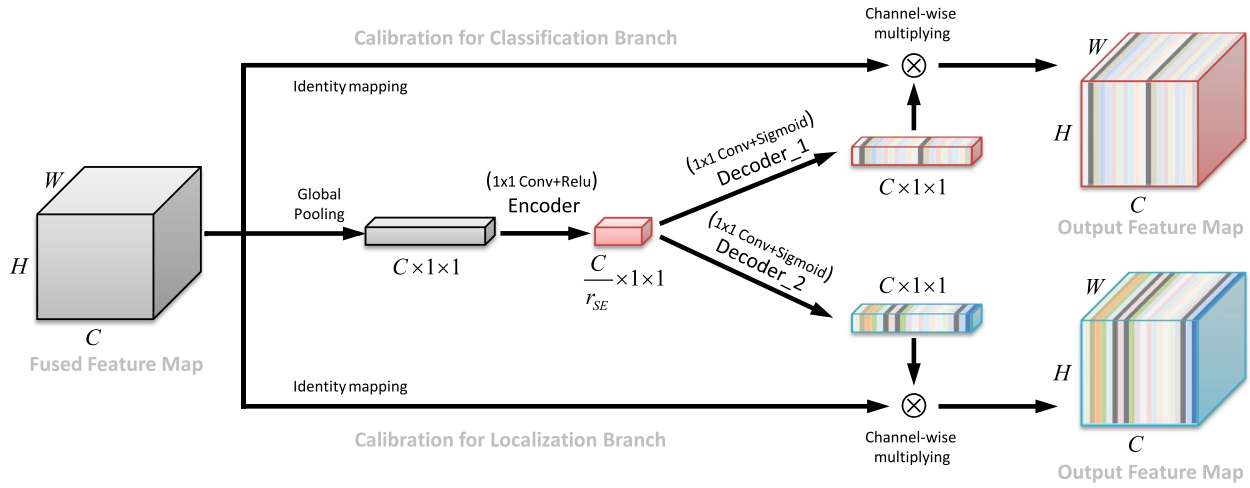


Fig. 4. Task-wise SE module. C is the number of channels of feature map and r_{SE} is the reduction factor. The different channel colors of the output feature maps of the two branches indicate that they are calibrated by different calibration vectors.

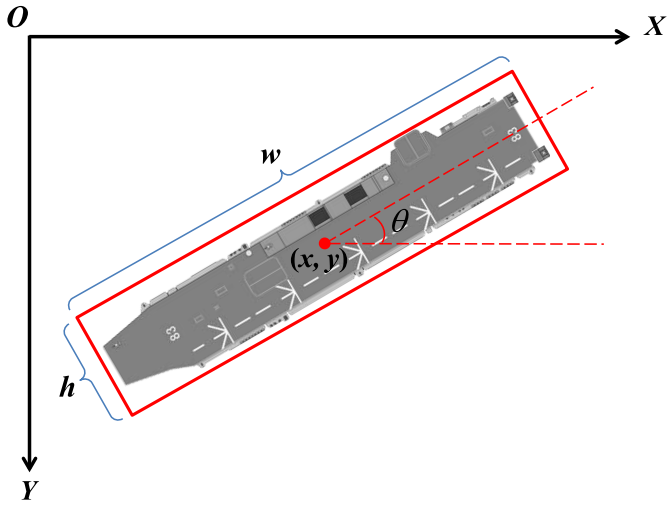


Fig. 5. Representation of RBox.

D. Prior Anchor Generation With RBox

After obtaining the multiscale feature maps, the prior anchors are generated on the image at the anchor points corresponding to each pixel of feature maps. The interval of anchor points is determined by the spatial relationship between the input image and the feature maps, which is well illustrated in previous studies [51]. Most previous researches applied anchors based on BBox to detect ship targets. However, recent studies have shown that RBox can depict the orientation angle of targets much better. Furthermore, the RBox can also effectively improve the detection performance, benefiting from its natural adaptation to densely arranged ship targets. Consequently, we use RBox, which is shown in Fig. 5 for ship detection. An RBox can be defined by the following five parameters: the coordinates of its central point (x, y), the height h and width w, and the rotation angle θ. The rotation angle θ of the RBox is defined as the angle between the long axis of the RBox and the horizontal axis, which is limited from -90° to 90°.

Each anchor generated on the images will be labeled as positive sample or negative sample according to their IoU with the ground truth, which are then used in the training process. For RBox-based anchors, skew IoU [50], [59] is adopted to compute the IoU between two RBoxes.

E. AIT Training Method

During the training stage, the loss function can be represented as the sum of classification loss and localization loss

$$\begin{aligned} L_{\text{total}} &= L_{\text{class}} + L_{\text{localization}} \\ &= \frac{1}{N_p} L_{\text{conf}}(\mathbf{c}) + \frac{1}{N_p} L_{\text{reg}}(\mathbf{R}, \mathbf{G}) \end{aligned} \tag{5}$$

where N_p is the number of positive samples, $L_{\text{conf}}(\mathbf{c})$ is the confidence loss generated by the classification branch, and $L_{\text{reg}}(\mathbf{R}, \mathbf{G})$ is the regression loss generated by the localization branch. Let N denotes the total number of anchor boxes, then $\mathbf{c} = [c_1, c_2, \ldots c_N]^T$ is the confidence vector predicted by model. $\mathbf{R} = [\mathbf{r}_1; \mathbf{r}_2; \ldots; \mathbf{r}_{N_p}]$ denotes the predicted localization matrix for positive samples, where $\mathbf{r}_i = [r_i^x, r_i^y, r_i^h, r_i^w, r_i^\theta]^T$ records the predicted regression results of the ith anchor box. $\mathbf{G} = [\mathbf{g}_1; \mathbf{g}_2; \ldots; \mathbf{g}_{N_p}]$ represent ground truth localization matrix for positive samples, where $\mathbf{g}_i = [g_i^x, g_i^y, g_i^h, g_i^w, g_i^\theta]^T$ denotes the ground truth location of the ith anchor box. We use the Smooth-L1 function to calculate the regression loss L_{reg}, which is defined as

$$L_{\text{reg}} = \sum_{i \in \text{Pos}} \sum_{m \in \{x,y,h,w,\theta\}} \text{smooth}_{L1}\left(r_i^m - g_i^m\right) \tag{6}$$

where

$$\text{smooth}_{L1}(s) = \begin{cases} 0.5s^2, & \text{if } |s| < 1 \\ |s| - 0.5, & \text{else} \end{cases}. \tag{7}$$

Focal Loss [57] was introduced as the confidence loss for classification branch to alleviate the imbalance of positive and

negative samples in our one-stage model, which can be represented as

$$\begin{aligned} FL &= -\sum_{i\in \text{Pos}} (1-c_i)^{\gamma} \log(c_i) - \sum_{i\in \text{Neg}} c_i^{\gamma} \log(1-c_i) \\ &= -\sum_{i} (1-p_i^t)^{\gamma} \log(p_i^t) \end{aligned} \tag{8}$$

where $\gamma = 2$ and $p_i^t \in [0, 1]$ and can be expressed as

$$p_i^t = \begin{cases} c_i & \text{if} \quad i \in \text{Pos} \\ 1 - c_i & \text{if} \quad i \in \text{Neg} \end{cases}. \tag{9}$$

In the common practice [51], anchor that has the largest IoU with the ith ground truth target will be matched with the ith ground truth target and labeled as positive sample. In addition, anchors that have an IoU larger than T_{IoU} with any ground truth target will be matched with that ground truth target and labeled as positive samples. Except for the positive samples, the rest of anchor box are labeled as the negative samples or ignored. Previous studies [49], [51] used a fixed T_{IoU} to distinguish positive and negative samples when training the RBox-based model because it is widely used in BBox-based model [31], [37], [57]. However, this simple copy will cause an extremely large variance of the number of positive samples generated by different targets in the RBox-based model, which has not been mentioned in previous studies. The high variance of the number of positive samples can be regarded as a kind of intraclass imbalance problem, which will make the model pay too much attention to the targets with more positive samples and ignore the targets with fewer positive samples.

An intuitive idea is to directly lower the fixed IoU threshold to obtain more positive samples, but this will introduce too many poor matched positive samples to those targets that already have enough well-matched positive samples, which cannot effectively reduce the variance of number of the positive samples.

Based on the above analysis, a new training method based on AIT is proposed to reduce the variance of the number of positive samples. For a specific target i, we first calculate the IoU between all anchors and the ground truth RBox of target i. Then, we select the first N_a anchors with the largest IoU as the candidate anchors. Finally, let V_j be the IoU of the jth anchor in the candidate anchors and the T_{IoU} of target i will be calculated as follows:

$$T_{\text{IoU}}^i = \frac{1}{N_a} \sum_{j=1}^{N_a} V_j. \tag{10}$$

The anchors with IoU $\geq T_{\text{IoU}}^i$ will be regarded as positive samples and the anchors with IoU $< T_{\text{IoU}}^i - 0.1$ will be regarded as negative samples. The rest of the anchors will be ignored. T_{IoU}^i will become larger when there are a large number of anchors that match the target well, thereby avoiding introducing too many positive samples that are poorly matched to the target. T_{IoU}^i will decrease when most of the anchors are not well-matched with the target, which ensures that this target has enough positive samples.

III. Experiments and Discussion

In this section, experiments are performed to evaluate the performance of the proposed methods. The experimental datasets used in this article and the corresponding evaluation methods will be introduced first. Then, detailed ablation experiments will be performed to prove the effectiveness of each improvement. Finally, the comparison with other state-of-the-art methods implies the significance of the proposed method.

A. Experimental Data

Two different datasets are collected to evaluate the proposed method: SSDD+ and GF3-Ship. SSDD+ contains different scenarios and a large number of targets with different scales, which will be used to analyze the performance of the proposed method in detail. GF3-Ship consists of 882 high-resolution SAR image chips produced by the Chinese GF-3 satellite, which is smaller than SSDD+ and will be used to evaluate the performance of the proposed methods in high-resolution images and explore the potential of the GF-3 satellite for ship monitoring.

SSDD+ can only be obtained by applying Li *et al.* [36]. The download link of the SAR image used to make GF3-Ship dataset was released in [60]. Details of these datasets are described in the following.

1) SSDD+: SSDD [36] is the first public dataset for ship target detection in SAR images, which has been widely used to compare the performance of different detectors. SAR images of resolution from 1 to 15 m are collected from Radarsat-2, Sentinel-1, and TerraSAR-X to form the SSDD dataset, which contains multiscale ships labeled with BBox in various environments, including different scenes, sensor types. The polarization modes of these samples include HH, HV, VV, and VH.

On the basis of SSDD, Li *et al.* [36] introduced a ship dataset labeled with RBox in order to facilitate the research of RBox-based detection methods in SAR images, which is called SSDD+ and has the same samples as SSDD. It should be noted that some targets in the original SSDD+ have very poor label quality and low-quality labels will seriously affect the results of the experiments, so we corrected the labels that have large errors in the original SSDD+ before our experiments. There are totally 1160 images in SSDD+, which are randomly divided into training set and test set with the proportion of 8:2 for the training and testing of the proposed method.

2) GF3-Ship: GF3-Ship is a small dataset for SAR ship detection research, which is composed of images under different levels of sea condition. The sea condition information in the GF3-Ship dataset makes it possible to study the robustness of the model under different sea conditions. GF3-Ship is made from 31 large-scale SAR images published by Sun *et al.* [60]. These large-scale SAR images generated by the GF3 satellite contain a large number of ship targets and each image has a size around 3000×3000. The horizontal BBox of each ship target is given by experts after examining the corresponding SAR images. Based on this dataset, we further mark out the corresponding RBox according to the BBox of each target and crop out 882 slices with size of 800×800 from 30 images to form the GF3-Ship dataset. The remaining one large-scale image is saved for large scene validation. The detailed parameters of GF3-Ship dataset are shown in Table III. Considering that many slices are very similar in GF3-Ship dataset, 60% of the slices are

TABLE III
DETAILED PARAMETERS OF GF3-SHIP DATASET

Parameter	Value
Number of Images	882
Image Size	800 × 800
Resolution	1m, 3m
Mode	Spotlight, UFS
Polarization Mode	VV
Sea Condition	Level 0 - Level 4

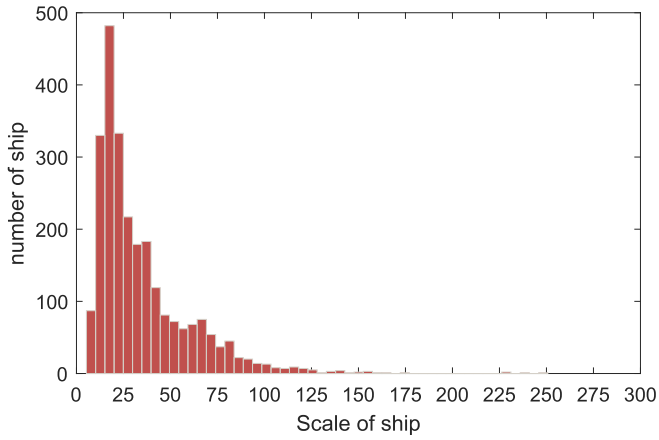


Fig. 6. Statistical histogram of targets' scales s_i^B in SSDD+.

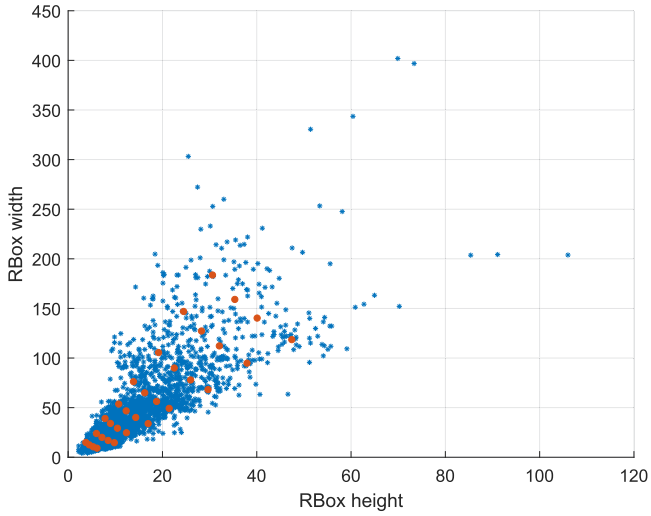


Fig. 7. Scatter diagram of the size of targets in SSDD+. The blue dot represents the size of a target's RBox, and the red dot represents the size of a prior anchor.

TABLE IV
ANCHOR SETTING FOR DIFFERENT FEATURE MAPS BASED ON RBOX

Feature Map Stage	Anchor Parameter		
	RBox Scale	Aspect Ratio	Angle
Stage_2	[7.5, 12]	[1.5, 2, 2.8, 4]	$[0°, \pm30°, \pm60°, \pm90°]$
Stage_3	[17.5, 24]	[2, 2.8, 3.8, 5]	$[0°, \pm30°, \pm60°, \pm90°]$
Stage_4	[32.5, 45]	[2.3, 3, 4, 5.5]	$[0°, \pm30°, \pm60°, \pm90°]$
Stage_5	[60, 75]	[2.5, 3.5, 4.5, 6]	$[0°, \pm30°, \pm60°, \pm90°]$

randomly sampled and used for training and the rest 40% of the slices are used for testing.

B. Implement Details

1) Data Preprocessing: To enhance the robustness of the model, we use data augmentation in data preprocessing. During the training stage, each image will be first padded into a size of 800 × 800 with zeros if its size is smaller than 800 × 800. Then, we randomly select a ship target in that image and randomly crop a 320 × 320 slices under the premise of including that target in this slice. Finally, the slice is randomly flipped horizontally with a probability of 0.5. During the testing stage, the test image is inferred at its original size.

2) Hyperparameter Setting: In standard RetinaNet, feature maps on stages 3–7 with $\text{ITS}_n \in \{24, 48, 96\,192\,384\}$ are used for detection. However, through the statistical results of the targets scale distribution of SSDD+ shown in Fig. 6, we find that that nearly 47% of ship targets in SSDD+ have scales smaller than ITS_3, whereas less than 4% of ship targets in SSDD+ have scales larger than ITS_5, which leads to serious scale misalignment problem. Therefore, it is necessary to recalibrate the scales of the output feature maps of the feature extraction network according to the distribution of the target scale in the dataset.

Based on the proposed method, we recalibrate the scales of the output feature maps from $\text{ITS}_n \in \{24, 48, 96\,192\,384\}$ to $\text{ITS}_n \in \{12, 24, 48, 96\}$ according to the distribution of ship scale in SSDD+, which means that the feature map on Stage_2, Stage_3, Stage_4, and Stage_5 of ResNet50 is used for detection. The calibration process of GF3-Ship is similar to SSDD+. The detailed calibration process on SSDD+ is explained in Section III-D.

In order to fit the scale distribution of ship targets as close as possible without introducing too much computation, we selected eight scales for the RBox-based anchor, and each scale contained seven uniformly distributed orientation angles and four aspect ratios. The size distribution of anchors and targets in SSDD+ is shown in Fig. 7. After the crossover experiments, we evenly distribute anchors of eight different scales to each output feature map, which is shown in Table IV. Note that the RBox scale of anchor in Table IV is defined as $s_i^R = \sqrt{h_i^R \times w_i^R}$, where h_i^R and w_i^R are the height and width of the RBox of anchor i, respectively.

The settings of the other hyperparameters mentioned above are shown in Table V as a reference for SSDD+ dataset, which led to best results in our crossover experiments. For the hyperparameters involved in the basic framework such as T_{conf} and T_{nms}, we use the default configuration given in the previous literature. These configurations have been proven to provide a stable performance in our cross experiments. The new hyperparameters introduced in our proposed methods mainly include the number of BBox centers K in the scale calibration method, the number of candidate anchor boxes N_a in the AIT training method and the scaling factor r_{SE} of the task-wise SE module. These newly introduced hyperparameters have undergone multiple crossover

TABLE V
SETTING OF HYPERPARAMETERS

Parameter	Recommended Value Range	Our Setting
T_{conf}	[0.02-0.1]	0.05
T_{nms}	[0.1-0.5]	0.2
K	[6-12]	6
N_a	[30-80]	40
r_{SE}	[4-16]	8

TABLE VI
ABLATION EXPERIMENTS ON SSDD+

Scale Calibration	TA-FPN	AIT	AP_{50} (%)	BEP_{50} (%)	Inferance Time (ms/Image)
×	×	×	87.78	88.62	46.47
√	×	×	92.39	92.54	57.94
×	√	×	88.64	89.04	53.11
×	×	√	91.96	91.29	46.56
√	√	×	93.41	93.23	62.87
√	×	√	94.27	93.75	59.16
×	√	√	92.49	92.37	53.26
√	√	√	**94.66**	**94.03**	62.77

experiments to obtain a reasonable range of values for each of them. Hyperparameters beyond the recommended value range in Table V may cause significant performance degradation of the model or unnecessary computational complexity. Detailed analysis of the effects of these hyperparameters is shown in Section III-D.

3) Optimizer Setting: All the models are trained with a stochastic gradient descent algorithm over an Intel E5-2680 V3 processor and an Nvidia GTX1080Ti GPU. Focal Loss function [57] and Smooth-L1 Loss function are used to calculate classification loss and regression loss, respectively. The mini-batch size is 4 in one iteration. The models are trained for 140k iterations with an initial learning rate of 0.0005, which is then divided by 10 at 80k and 120k iteration, respectively.

C. Evaluation Criteria

In order to compare different models properly, we choose average precision (AP) and break-even point (BEP) to quantify the performance of the models.

1) Average Precision: AP [38], [51] is the standard metric for target detection algorithms, which comprehensively considers the precision rate P_d and recall rate R_d of the model at different confidence levels and can be expressed as

$$\mathrm{AP}_d = \int_0^1 P_d(R_d)\mathrm{d}\mathrm{R}_d \tag{11}$$

where d is the IoU threshold used to distinguish whether a detection result is true positive or a false positive. If the IoU between a predicted RBox and a ground truth RBox is higher than $d\%$, the predicted RBox is a true positive, otherwise it is a false positive. The value of AP_d can ranges from 0 to 1. The AP_d of an ideal detector will be equal to 1. Following the common practice, $d = 30$ and $d = 50$ are used in the evaluation.

2) Break-Even Point: BEP [51] refers to the point where $P_d = R_d$, and the corresponding value of the recall (precision) rate is called as the BEP value and utilized as a metric to evaluate the detectors at a single confidence level. Higher BEP corresponds to better detection performance.

D. Evaluation of the Proposed Method

In this section, the contribution of each improvement will be quantitatively evaluated on SSDD+ and GF3-Ship through a series of ablation experiments to demonstrate the effectiveness of each modification. Discussions are then conducted with the aim of analyzing the impact of each improvement on the network model in detail. Finally, the proposed method is compared with other methods on two different datasets to show the advantages of the proposed method in ship detection.

1) Ablation Experiments Results: Several improvements have been made in the aspects of model design and training process, so it is necessary to study the actual effect of each improvement and their impact on each other. Experiments using different combinations of these improvements were performed and the experimental results on SSDD+ are shown in Table VI. The experimental results can be summarized as the following aspects.

First, it can be seen from Table VI that applying each improvement individually can effectively improve the performance of the benchmark model (Standard RetinaNet), which proves the significance of each improvement.

Second, the comparison of experiments containing two improvements with experiments containing only one improvement shows that the performance gain of the model is significantly cumulative with multiple improvements, indicating small overlap between different improvements.

Third, scale calibration provides the largest performance gain (about 5% of AP_{50}) in experiments that contain only a single improvement, which illustrate the importance of scale calibration of the backbone feature map. TA-FPN only contributed an AP_{50} improvement around 0.4% to the model with AIT, which indicates that a better structure may be needed to decouple the gradient flow of different tasks in the channel dimension.

Finally, it can be found from the last column of Table VI that the calibration of the feature map has a greater impact on the detection speed of the model. This is because that the feature map scale calibration introduces a higher resolution feature map for detection, which increases a lot of calculations. The introduction of TA-FPN will also reduce the detection speed because different subnetworks no longer share the same fusion feature. AIT has no effect on the detection speed because it does not change the calculation process of the model.

It is noted from the results that the highest AP and BEP are achieved by R-RetinaNet with all three improvements.

2) Effect of Scale Calibration: According to the proposed scale calibration method, we first cluster the BBoxes of the targets in SSDD+ to obtain six BBox centers: (189.96 × 60.91, 125.51 × 63.39, 95.39 × 44.30, 55.52 × 28.17, 44.63 × 24.86, and 19.49 × 16.17). Then, we can get the six scale

TABLE VII
CALIBRATION RESULTS UNDER DIFFERENT K

K	Calibration Result of Feature Map Scale
2	{3,4,5}
4	{3,4,5}
6	{2,3,4,5}
8	{2,3,4,5}
10	{2,3,4,5}
12	{2,3,4,5}
14	{2,3,4,5,6}
16	{2,3,4,5,6}

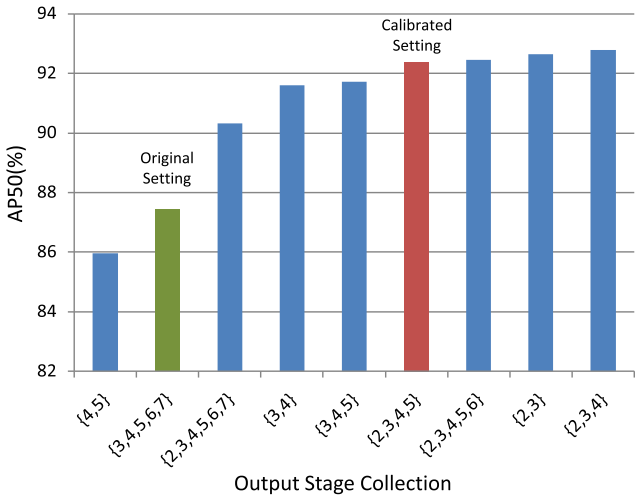


Fig. 8. Comparison of different output stage collections.

centers as (106.71, 89.20, 65.00, 39.55, 33.31, 17.76), and the $\mathrm{ITS}_{\min}$ and $\mathrm{ITS}_{\max}$ are known as $\mathrm{ITS}_{\min} = \mathrm{ITS}_2 = 12$ and $\mathrm{ITS}_{\max} = \mathrm{ITS}_5 = 96$. Finally, the scales of the output feature maps are recalibrated from $\mathrm{ITS}_n \in \{24, 48, 96\ 192\ 384\}$ (used in standard RetinaNet) to $\mathrm{ITS}_n \in \{12, 24, 48, 96\}$.

The calibration results under different K are shown in Table VII. It can be found that the calibration result of the model is quite insensitive to the variations of K from 6 to 12. It is worth noting that too small K cannot generate enough cluster centers to represent the scale distribution of the targets' BBox, whereas too large K will cause redundant calculations when performing the clustering. Consequently, $K = 6$ was selected for subsequent experiments.

To demonstrate the validity of scale recalibration, models using different scale (stage) settings for the output feature maps were tested. All controlled trials use the same anchor scale setting and allocation principles mentioned in Section III-B. All eight anchor sizes will be evenly distributed to the feature maps in each configuration. The experimental results are sorted by their AP value and shown in Fig. 8. It can be seen that the performance of the calibrated setting is far better than that of the original setting and slightly inferior to that of the optimal

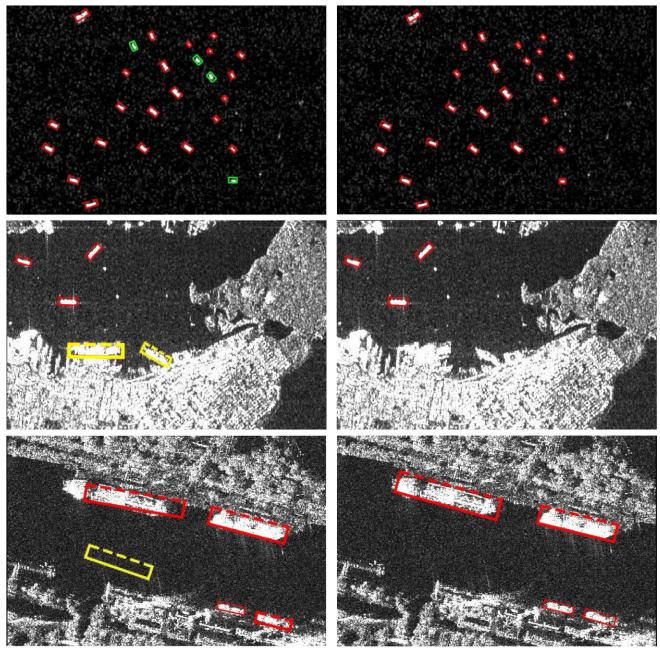


Fig. 9. Comparison of detection results before and after scale calibration. Note that the red boxes represent true positive targets, the yellow boxes represent false positive targets, and green boxes represent missed targets. Column (a) gives the results before scale calibration. Column (b) gives the results after scale calibration.

configuration {2,3,4}, which proves that feature map scale alignment can effectively improve model performance while avoiding exhaustive configuration search. Note that the calibrated setting has the highest test speed in the settings with $\mathrm{AP}_{50} > 92\%$. Fig. 9 shows some of the detection results before and after scale calibration on SSDD+. It can be seen from Fig. 9 that the scale calibration effectively improves the recall rate of the targets that have smaller size compared with other targets in SSDD+.

On the one hand, the proposed method can effectively guide the scale correction of the output feature map to improve the performance of the model, which proves the necessity to align the scale distribution of the output backbone feature maps with the scale distribution of the targets. On the other hand, it is worth noting that the proposed scale calibration method cannot guarantee the optimal scale setting of the feature map, which may because that the definition of the target scale in Section II-B is not suitable for the representation of the true scale of the ship target in the horizontal or vertical direction. Therefore, it is necessary to design a better target scale representation in order to increase the robustness of the scale calibration method in the future work.

3) Effect of TA-FPN: The training of the detection model based on CNN is a typical multitask learning process, which includes target classification learning and coordinate regression learning. The convolution calculation process before the subnetworks shown in Fig. 1 is fully shared by the classification subnetwork and the regression subnetwork in the previous methods, which greatly improves the efficiency of the model. However, classification tasks and localization tasks have completely

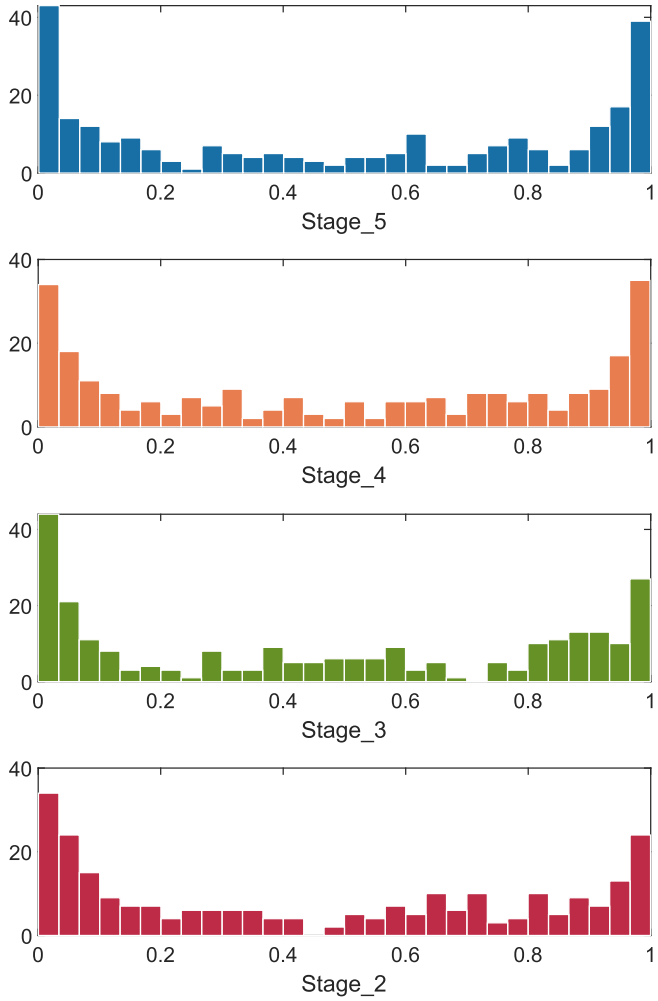


Fig. 10. Statistical histogram of element values in $\mathbf{d}_n$ at different stages.

opposite requirements on translation transformation and scale transformation of features.

In order to demonstrate the difference of requirement for the shared features between different learning tasks, we explored the behavior of the two decoders in each task-wise SE module of a converged model. First, the output vectors of decoder_1 and decoder_2 are averaged over 1160 images in SSDD+, respectively. Then, the difference between the average output vectors of the two decoders is calculated as follows:

$$\mathbf{d}_n = |\bar{\mathbf{a}}_{n1} - \bar{\mathbf{a}}_{n2}| = [d_{n1}, d_{n2}, \ldots, d_{nm}] \tag{12}$$

where $\bar{\mathbf{a}}_{n1}$ is the average output vector of decoder 1 at stage n and $\bar{\mathbf{a}}_{n2}$ is the average output vector of decoder 2 at stage n. $\mathbf{d}_n$ is the difference between the outputs of two decoders at stage n. $0 \leq d_{nm} \leq 1$ represents the difference between the calibration values of two decoders for the mth channel of the feature map at stage n.

The statistical histograms of the element values in $\mathbf{d}_n$ are shown in Fig. 10. It can be seen that almost half of the differences are greater than 0.5, which means that many channels of the feature map are suppressed in one branch, whereas enhanced in the other branch, indicating that the feature preference of classification branch and localization branch is significantly different. The different behaviors of two decoders in each task-wise SE module prove the different requirements for features and the necessity of decoupling in channel dimension. We also tested different calibration strategies on SSDD+ to prove the advantages of the proposed method, including global optimization, classification optimization only and localization optimization only. The experimental results are shown in Table VIII. It can be seen that the proposed method is superior to other optimization strategies. Besides, it can be found that the model that uses different calibration strategies in the two branches (classification optimization only, localization optimization only or TA-FPN) is always better than the model that uses the same calibration strategy in the two branches (original FPN or global calibration), which indicates the necessity of decoupling different tasks in the channel dimension.

TABLE VIII
PERFORMANCE OF DIFFERENT CALIBRATION STRATEGIES

Calibration Strategy	$\mathbf{AP_{50}}$ (%)	$\mathbf{BEP_{50}}$ (%)	Inferance Time (ms/Image)
Original FPN	92.39	92.54	57.94
Global Calibration	92.95	92.74	63.23
Classification Calibration	93.32	92.16	62.47
Localization Calibration	93.25	92.54	62.88
TA-FPN	**93.41**	**93.23**	62.87

TABLE IX
INFLUENCE OF r_{SE} ON MODEL PERFORMANCE

r_{SE}	2	4	8	16	32	64
$\mathbf{AP_{50}}$	93.19	92.92	93.41	93.32	92.53	92.81

The performance influence of r_{SE} on model after scale calibration process is shown in Table IX. Increasing r_{SE} can reduce the amount of parameters of the task-wise SE module, but a too large r_{SE} will result in insufficient information input to the decoder, which will damage the performance of the model. Here, we set $r_{\mathrm{SE}} = 8$, which provides the best performance in our experiments.

4) Effect of AIT: Although the introduction of RBox makes the model better adapt to the densely arranged target, it also causes a huge increase in the variance of the number of positive samples generated by different targets when using a fixed IoU threshold, which was totally ignored in previous studies. In order to analyze the difference in the distribution of positive samples generated by the BBox and RBox-based models at a fixed IoU threshold during the training stage, we randomly shift each target in SSDD+ 100 times on the image and record the number of positive samples generated by each target after every shift. Since random cropping is used in the training stage for data augmentation, and the position of the target in the input image will change after each cropping, therefore random shift is used

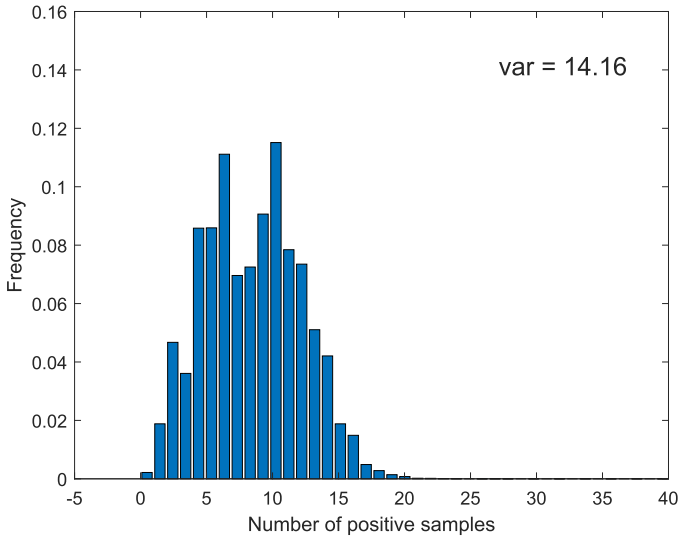


Fig. 11. Positive samples distribution of BBox-based model with fixed IoU threshold.

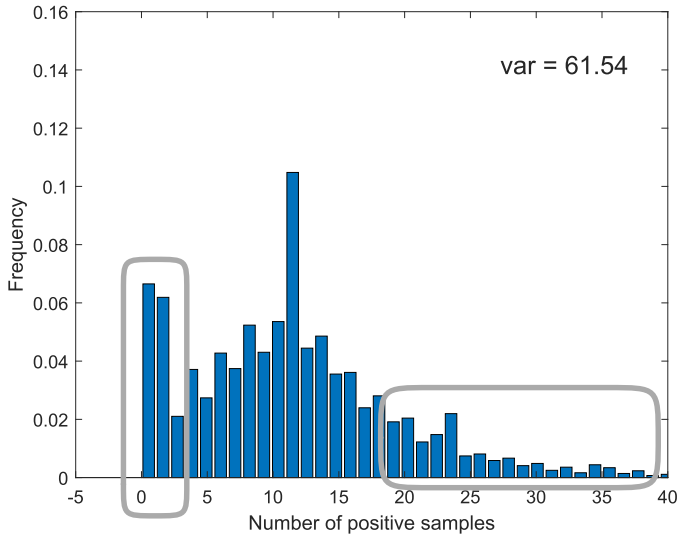


Fig. 12. Positive samples distribution of RBox-based model with fixed IoU threshold.

to simulate the effect of random cropping when recording the number of positive samples produced by each target.

Figs. 11 and 12 show the normalized frequency distribution histogram and variance of the number of positive samples generated at $T_{\mathrm{IoU}} = 0.5$ by the BBox-based and RBox-based RetinaNet, respectively. Note that both models use hyperparameter configurations that maximize their detection performance. It can be seen from Fig. 12 that there are two main reasons for the rise of variance of the number of positive samples in the RBox-based model.

The first reason is that a large number of targets generate very few positive samples (highlighted by the rectangle on the left in Fig. 12), which is mainly caused by those targets that have large aspect ratios. Another reason is that some targets generate a lot of positive samples (highlighted by the rectangle on the right in Fig. 12), which is mainly caused by targets with smaller aspect ratios or target whose center is very close to the anchor point. These targets have larger IoU with anchors that has similar size but different orientation angles, which increases the number of positive samples. Several typical targets in SSDD+ dataset are shown in Fig. 13 to illustrate the imbalance of positive samples.

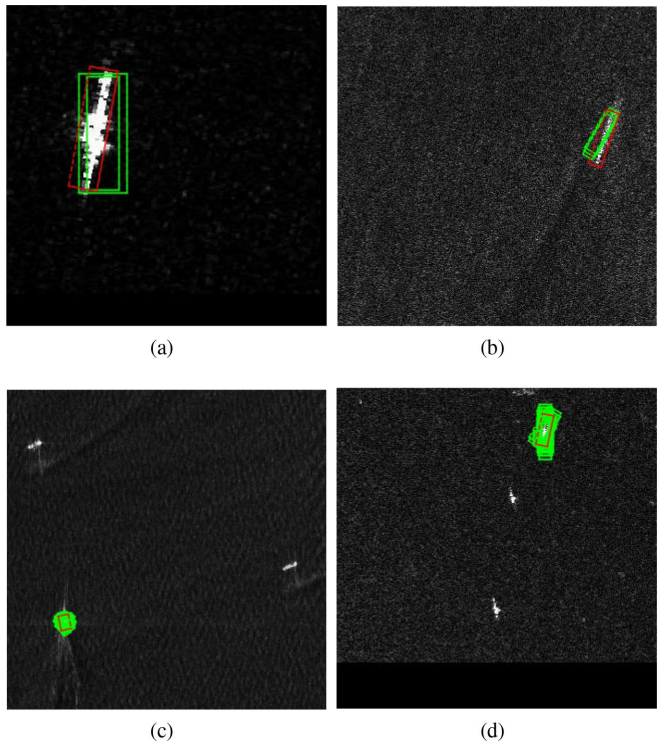


Fig. 13. Several typical targets in SSDD+ dataset. Note that the red box represents the ground truth of the target, and the green box represents the anchor box that has an IoU greater than 0.5 with the ground truth. The anchor setting follows Table IV. (a) Target with two positive samples. (b) Target with three positive samples. (c) Target with 28 positive samples. (d) Target with 35 positive samples.

The high variance of positive samples makes the model pay more attention to the targets with more positive samples in the training stage, thus ignoring the targets with fewer positive samples. In order to rebalance the contribution of different targets in the model training stage, AIT was introduced during the training stage. The positive samples distribution of RBox-based model with AIT ($N_a = 40$) is shown in Fig. 14. It can be seen that AIT successfully reduces the variance of the number of positive samples, which makes the RBox-based model can strike a balance between different targets like the BBox-based model. The balance of the contribution of different ship targets to the model during the training process can drive the model to focus on learning the common characteristics of different ship targets, instead of focusing too much on some specific ship targets that generate more positive samples.

AIT with different N_a was tested in order to study the effect of different N_a on the model performance. The experimental results are shown in Fig. 15. The results show that the value between 40 and 120 does not lead to significant changes in outcomes. However, too small N_a will cause a performance

TABLE X
PERFORMANCE OF DIFFERENT IOU THRESHOLD ON SSDD+

IoU Threshold	Mean of positive sample number	Variance of positive sample number	$AP_{50}(\%)$	$BEP_{50}(\%)$
$T_{IoU} = 0.6$	2.88	6.59	91.41	91.68
$T_{IoU} = 0.5$	11.61	61.54	93.91	93.23
$T_{IoU} = 0.4$	45.12	396.04	93.88	93.42
$T_{IoU} = 0.3$	157.91	3.52×10^3	92.31	91.86
$T_{IoU} = 0.2$	582.84	3.47×10^4	90.25	90.52
Adaptive IoU Threshold with $N_a = 40$	16.35	7.56	**94.66**	**94.03**

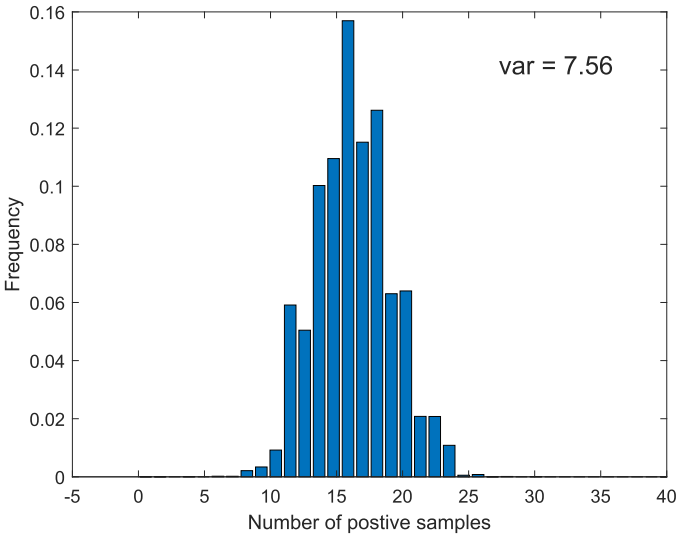


Fig. 14. Positive samples distribution of RBox-based model with AIT($N_a = 40$).

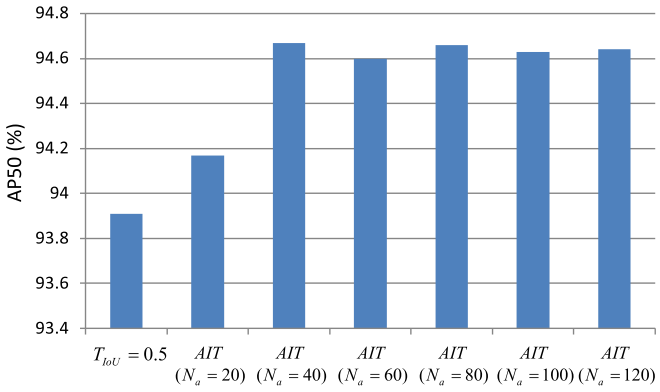


Fig. 15. Effect of different N_a on model performance.

drop, because small N_a will lead to insufficient overall number of positive samples in the training stage. The performance comparison between AIT and the fixed IoU threshold method is shown in Table X. It can be seen from Table X that for the fixed IoU threshold training method, the variance of the number of positive samples will increase rapidly when the IoU threshold is set too small, making the model unable to balance different targets. When the IoU threshold is set too large, the mean of the number of positive samples will become too small, which will cause a serious imbalance between positive and negative samples. Compared with fixed IoU threshold, AIT can effectively reduce the variance of the number of positive samples while maintaining the overall number of positive samples, which improves the detection performance.

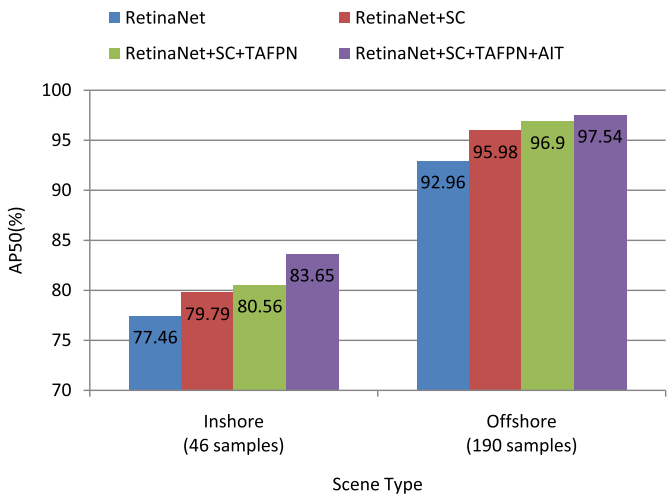


Fig. 16. Model performance on SSDD+ dataset in inshore scene and offshore scene. SC means scale calibration method.

5) Comparison of Inshore Scene and Offshore Scene: When using deep learning technology for ship target detection, the ship detection in the inshore scene is more challenging than the ship detection in the offshore scene due to the interference of the land area. On the one hand, the detector may recognize the objects on the land in the inshore scene as ships and cause false alarms. On the other hand, when a ship is close to the port or closely aligned with other ships, the detector may treat it as a part of the port or other ships, leading to missed detections. Therefore, the detection performance of inshore ships plays an important role in the evaluation of a detector.

Since SSDD+ has a larger sample size, we choose SSDD+ to evaluate the detection performance of the model in inshore scene and offshore scene. Fig. 16 shows the test results of the model in the inshore scene and the offshore scene under different model configuration.

For clarity, Fig. 16 only lists four different model configurations. On the one hand, it can be found that the detection accuracy of inshore scenes under different model configurations

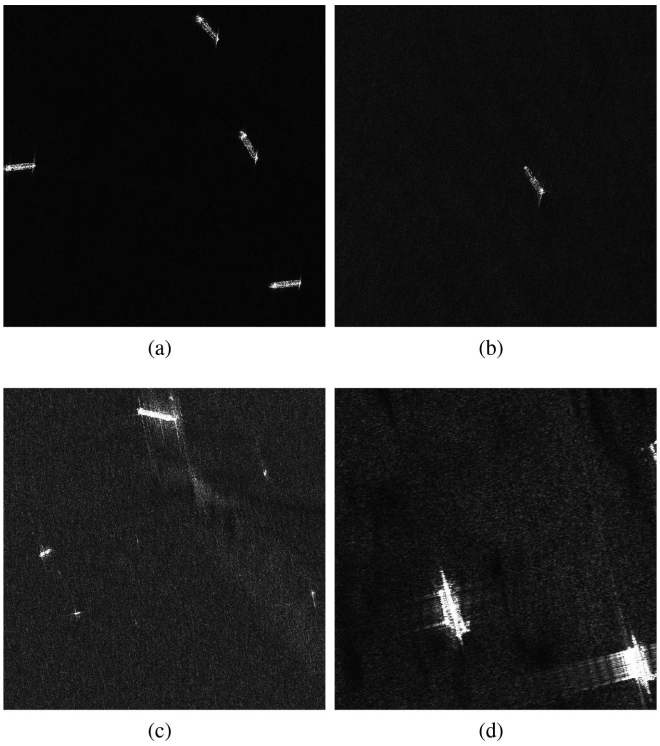


Fig. 17. Ship targets in GF3-Ship dataset under different sea conditions. (a) Ship targets under sea condition of level 1. (b) Ship targets under sea condition of level 2. (c) Ship targets under sea condition of level 3. (d) Ship targets under sea condition of level 4.

is much lower than that of offshore scenes, which proves that the detection task of inshore ships are more challenging than offshore ships. On the other hand, since the three proposed improvements were not designed for specific scenes, all three improvements can effectively boost the detection performance of inshore scene and offshore scene when compared with the benchmark model (RetinaNet), which verifies the robustness of the proposed improvements in various scenes.

In addition, compared with offshore scenes, removing AIT will result in a more significant reduction in the detection accuracy of inshore scenes. Since the model trained with AIT is more focused on the common characteristics of different ship targets, this may indicate that learning the common characteristics of different targets is more essential to ship detection in inshore scenes.

6) Model Performance Under Different Sea Conditions: The complex motion state of ship under high sea condition has caused great difficulties for SAR imaging. This makes ship targets in SAR images under high sea condition often have poor resolution and high sidelobes, which can be seen from Fig. 17. As the level of sea condition increases, the imaging quality of ship targets gradually deteriorates. Deterioration of image quality brings a huge challenge to the ship detection task. Therefore, it is necessary to explore the robustness of the detector under different sea conditions.

In the GF3-Ship dataset, there are a total of five different sea conditions, ranging from level 0 to level 4. As the sample size of level 0 is too small (eight samples), we decided to combine

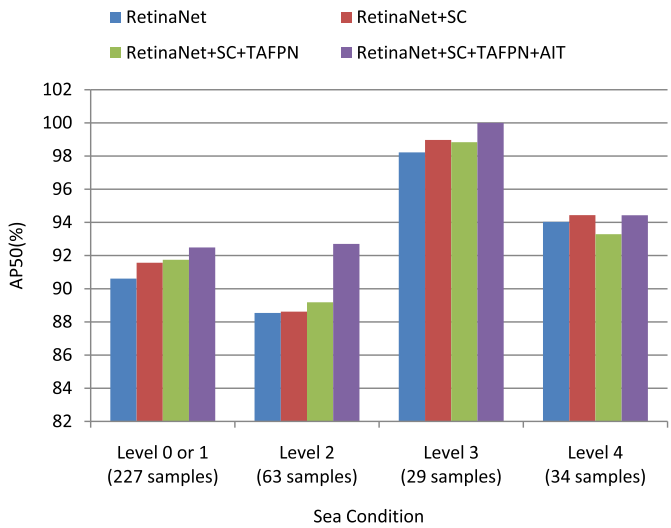


Fig. 18. Model performance on GF3-Ship dataset under different sea conditions.

level 0 and level 1 into one group during the model evaluation process. The evaluation results of the proposed methods under different sea conditions are shown in Fig. 18.

Like Fig. 16, here we only list four different model configurations in Fig. 18 for clarity. It can be seen from Fig. 18 that the overall performance of different models has not gradually deteriorated as the sea condition level increases. This may be because the number of test samples under high sea conditions (levels 3 and 4) in the GF3-Ship dataset is too small to provide reliable experimental results under high sea conditions. Therefore, in order to assess the impact of sea conditions on the performance of the model more accurately, it is necessary to establish a dataset with a large number of samples in different sea conditions. Nevertheless, from the experimental results of low-level sea condition (level 0–1 and level 2) which have sufficient samples, it can be seen that the increase in sea condition level does have a certain impact on the detection accuracy of the model. In addition, in the case of sufficient samples, the model including all three improvements achieves the best detection accuracy under different sea conditions, which verifies the effectiveness of the proposed improvements.

7) Comparison With the State-of-the-Art Methods: In this section, the proposed methods are compared with several state-of-the-art SAR ship detectors based on RBox under our implementation to demonstrate the advantages of the proposed methods. Table XI shows the performance of these methods on SSDD+ and Table XII shows the performance of these methods on GF3-Ship. Note that the inference time is measured at a resolution of 320×320. Figs. 19 and 20 show the comparison of some detection results of different one-stage models on SSDD+ and GF3-Ship, respectively. The PR curves of different models at $d = 50$ are shown in Figs. 21 and 22.

As can be seen from Tables XI and XII, due to the use of multiscale feature maps for training and prediction, SDOE and DRBox-v2 are better than DRBox-v1 and their detection accuracy on SSDD+ is similar to basic R-RetinaNet ($AP_{50} =$

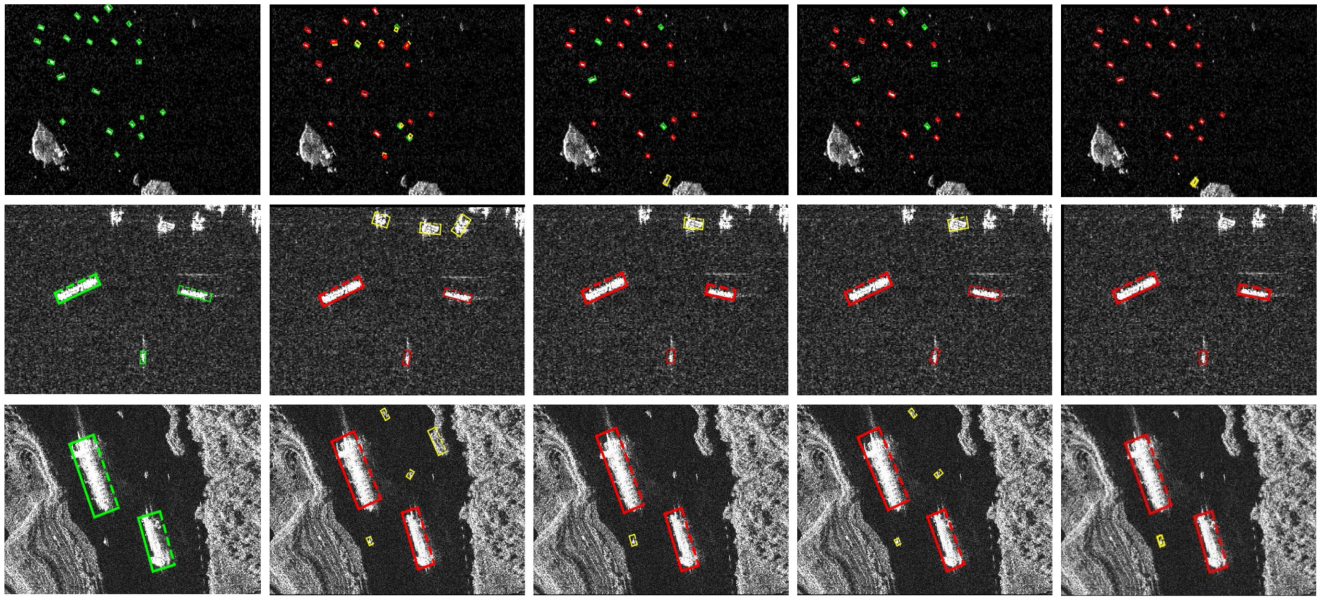


Fig. 19. Comparison of some detection results of different models on SSDD+. The meaning of the color of the RBox is the same as Fig. 9. (a) Ground-truth. (b) DRBox-v1. (c) SDOE. (d) DRBox-v2. (e) Our methods.

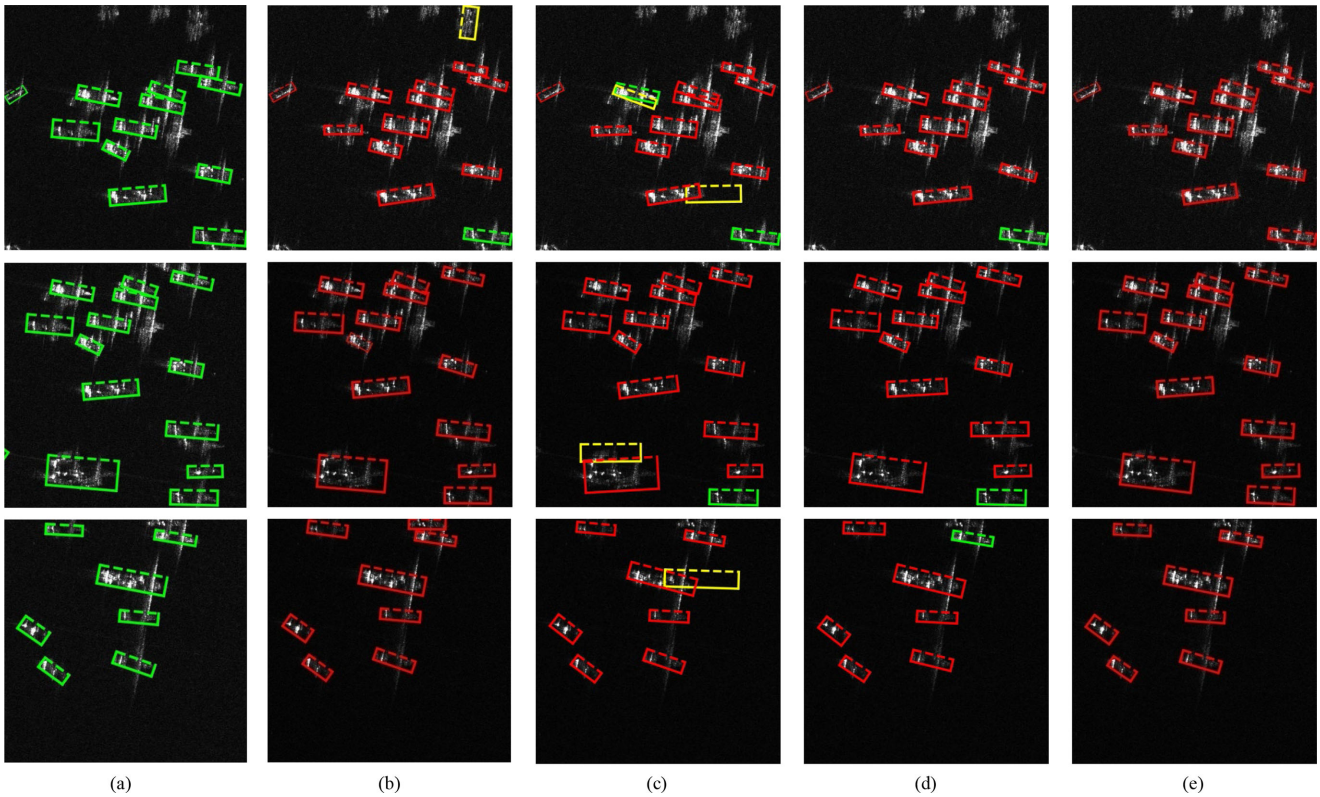


Fig. 20. Comparison of some detection results of different models on GF3-Ship. The meaning of the color of the RBox is the same as Fig. 9. (a) Ground-truth. (b) DRBox-v1. (c) SDOE. (d) DRBox-v2. (e) Our methods.

TABLE XI
COMPARISON OF DIFFERENT RBOX-BASED METHODS ON SSDD+

Method	Framework	Backnone	AP_{30} (%)	AP_{50} (%)	BEP_{50} (%)	Inference Time per Image (ms)
DRBox-v1 [51]	One-Stage	VGG16	86.41	-	-	-
DRBox-v1(Our implementation)	One-Stage	ResNet50	93.21	81.4	82.59	29.17
SDOE [49]	One-Stage	VGG16	-	84.2	-	**25**
SDOE(Our implementation)	One-Stage	ResNet50	93.88	85.17	87.23	35.33
DRBox-v2 [51]	One-Stage	VGG16	92.81	-	-	-
DRBox-v2(Our implementation)	One-Stage	ResNet50	95.24	85.74	83.75	34.21
MSR2N [52]	Two-Stage	ResNet50	93.93	90.11	90.87	103.27
R-RetinaNet	One-Stage	ResNet50	94.15	87.78	88.62	46.47
R-RetinaNet + SC + TA-FPN + AIT	One-Stage	ResNet50	**97.72**	**94.66**	**94.03**	62.77
R-RetinaNet	One-Stage	ResNet101	95.64	89.48	91.46	63.18
R-RetinaNet + SC + TA-FPN + AIT	One-Stage	ResNet101	97.36	94.45	93.88	77.05

TABLE XII
COMPARISON OF DIFFERENT RBOX-BASED METHODS ON GF3-SHIP

Method	Framework	Backnone	AP_{30} (%)	AP_{50} (%)	BEP_{50} (%)	Inference Time per Image (ms)
SDOE(Our implementation)	One-Stage	ResNet50	91.05	89.54	90.96	36.32
DR-Box-v1(Our implementation)	One-Stage	ResNet50	91.55	90.13	90.02	**28.72**
DRBox-v2(Our implementation)	One-Stage	ResNet50	92.79	91.68	91.88	34.54
MSR2N [52]	Two-Stage	ResNet50	**93.26**	92.17	**92.44**	101.28
R-RetinaNet	One-Stage	ResNet50	92.57	90.44	89.59	46.66
R-RetinaNet + SC + TA-FPN + AIT	One-Stage	ResNet50	93.21	**92.61**	91.73	63.95
R-RetinaNet	One-Stage	ResNet101	92.22	90.84	91.19	62.55
R-RetinaNet + SC + TA-FPN + AIT	One-Stage	ResNet101	92.73	92.41	92.38	76.01

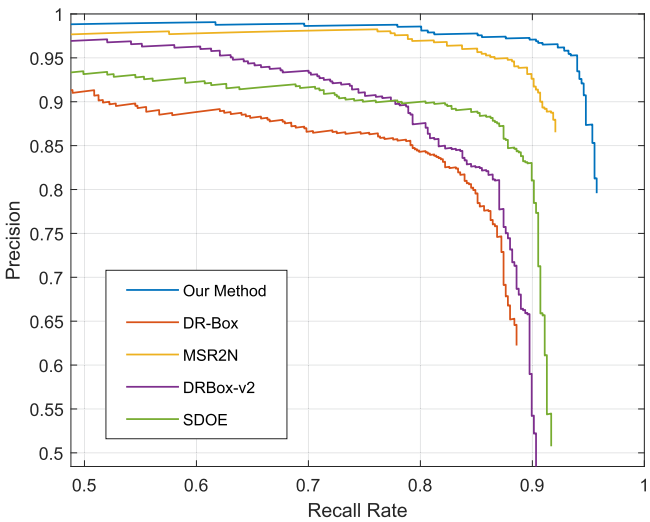


Fig. 21. PR curves of different models on SSDD+ ($d = 50$).

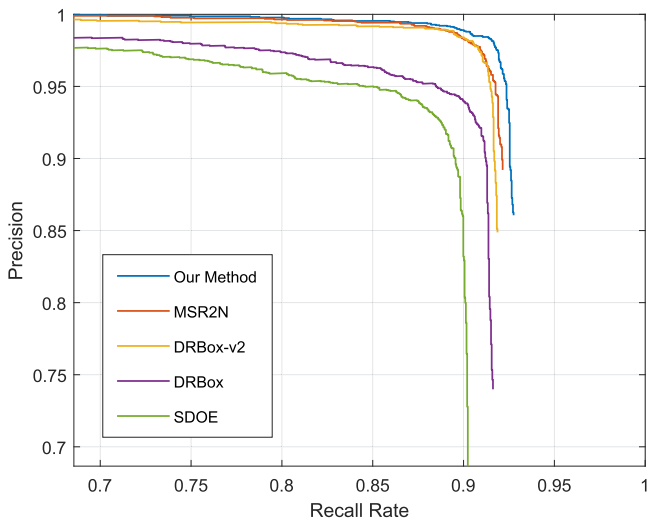


Fig. 22. PR curves of different models on GF3-Ship ($d = 50$).

87.78%) that does not apply any additional improvements. For MSR2N, the multilevel bounding box regression enables it to obtain higher detection accuracy than the basic R-RetinaNet and other one-stage models, but the multilevel bounding box regression also severely reduces the detection speed of MSR2N.

With the help of the proposed methods, our improved R-RetinaNet not only approaches or even surpasses MSR2N in AP, but also outperform the multistage MSR2N in detection speed because of the high detection efficiency of one-stage model. In addition, it can be seen from Figs. 21 and 22 that the precision of our method under different recall rates is close to or better than MSR2N.

Compared with other one-stage state-of-the-art methods, it can be seen from Fig. 19 that our model has a higher recall rate on small targets. This may be due to the fact that other one-stage models only use some feature maps with a smaller resolution, resulting in that smaller targets cannot be matched to feature maps with sufficient resolution. Our model is not susceptible to the interference of near-shore nonship objects, which results in a lower false alarm rate. This phenomenon may be because TA-FPN can make the model have better classification performance and avoid the model from misclassifying the background as ship targets. Besides, Fig. 20 also proves that the proposed method has better detection performance on high-resolution SAR images than other one-stage state-of-the-art methods.

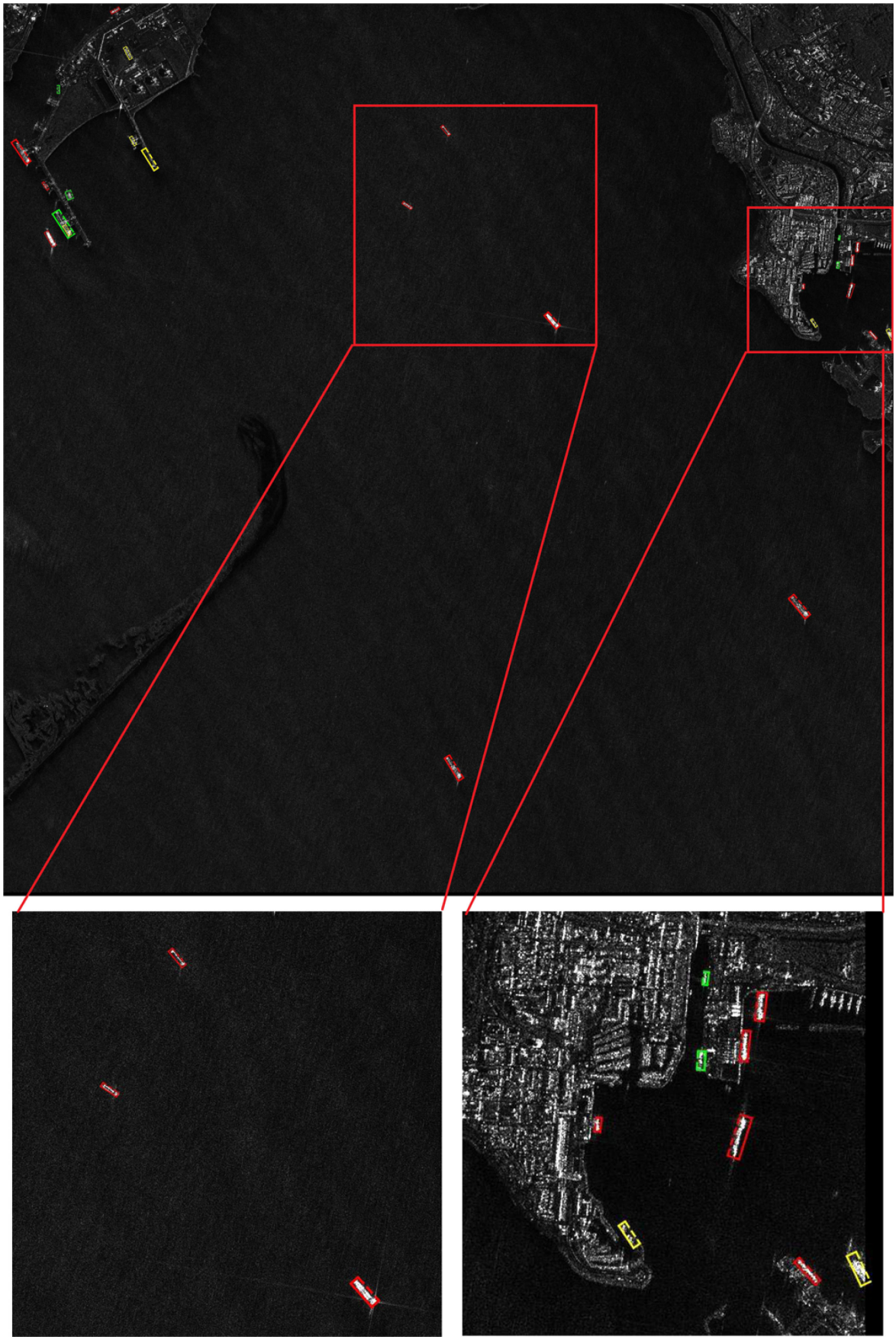


Fig. 23. Detection results of large scene SAR image. Note that the red boxes represent true positive targets, the yellow boxes represent false positive targets, and green boxes represent missed targets.

TABLE XIII
DESCRIPTIONS OF THE LARGE SCENE IMAGE

Parameter	Value
Sensor	GF3 satellite
Waveband	C
Image Size	3000×3000
Resolution	3m
Polarization Mode	VV
Sea Condition	Level 2

In addition to ResNet50, we also used the deeper backbone network ResNet101 to test the proposed method. The experimental results show that the proposed methods can significantly boost the detection performance on different backbone networks. At the same time, the comparison between different backbone networks shows that using a larger backbone network cannot provide a stable performance improvement. This may be because the smaller backbone network is sufficient to fit the training data.

Overall, experimental results show that the proposed method can provide competitive performance on both SSDD+ and GF3-Ship while maintaining a relatively fast inference speed.

8) Validation on Large Scene SAR Image: In order to test the practicability of the proposed model, we use a large scene image retained when making the GF3-Ship dataset to test the proposed model trained on the GF3-Ship dataset. This scene contains inshore and offshore ships of different scales. The image parameter information is shown in Table XIII.

First, the large scene SAR image is vertically and parallelly cropped by 800×800 pixels sliding window to provide a suitable input size for the model; each successively cropped image has an overlapped ratio of 25% to ensure the stitching process can be implemented. Second, 25 cropped SAR images are input into the proposed model to get the detection results. Third, detection results are stitched to form the detected panoramic SAR image. The test results are shown in Fig. 23.

It can be seen from Fig. 23 that the proposed model performs well in detecting the offshore targets. As for the inshore scene, false alarms and missed detections still exist. This may be due to the small number of training samples in the GF3-Ship dataset. Fewer training samples may cause the model to overfit on the training set, which in turn affects the generalization performance of the model.

9) Potential Application of the Proposed Methods on Optical Images: The proposed scale calibration method is used to solve the problem of misalignment between the feature map scale and the target scale distribution. TA-FPN is used to alleviate the learning conflict between classification tasks and regression tasks. The adaptive IoU training method is mainly used to solve the problem of imbalance of positive samples of large aspect ratio targets in the training phase. The above three problems are not unique to the SAR image ship detection task. The same problem may also exist in the task of optical image target detection, so the applicability of the proposed methods in the task of target detection from optical images is also worthy of further exploration.

IV. CONCLUSION

In this article, an RBox-based neural network detection method is proposed for SAR image ship detection. Experiments show that the proposed feature map scale calibration method can effectively align the scale distribution of the output feature map of the backbone network with the scale distribution of the targets, which greatly improves the performance of the model; the proposed TA-FPN can automatically adapt shared features to different learning tasks, which alleviates the conflict between different learning tasks. In addition, the proposed AIT training method effectively suppresses the positive sample intraclass imbalance problem in the RBox-based detection method and reduces variance of the number of positive samples. Compared with other one-stage RBox-based state-of-the-art methods, our model obtained the highest AP, which proves the superiority of the proposed methods.

Furthermore, the detection method proposed in this article uses a detection architecture based on anchor. Anchor-based architecture is suitable for natural scene images with dense targets. However, targets in SAR images are often very sparse. Therefore, the application of anchor-free architecture that is suitable for sparse targets in SAR ship detection should become the key direction of our future research.

REFERENCES

[1] S. Brusch, S. Lehner, T. Fritz, M. Soccorsi, A. Soloviev, and B. van Schie, "Ship surveillance with TerraSAR-X," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1092–1103, Mar. 2011.

[2] Q. An, Z. Pan, and H. You, "Ship detection in Gaofen-3 SAR images based on sea clutter distribution analysis and deep convolutional neural network," *Sensors*, vol. 18, no. 2, pp. 334–355, 2018.

[3] W. Ao, F. Xu, Y. Li, and H. Wang, "Detection and discrimination of ship targets in complex background from spaceborne ALOS-2 SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 2, pp. 536–550, Feb. 2018.

[4] W. G. Pichel, P. Clemente-Colón, C. Wackerman, and K. S. Friedman, "Ship and wake detection," Synthetic Aperture Radar, Marine User's Manual, U.S. Dept. Commerce, Washington, DC, USA, pp. 277–303, 2004.

[5] K. Eldhuset, "An automatic ship and ship wake detection system for spaceborne SAR images in coastal regions," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 4, pp. 1010–1019, Jul. 1996.

[6] C. C. Wackerman, K. S. Friedman, W. G. Pichel, P. Clemente-Colón, and X. Li, "Automatic detection of ships in RADARSAT-1 SAR imagery," *Can. J. Remote Sens.*, vol. 27, no. 5, pp. 568–577, 2001.

[7] T. Li, Z. Liu, R. Xie, and L. Ran, "An improved superpixel-level CFAR detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 184–194, Jan. 2018.

[8] J. A. Lorenzzetti, R. L. Paes, and D. M. Gheradi, "A performance comparison of a CFAR ship detection algorithm using EnviSat, RadarSat, COSMO-SkyMed and Terra SAR-X images," in *Proc. 3rd Int. Workshop ESRIN*, 2010, vol. 679, p. 32.

[9] R. Pelich, N. Longépé, G. Mercier, G. Hajduch, and R. Garello, "AIS-based evaluation of target detectors and SAR sensors characteristics for maritime surveillance," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 3892–3901, Aug. 2015.

[10] G. Gao, Y. Luo, K. Ouyang, and S. Zhou, "Statistical modeling of PMA detector for ship detection in high-resolution dual-polarization SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4302–4313, Jul. 2016.

[11] G. Gao, K. Ouyang, Y. Luo, S. Liang, and S. Zhou, "Scheme of parameter estimation for generalized gamma distribution and its application to ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1812–1832, Mar. 2017.

[12] J. Li and E. G. Zelnio, "Target detection with synthetic aperture radar," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 2, pp. 613–627, Apr. 1996.

[13] X. Leng, K. Ji, X. Xing, S. Zhou, and H. Zou, "Area ratio invariant feature group for ship detection in SAR imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2376–2388, Jul. 2018.

[14] X. Leng, K. Ji, S. Zhou, and H. Zou, "Noncircularity parameters and their potential in ship detection from high resolution SAR imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 1876–1879.

[15] R. Wang, J. Zhang, J. Chen, L. Jiao, and M. Wang, "Imbalanced learning-based automatic SAR images change detection by morphologically supervised PCA-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 554–558, Apr. 2019.

[16] S. Song, B. Xu, Z. Li, and J. Yang, "Ship detection in SAR imagery via variational Bayesian inference," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 319–323, Mar. 2016.

[17] A. Marino, N. Walker, and I. Woodhouse, "Ship detection with RadarSat-2 Quad-Pol SAR data using a notch filter based on perturbation analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2010, pp. 3704–3707.

[18] G. Ferrara, M. Migliaccio, F. Nunziata, and A. Sorrentino, "GK-based observation of metallic targets at sea in full-resolution SAR data: A multipolarization study," *IEEE J. Ocean. Eng.*, vol. 36, no. 2, pp. 195–204, May 12, 2011.

[19] M. Yeremy, J. Campbell, K. Mattar, and T. Potter, "Ocean surveillance with polarimetric SAR," *Can. J. Remote Sens.*, vol. 27, no. 4, pp. 328–344, 2001.

[20] D. J. Crisp, "The state-of-the-art in ship detection in synthetic aperture radar imagery," Defence Sci. Technol. Org. Salisbury (Australia) Info Sci. Lab, Salisbury, SA, Australia, Tech. Rep. DSTO-RR-0272, 2004.

[21] J. Chen, Y. Chen, and J. Yang, "Ship detection using polarization cross-entropy," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 723–727, Oct. 2009.

[22] R. Touzi and F. Charbonneau, "Characterization of target symmetric scattering using polarimetric SARs," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2507–2516, Nov. 2002.

[23] F. Nunziata, M. Migliaccio, and C. E. Brown, "Reflection symmetry for polarimetric observation of man-made metallic targets at sea," *IEEE J. Ocean. Eng.*, vol. 37, no. 3, pp. 384–394, Jul. 2012.

[24] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with SVD networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5832–5845, Oct. 2016.

[25] W. Fan, F. Zhou, X. Bai, M. Tao, and T. Tian, "Ship detection using deep convolutional neural networks for PolSAR images," *Remote Sens.*, vol. 11, no. 23, pp. 2862–2889, 2019.

[26] A. Marino, M. J. Sanjuan-Ferrer, I. Hajnsek, and K. Ouchi, "Ship detection with spectral analysis of synthetic aperture radar: A comparison of new and well-known algorithms," *Remote Sens.*, vol. 7, no. 5, pp. 5416–5439, 2015.

[27] A. Renga, M. D. Graziano, and A. Moccia, "Segmentation of marine SAR images by sublook analysis and application to sea traffic monitoring," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1463–1477, Mar. 2019.

[28] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[29] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[30] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 779–788.

[33] W. Liu *et al.*, "Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 21–37.

[34] J. Zhao, Z. Zhang, W. Yu, and T.-K. Truong, "A cascade coupled convolutional neural network guided visual attention method for ship detection from SAR images," *IEEE Access*, vol. 6, pp. 50693–50708, 2018.

[35] M. Kang, X. Leng, Z. Lin, and K. Ji, "A modified faster R-CNN based on CFAR algorithm for SAR ship detection," in *Proc. Int. Workshop Remote Sens. Intell. Process.*, 2017, pp. 1–4.

[36] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. SAR Big Data Era, Models, Methods Appl.*, 2017, pp. 1–6.

[37] J. Jiao *et al.*, "A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection," *IEEE Access*, vol. 6, pp. 20881–20892, 2018.

[38] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Nov. 2019.

[39] Y. Wang, C. Wang, and H. Zhang, "Combining a single shot multibox detector with transfer learning for ship detection using Sentinel-1 SAR images, "*Remote Sens. Lett.*, vol. 9, no. 8, pp. 780–788, 2018.

[40] C. Chen, C. He, C. Hu, H. Pei, and L. Jiao, "A deep neural network based on an attention mechanism for SAR ship detection in multiscale and complex scenarios," *IEEE Access*, vol. 7, pp. 104848–104863, 2019.

[41] Z. Lin, K. Ji, X. Leng, and G. Kuang, "Squeeze and excitation rank faster R-CNN for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 751–755, May 2019.

[42] X. Zhang *et al.*, "A lightweight feature optimizing network for ship detection in SAR image," *IEEE Access*, vol. 7, pp. 141662–141678, 2019.

[43] Y. Wang, C. Wang, and H. Zhang, "Ship classification in high-resolution SAR images using deep learning of small datasets," *Sensors*, vol. 18, no. 9, pp. 2929–2944, 2018.

[44] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery," *Remote Sens.*, vol. 11, no. 5, pp. 531–545, 2019.

[45] Z. Deng, H. Sun, S. Zhou, and J. Zhao, "Learning deep ship detector in SAR images from scratch," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4021–4039, Jun. 2019.

[46] T. Zhang, X. Zhang, J. Shi, and S. Wei, "Depthwise separable convolution neural network for high-speed SAR ship detection," *Remote Sens.*, vol. 11, no. 21, p. 2483, pp. 2483–2520, 2019.

[47] H. Su *et al.*, "HQ-ISNet: High-quality instance segmentation for remote sensing imagery," *Remote Sens.*, vol. 12, no. 6, pp. 989–1013, 2020.

[48] S. Wei *et al.*, "Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet," *Remote Sens.*, vol. 12, no. 1, pp. 167–196, 2020.

[49] J. Wang, C. Lu, and W. Jiang, "Simultaneous ship detection and orientation estimation in SAR images based on attention module and angle regression," *Sensors*, vol. 18, no. 9, pp. 2851–2868, 2018.

[50] C. Chen, C. He, C. Hu, H. Pei, and L. Jiao, "MSARN: A deep neural network based on an adaptive recalibration mechanism for multi-scale and arbitrary-oriented SAR ship detection," *IEEE Access*, vol. 7, pp. 159262–159283, 2019.

[51] Q. An, Z. Pan, L. Liu, and H. You, "DRBox-v2: An improved detector with rotatable boxes for target detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8333–8349, Nov. 2019.

[52] Z. Pan, R. Yang, and Z. Zhimin, "MSR2N: Multi-stage rotational region based network for arbitrary-oriented ship detection in SAR images," *Sensors*, vol. 20, no. 8, pp. 2340–2358, 2020.

[53] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 4203–4212.

[54] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11563–11572.

[55] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 784–799.

[56] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7132–7141.

[57] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2980–2988.

[58] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2117–2125.

[59] Y. Jiang *et al.*, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*.

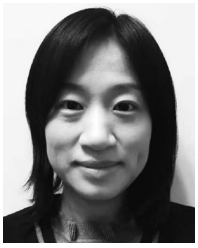
[60] X. Sun, Z. Wang, and S. Yuanrui, "AIR-SARShip-1.0: High-resolution SAR ship detection dataset," *J. Radars*, vol. 8, no. 6, pp. 852–862, 2019.

Rong Yang received the B.S. degree in electronic information science and technology from Xidian University, Xi'an, China, in 2016. He is currently pursuing the Ph.D. degree in communication and information system with Space Microwave Remote Sensing System Department, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

He is currently with the University of Chinese Academy of Sciences, Beijing, China. His research interests include object detection and semantic segmentation in synthetic aperture radar image.

Zhenru Pan is currently pursuing the Ph.D. degree in communication and information system with Space Microwave Remote Sensing System Department, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

She is currently with the University of Chinese Academy of Sciences, Beijing, China. Her research interests include synthetic aperture radar ship imaging, detection, and identification.



Xiaoxue Jia received the Ph.D. degree in communication and information system from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2010.

She is currently an Associate Research Fellow with the Aerospace Information Research Institute, Chinese Academy of Sciences. Her current research focuses on SAR signal processing techniques.

Lei Zhang was born in Jilin, China, in 1985. He received her Ph.D. degree in communication and information system from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2014.

He is currently an Associate Research Fellow with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include high-resolution synthetic aperture radar imaging and signal processing.

Yunkai Deng (Member, IEEE) received the M.S. degree in electrical engineering from the Beijing Institute of Technology, Beijing, China, in 1993.

In 1993, he joined the Institute of Electronics, Chinese Academy of Sciences (IECAS), Beijing, China, where he was involved in antenna design, microwave circuit design, and spaceborne/airborne synthetic-aperture radar (SAR) technology. Since 1993, he has been a Research Fellow with the Department of Space Microwave Remote Sensing System, IECAS. He has been the Leader of several spaceborne/airborne SAR programs and developed some key technologies of spaceborne/airborne SAR. Since 2012, he has been a Principal Investigator with the Helmholtz-Chinese Academy of Sciences (CAS) Joint Research Group, Beijing, China, concerning spaceborne microwave remote sensing for prevention and forensic analysis of natural hazards and extreme events. He is currently a Research Scientist with the University of Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 100 articles, since 2002, of which more than 100 peer-reviewed and well-known journal articles. His research interests include spaceborne/airborne SAR technology for advanced modes, multifunctional radar imaging, and microwave circuit design.

Mr. Deng is a member of the Scientific Board. He was a recipient of several prizes, including the First and Second Class Rewards of National Defense Science and Technology Progress, in 2007, the First Class Reward of the National Scientific and Technological Progress, in 2008, the achievements of the Outstanding Award of the CAS, in 2009, and the First Class Reward of Army Science and Technology Innovation, in 2016, for his outstanding contribution in SAR field.