A Generating-Anchor Network for Small Ship Detection in SAR Images

Tingxuan Yue, Yanmei Zhang, Pengyun Liu, Yanbing Xu, and Chengcheng Yu

***Abstract*—Synthetic aperture radar (SAR) ship detection especially for small ships has issues, such as dense distribution of ships, interference from land and small islands. To address these issues, many deep learning methods, including anchor-based and anchor-free methods, have been successfully migrated from optical scenes to SAR images. However, when the preset scale of anchors does not match well with the ships, it will seriously reduce the detection precision. Due to the lack of anchor-based refinement process, anchor-free methods may generate missing or false alarms in complex scenarios. In this article, a two-stage ship detection network which can generate anchors is proposed. First, our method generates high-quality anchors by network, which is more beneficial for the network to capture small ships. In addition, the generated anchors are centrally set in the region of ships, which reduces the number of anchors unrelated to ships. Second, the receptive field enhancement module is inserted into the feature pyramid network. It sets different dilation ratios of atrous convolution according to the scale of the feature map, which further enriches the semantic information of the elements in the feature map. Therefore, the network can use the information of a wider region effectively to detect ships. Finally, to verify the effectiveness of our method, extensive experiments are carried out on SAR ship detection dataset and high-resolution SAR images dataset. The results show that our method has more strong ability of detecting small ships, and achieves better detection performance than some state-of-the-art methods.**

***Index Terms*—Deep learning (DL), ship detection, small-scale target, synthetic aperture radar (SAR).**

I. Introduction

With the propagation characteristics of electromagnetic waves, weather conditions have less impact on SAR than optical remote sensing sensor, and the SAR can work all day. As a valuable research topic in the field of SAR image processing, ship detection plays an important role in sea surface monitoring and fishery management [1]. Unfortunately, detecting ships in complex environments (such as areas near land and little islands) is still not a completely resolved task for researchers. What is more, detection of small ships is also a great challenge [2].

The backscatter signal of ships is typically stronger than the sea surface, resulting in its area being brighter than the surrounding background in SAR images [3]. The constant false alarm rate (CFAR) is generally introduced into detection [3], [4]. For detecting ships, a bilateral CFAR algorithm combined the intensity (i.e., brightness) distribution and spatial information of the SAR image [5]. The two-parameter CFAR detector with polarimetric whitening filter was derived under the distribution of clutters including Wishart, K-Wishart, F-Wishart, etc., [3]. However, the complex background can affect the performance of CFAR detector [5], [6], e.g., radio frequency interference [7] may cause mismatch of clutter model, and the CFAR detector may suffer performance degradation in multitarget scenarios [8]. The texture and contour of the SAR images have been extracted from the gray value as another type of features for ship detection [9]. Gao [10] investigated the effectiveness of multiple features (like spatial boundary features, fractal dimension feature), and extracted signal-to-noise-ratio (SNR) features for SAR target detection.

Wang et al. [11] obtained the complete structures of the bright area via superpixel segmentation and Bayesian framework, then the morphological features was used to distinguish target from clutter. The fisher vector (FV) represents more characteristics of superpixel than its intensity values, and it contains the zero-order, first-order, and second-order feature for ship detection [8]. Subsequently, the classifier completes the detection of ships in the feature space [12]. The widely used classifiers, like support vector machine (SVM), adaptive boosting, and so on, achieve accurate detection performance in the suitable scenarios. Nevertheless, due to the influence of speckle noise and small islands, false alarms often occur in these traditional methods based on image processing. Whenever an unknown scattering of ships appears or characteristic of interference changes turbulently, it takes a relatively long time for scholars to design new features manually.

The vigorous development of deep learning (DL) technology has promoted computer vision (CV) to a new stage. The powerful learning ability of neural networks eliminates the need for scholars to design features manually. Influenced by the significant performance in CV field, the convolutional neural network (CNN) has been introduced to detect targets in remote sensing images [13]. The anchor-based CNNs develop along two paths: 1) the single-stage with high operating speed and 2) the two-stage with high precision. The representative algorithms of single-stage methods are You Only Look Once (YOLO) [14], RetinaNet [15] and Single Shot Detection (SSD) [16]. In order to make YOLOv4 network [17] more suitable for ship detection in SAR images, Gao et al. [18] introduced scale-equalizing pyramid convolution module and convolutional block attention

Manuscript received 23 June 2022; revised 15 August 2022; accepted 1 September 2022. Date of publication 6 September 2022; date of current version 15 September 2022. This work was supported by the National Key Research and Development Program of China under Grant 61404130221. *(Corresponding author: Yanmei Zhang.)*

The authors are with the Beijing Institute of Technology, Beijing 100081, China (e-mail: yue.tx@163.com; zhangym0726@163.com; bit_lpy@163.com; 3120185388@bit.edu.cn; 3120170372@bit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3204578

module into the network, and modified the head of the YOLOv4. Represented by faster R-CNN [19], the two-stage method with the region proposal network (RPN) trades speed for an increase in detection precision. To alleviate the multiscale problem within target in the ship detection, Deng et al. [2] proposed a ship detection network that incorporates a design of multiscale filter based on the two-stage network structure, and the redesigned backbone network is compact in order to improve the training efficiency. The two-stage network can be regarded as the first stage of rough classification and the second stage of fine classification, which can further improve the accuracy of ship detection. Hou et al. [20] used the SSD network as the first stage, then constructed RefineDNet to improve the confidence of potential objects from the first stage. Zhang et al. [21] proposed an SAR ship detector based on the faster R-CNN incorporating four balanced strategies and verified effectiveness of the detector for solving scene imbalance and sample imbalance on multiple public datasets. Besides the above anchor-based CNNs, the anchor-free CNNs, like CenterNet [22], FCOS [23] have also been introduced to ship detection. The CenterNet has been used as a low-computation ship detector in [9]. The detector added the spatial shuffle-group enhance attention module for capturing features accurately under the interference of noise. Hu et al. [24] introduced deformable convolution into FCOS to capture more effective information for ships, and the nonlocal attention mechanism in the network effectively balanced the local information of the feature map. Gao et al. [27] proposed a novel feature aggregation scheme to enhance representation ability of the features, and the feature reuse strategy of the scheme improved the generalization ability of the model. Fu et al. [26] retained the overall architecture of the FCOS, and proposed a module for mitigating interference from objects adjacent to the ships. And an intersection over union (IoU) prediction branch was inserted into head of the network for the bounding box regression of small-scale ships.

Due to the small radar cross-section, small ships are little-scale and weak-intensity in SAR images, which are very easy to be confused with islands and speckle noise for CNNs. In [27], the inception module was adopted to increase the receptive field that can capture small ships more effectively. Cui et al. [28] integrated spatial and channel attention into the feature pyramid network (FPN) structure, which strengthened the important information in the small-scale feature map and improved the detection precision of small targets. In order to improve the ability of network to detect small ships, Wang et al. [29] added a nonlocal attention mechanism as a module on the SSD to enrich the semantic information of feature maps. The coordinate attention module was used to capture horizontal and vertical correlations on feature maps in [30], then these feature maps are processed by receptive field boosting to effectively reduce false alarms. For improving the positioning accuracy of small ships and reduce false alarms of nonships, Chen et al. [31] derived a shape similarity IoU loss to instead the original loss function of bounding box regression. Su et al. [32] used multiscale pooling operation to upgrade location information of small ships at the high-level features. Zeng et al. [33] pioneered the utilization of low-level feature to match the receptive field of small ships, the low-level features used to contain regional and texture information for capturing small ships.

In the abovementioned ship detection network, the researchers extended the DL algorithm based on optical target detection at close range to the ship detection in SAR images. With the advantages of high accuracy, the ship detectors based on the DL algorithm have become a research hotspot in the field of remote sensing. However, SAR images contain less information compared with optical images and have interference, such as clutter, the inherent shortcomings of these methods may be further revealed in SAR ship detection. The anchor-based methods regress the bounding box of target based on the designed anchor, but the designed anchors are usually placed uniformly over the entire feature map, resulting in a huge computational cost. Ships are sparsely distributed in SAR images, which means that the proportion of images occupied by ships, especially small ships, is generally small. Sampling of the corresponding ocean region on the feature map by the anchor will generate a large number of negative samples unrelated to the ships, which wastes computing resources. The detection performance of anchor-based methods are very sensitive to the setting of anchor hyperparameters. The methods with fixed scale and aspect ratio of anchors, such as faster R-CNN and RetinaNet, cannot be applied to all resolutions of ships datasets. Methods that spend the effort to design anchors manually, such as YOLOv3 [34], may have an unstable performance for large-scale variation within the class. If the scale of preset anchors is not small enough, the ability of the network to detect small ships will be severely affected. Although the anchor-free method avoids the effort of adjusting the hyperparameters and reduces the amount of computation caused by anchors, the anchor-free methods lack further refinement based on anchors, resulting in a lack of ability to handle complex scenes and cases [35]. When the ships are parked close to the shore or are distributed densely, the performance of the anchor-free method to predict the bounding box will decrease. Furthermore, the SAR images are grayscale images, which lack color information that helps to directly regress the bounding box without setting anchors.

To overcome the obstacles of the above anchor-based and anchor-free methods while obtaining stronger detection capability for small ships, we propose an SAR ship detection network capable of generating anchors. Our method predicts the shape of the anchor and its location on the feature map. The major contributions of our work can be summarized as follows.

1) To reduce false candidates unrelated to ships while preserving the design of anchor for high-accuracy detection performance, we propose a generating anchors module (GAM). The GAM receives the multiscale feature maps from the FPN and predicts the position and shape of the anchor. The anchor generated by the network can more effectively handle ships with various aspect ratios.

2) Feature maps with rich semantic ranges can provide spatial interaction information in the scene. Therefore, we design a receptive fields enhancement module (RFEM) for improving the ability of the network to locate ships. The feature maps with multisize receptive fields from the RFEM are merged into a new feature through channel then fed into the FPN.

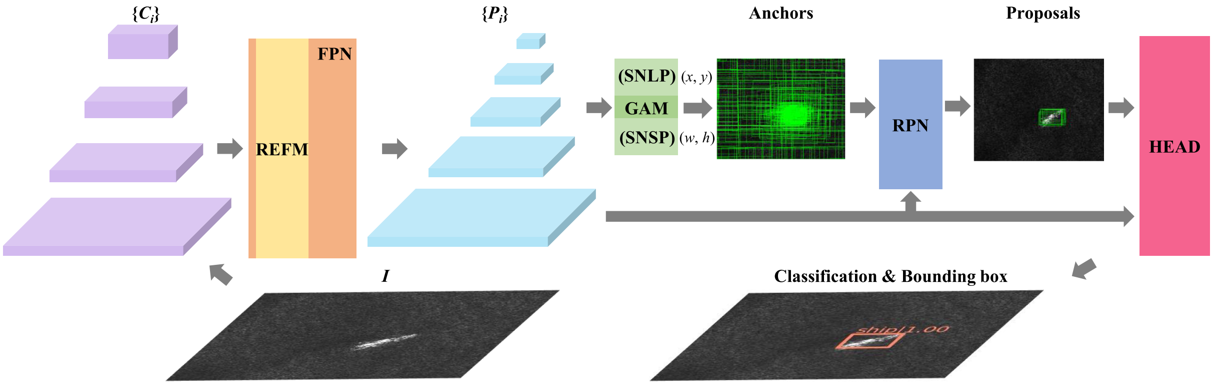


Fig. 1. Overall architecture of our proposed ship detector.

3) To verify the effectiveness of our method, we conduct extensive experiments on two widely used SAR image ship datasets: SAR ship detection dataset (SSDD) [36] and high-resolution SAR images dataset (HRSID) [37]. Our method attains AP of 64.8 and 66.5, AP_{50} of 95.7 and 91.1 on SSDD and HRSID, respectively, as well as achieves better detection performance for small ships than popular detectors.

The rest of this article is organized as follows: The Section II presents the details of our method, and Section III analyzes the effectiveness of our method with experimental data. Finally, Section IV concludes this article.

II. Proposed Method

This section describes the proposed method in detail, and Fig. 1 shows the overall architecture of our proposed ship detection network. An SAR image is input into the convergent network that has been trained. It is the first extracted feature by the backbone network, and the network obtains multiple feature maps of different scales. The resolution of these feature maps is gradually reduced, and the semantic content of which is enriched scale by scale. They are then fed into the RFEM, and the receptive fields of elements in each scale feature map are enhanced, facilitating the GAM to generate anchors that are more similar in shape to ships. These feature maps processed by the RFEM enter into the GAM after completing multiscale feature fusion in the FPN. The subnetwork of location prediction (SNLP) in the GAM filters the location of the center point, where the anchor is set on the feature map. And the subnetwork of shape prediction (SNSP) in the GAM generates the height (h) and width (w) of the anchor corresponding to the location from the SNLP. Subsequently high-quality anchors and feature maps output by FPN are fed into RPN and the head of the network to complete proposal extraction, bounding box refinement, and target classification in turn.

A. Basic Framework

Compared with the single-stage network, the two-stage network has one more RPN. Although the two-stage network has the inherent disadvantage of slow inference speed, the authors in [21] and [28] choose to develop ship detectors based on a two-stage network framework due to its high detection precision. With the improvement of computing power, the time-consuming gap between the two-stage networks and the single-stage networks will be further narrowed. And our method can reduce the amount of computation caused by invalid candidates, so our method adapts the two-stage network as the basic frame. The two-stage network framework with the FPN inserted is shown in Fig. 2. The feature maps $\{C_2, C_3, C_4, C_5\}$ obtained from the bottom-up path in the backbone fuse with the feature maps $\{A_2, A_3, A_4, A_5\}$ generated by top-down paths in the FPN via lateral connections. The new feature maps $\{P_2, P_3, P_4, P_5\}$ contain the semantic information in the higher layers and retain the location information of target in the lower layers

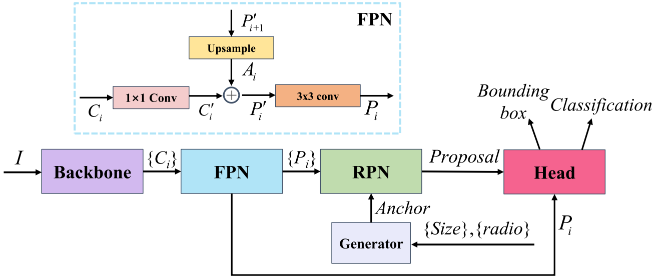


Fig. 2. Two-stage network framework with the FPN inserted.

$$A_i = \text{Upsample}(P'_{i+1}), i = 2, 3, 4 \tag{1}$$

$$P'_i = \begin{cases} A_i \oplus \text{Conv}_{1\times1}(C'_{i+1}), i = 2, 3, 4 \\ \text{Conv}_{1\times1}(C_{\text{i}}), i = 5 \end{cases} \tag{2}$$

$$P_{\text{i}} = \begin{cases} \text{Conv}_{3\times3}(P'_i), i = 2, 3, 4, 5 \\ \text{Downsample}(P_{\text{i-1}}), i = 6. \end{cases} \tag{3}$$

Since the number of channels among C_i is inconsistent, the FPN first obtain C'_i with the same number of channels through 1×1 convolution. To detect large ships, we downsample P_5 to obtain P_6, so the downsampling factor of $\{P_2, P_3, P_4, P_5, P_6\}$ corresponding to the input sample image $I \in \mathbb{R}^{C\times H\times W}$ is $s_i = \{4, 8, 16, 32, 64\}$. Therefore, the number of anchors $N_{\text{anchor\_}i}$ set on the feature map P_i is as follows:

$$N_{\text{anchor\_}i} = \text{Ceil}(H/s_i) \times \text{Ceil}(W/s_i) \times N_{\text{size}} \times N_{\text{ratio}} \tag{4}$$

where N_{size} and N_{ratio} are the number of two preset hyperparameters (the size and the aspect ratio of anchor), respectively. The RPN judges whether the anchors contain targets, and regresses the anchors containing targets as proposals. The nonmaximum suppression (NMS) is used to filter proposals for obtaining regions of interest (ROIs). Then the network assigns the ROIs to the feature map P_i according to the scale. The detection head of the network extracts the features of the corresponding ROIs to complete the classification of targets and the refinement of bounding boxes. In the two regression of bounding box in the two-stage network, the network does not directly predict the center coordinates (x, y), height and width (w, h) of the box. When the RPN generates proposals, the network outputs the difference between the anchor and the ground truth (GT) box, i.e., the panning amount and the transformation scale (x_t, y_t, w_t, h_t). The parameters (x, y, w, h) of the proposal can be decoded from the parameters of the anchor (x_a, y_a, w_a, h_a).

$$x = x_t w_a + w_a, y = y_t h_a + h_a \tag{5}$$

$$w = w_a e^{w_t}, h = h_a e^{h_a}. \tag{6}$$

The e in (6) is the base of the natural logarithmic function. When refining the bounding box, the network predicts the difference between the ROI and the GT box, and the decoding method is the same as (5) and (6).

B. GAM

In the anchor-based network, the anchors are densely set on the image, most of the anchors in the SAR ship detection are set in the ocean area, which causes the RPN to waste a lot of time for judging whether the anchor contains ships. The ships are long strips and sail at any direction in the ocean, so the aspect ratio of the GT boxes is usually quite different. The preset and clustered from dataset anchors are not robust enough for ship detection. Inspired by [35], we propose the GAM with supervision of aspect ratio to generate anchors. The location(x, y) and (w, h) of the ship's bounding box on an SAR image I follow a conditional probability density distribution

$$p(x, y, w, h|I) = p(x, y|I)p(w, h|x, y, I). \tag{7}$$

The $p(x, y|I)$ means that the ships appear in a specific location on the image, i.e., the probability of placing an anchor on each point of the feature map is different. And the $p(w, h|x, y, I)$ means that the shape of the ships bounding box is related to the location of the ship on the image, that is, the (w, h) of anchor on each location has a relationship between the location on the feature map. Based on (7), the GAM structure is shown in the Fig. 3. This module contains two branches: the SNLP and the SNSP. In the SNLP, each position (x, y) in the feature map P_i corresponds to the coordinate $((x + \frac{1}{2})s_i, (y + \frac{1}{2})s_i)$ on the input image I. The $p(x, y|P_i)$ indicates the probability that the ship exists in this location. The 1×1 convolution is applied to P_i for obtaining the score map of ships existence. The score map is processed by the sigmoid layer to generate the probability map

$$p(x, y|P_i) = \text{Sigmoid}(\text{Conv}_{1\times 1}(P_i)). \tag{8}$$

We take the location on the $p(x, y|P_i)$ where the value is higher than the predefined threshold ε as the (x_a, y_a) to place the anchors.

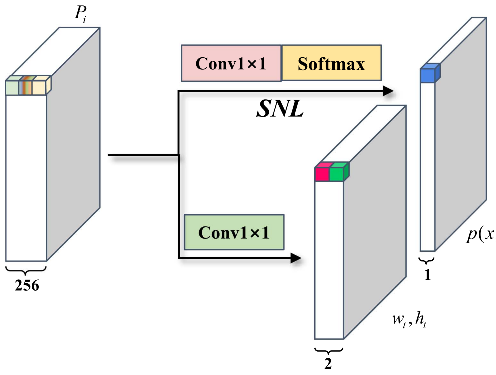


Fig. 3. Structure of the GAM.

The first term of the product in (7) is obtained by the SNLP, and the SNSP predicts the (w_a, h_a) of the anchor at (x_a, y_a). According to [15] and [19], the (6) can be used to obtain a more stable shape of anchor, therefore the SNSP predicts the transformation scale (w_t, h_t) and the (w_a, h_a) of anchor is as follows:

$$w_a = \sigma s_i e^{w_t}, h_a = \sigma s_i e^{h_t} \tag{9}$$

where σ is the scale factor, and the w_t and h_t come from a two-channel map generated by applying a 1×1 convolution on P_i. The (w_a, h_a) combines the location of anchor center (x_a, y_a) from the SNLP to obtain the anchor (x_a, y_a, w_a, h_a) that can better capture ships. It is worth noting that only one anchor is associated with each location, so the $N_{\text{anchor\_}i}$ on P_i changes from (4) to

$$N_{\text{anchor\_}i} = \text{Count}(p(x, y|P_i) > \varepsilon). \tag{10}$$

Compared to (4), the number of anchors drops significantly after applying the GAM. The number of positive and negative samples becomes more balanced.

C. RFEM

To improve the receptive field of elements in the feature map, pooling operation [38] or atrous convolution [39] can be performed. Although the pooling operation does not increase the number of parameters, it is easy to cause the feature map to be disturbed by noise, especially the strong interference of speckle in the SAR image. And the pooling operation reduces the resolution of the feature map. Therefore, atrous convolution is used in our method to enhance the receptive field while preserving the spatial information of the feature maps. For a 2-D feature map P, the Q obtained after atrous convolution can be expressed as

$$Q(i, j) = \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} P(i + rm, j + rn)W(m, n) \tag{11}$$

where (i, j) are coordinates on the feature map, and W and r are a convolution filter of size $K * K$ and dilated rate, respectively. When the dilation ratio is 1, the atrous convolution degenerates into an ordinary convolution.

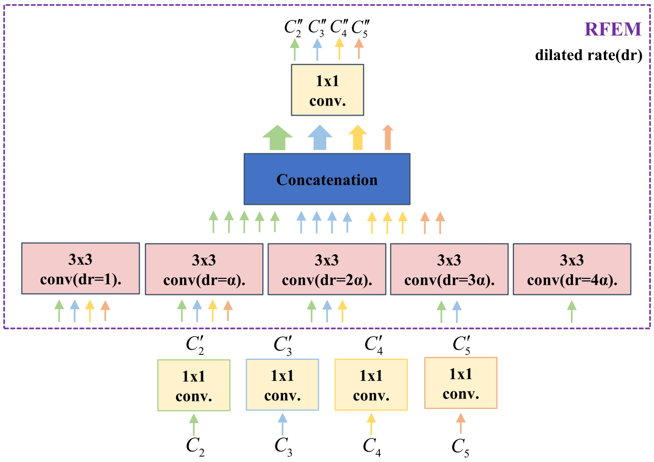


Fig. 4. Structure of the RFEM.

In the ASPP structure, the input feature map is responsible for predicting targets with size within a range, so the dilation rate of atrous convolution chosen to improve the receptive field are $\{6, 12, 18\}$. However, the FPN assigns targets to feature maps of different resolutions according to scale, and the proportion of images taken up by ships is relatively low in ship detection. Therefore, it is easy to introduce redundant information irrelevant to the ships by using the atrous convolution operation with large dilation rates on the high-level feature map. In the structure of our designed RFEM, the number of atrous convolutions is gradually reduced as the resolution of the feature map decreases. The P_5 of lowest resolution requires only one atrous convolution operation with the minimum dilated rate. The structure of RFEM is shown in Fig. 4, the RFEM is embedded after the channel reduction operation of $1{\times}1$ convolution in FPN. The $C'_2 \in \mathbb{R}^{256\times\frac{H}{4}\times\frac{W}{4}}$ with the highest resolution uses atrous convolution with dilated rates $\{1, \alpha, 2\alpha, 3\alpha, 4\alpha\}$ in parallel, and the α is the step size of the dilated rate. In order to ensure that the multiple output feature maps can be concatenated, these feature maps are made as the same shape like C'_2 by zero-padding operation before processing by atrous convolution. The feature map with 256×5 channels obtained by concatenatting is subjected to $1{\times}1$ convolution for the interaction between feature maps of different receptive fields, so the obtained $C''_2 \in \mathbb{R}^{256\times\frac{H}{4}\times\frac{W}{4}}$ has stronger representation ability. The operation for C'_3 is similar to that for C'_2, except that the dilated rates of atrous convolution is $\{1, \alpha, 2\alpha, 3\alpha\}$, by analogy, the dilated rates of the atrous convolution used for C'_5 is $\{1, \alpha\}$. The FPN and the RFEM are connected by cascade, so the C'_i in Fig. 2 is replaced by the output of RFEM C''_i in our method. The feature maps extracted from the backbone enhance the receptive field via the RFEM, then they complete the fusion of multiresolution features through a top-down path. The new feature maps output by FPN not only enable the network head to achieve better detection performance, but also promote GAM to generate higher quality anchors.

TABLE I
Three Types of Sample Labels on the Feature Map

Type	Region
Positive samples	Pixels in the center region $CR{=}R(x_g^{'}, y_g^{'}, \sigma_1 w_g^{'}, \sigma_1 h_g^{'})$
Ignore samples	Pixels in the ignore region $IR{=}R(x_g^{'}, y_g^{'}, \sigma_2 w_g^{'}, \sigma_2 h_g^{'})\backslash CR, \sigma_2 > \sigma_1$
Negative samples	Pixels in the outside region $OR = P_i \backslash (IR \bigcup CR)$

D. Loss

Our proposed ship detection network follows optimization approach of end-to-end via multitask loss. The multitask loss function Loss contains loss function of SNLP L_{SNLP} and loss function of SNSP L_{SNSP} from the GAM, loss function of classification L_{cls} and loss function of regression L_{reg} from the base framework. In the training of the network, the L_{cls} and L_{reg} are cross entropy loss and smooth L1 loss, respectively,

$$\text{Loss} = \lambda_1 L_{\text{SNLP}} + \lambda_2 L_{\text{SNSP}} + L_{\text{cls}} + L_{\text{reg}}. \tag{12}$$

The training of SNLP requires the region of ships occupation as label to calculate the L_{SNLP}, the label can be obtained directly from the GT box of ships. Since the higher initial IoU value appear when the center of anchor and GT box are closer, the locations in the center region of the GT boxes on the feature map are regarded as positive samples. In addition, we wish to set as few anchors as possible on the region far from the center of the GT boxes. First the GT box (x_g, y_g, w_g, h_g) must be mapped to the scale of the feature map P_i to get (x'_g, y'_g, w'_g, h'_g). The rectangular region is defined as $R(x, y, w, h)$ like the bounding box, the three types of sample labels on the feature map are defined, as Table I. The CR usually occupies a smaller portion on the feature map, so we use focal loss [15] as L_{SNLP}. Since the GAM is cascaded after the FPN, the assignment scheme of GT boxes in the FPN is still used when generating the binary label map.

In training, the basic framework assigns anchors for GT box to calculate the loss according to the maximum IoU value. But it is no longer applicable to the case where w and h are variables in the GAM. This problem is solved by approximating IoU with the variable IoU (vIoU) in the GAM

$$\text{vIoU}(a_{wh}, gt) = \max_{a_{\text{wh}} \in a_{\text{sample}}} \text{IoU}(a_{wh}, gt) \tag{13}$$

where $\text{IoU}(a_{wh}, gt)$ is the IoU between a anchor with (w, h) and GT box gt, and a_{sample} is the set of anchors with common (w, h) obtained by sampling. The nine pairs of sampling anchors in our experiments are the same as [15], i.e., the aspect ratio of anchors on the P_i are ratio $= \{0.5,1,2\}$, and the base scale of anchors base_scale $= 2^{\text{m}/3}, \text{m} = 0, 1, 2$. Compared with optical photos and optical remote sensing images, SAR images lack rich color boundary information to help the GAM predict the shape of anchors. Due to the interference, such as sea clutter, predicting the shape of anchor alone may lead to an error between the aspect ratio of anchor and ideal situation. Additionally the shape of the ships has higher requirements on the aspect ratio of the anchors.

As a result, the bounding box and GT box are not matched accurately enough, it is difficult to meet the requirements of the scene with a high IoU value between the bounding box and GT box. Therefore, we design the aspect ratio loss as the supervision for generating the anchor, and the L_{SNSP} is as follows:

$$L_{\mathrm{SNSP}} = L_1\left(1 - \min\left(\frac{w}{w_g}, \frac{w_g}{w}\right)\right) + L_1\left(1 - \min\left(\frac{h}{h_g}, \frac{h_g}{h}\right)\right) + \mathrm{MSE}\left(\frac{h_g}{w_g}, \frac{h}{w}\right) \tag{14}$$

where L_1 is the smooth L1 loss, and MSE is the mean square error loss.

III. Experiments

All of our experimental results are run by a computer with NVIDIA RTX 3090 GPU. The operating system is ubuntu 20.04, and the installed DL framework is Pytorch. Furthermore, our method is implemented based on the MMDetection Toolbox [40].

A. Datasets and Settings

To verify the effectiveness of our proposed module and test the performance of our ship detection network, we conduct hyperparameter experiments, ablation experiments, and comparison experiments with other popular networks on the SSDD. The SSDD has multiple data sources (RadarSat-2, TerraSAR-X, and Sentinel-1), and its resolution covers the range of 1–15 m. The SSDD includes 1160 samples ranging in side lengths between 500 pixels and 600 pixels, with a total of 2456 ships in these samples. Furthermore, to demonstrate the generalization of our method, we also conduct comparative experiments with other popular networks on the HRSID. The resolutions of samples in the HRSID are 0.5, 1, and 3 m, and the resolution of most samples is 3 m. The sample size of the HRSID is 800×800 pixels, and these samples come from TanDEM, TerraSAR-X and Sentinel-1. The HRSID has a total of 5604 samples and 16 951 ship labels, with an average of three ships on each sample. The two datasets also provide labels for inshore and offshore samples, which can test the performance of our method in complex environments. For the SSDD, we divide the training set and the testing set according to [36]. The raw images whose the last digits of the file number is 1 and 9 are used as the testing set, and the rest are used as the training set, i.e., 232 samples are used for testing and 928 samples are used for training. In order to facilitate the input of network, the samples in the SSDD are resized to 512×512 pixels. For keeping the aspect ratio of the samples, the zero-padding operation is used when resizing samples. The sample division plan of HRSID is in accordance with the [37], that is about 65% (3642 images) are the training set, and 35% (1962 images) are the testing set. According to the [21], we directly input the HRSID samples into network without resizing and padding.

In our experiments, the samples do not undergo any augmentation operations to fully demonstrate the performance of network. The backbone network of all networks except the HR-SDNet [41] is the ResNet-50 network [42] loaded with pretrained weights of the ImageNet from the torchvision. All networks are trained on GPU with batchsize of 1 for 12 epochs, and the configuration of the optimizer is in Table II. For promoting the network to converge, we also adopt a linear warm-up strategy of the learning rate. The number of warm-up iterations and the warm-up rate of this strategy are 500 and 0.001, respectively. The NMS is applied to filter redundant bounding boxes from the outputs of network, and the IoU threshold of NMS is set to 0.5. We set $\lambda_1 = 1.0$ and $\lambda_2 = 0.8$ to balance the loss terms in our method. The rest of hyperparameters follow the default settings of the MMDetection Toolbox.

TABLE II
Configuration of Optimizer

Parameter	value
Type	SGD
Initial learning rate	0.0025
Epochs of learning rate decay	[8, 11]
Decay rate of learning rate	0.1
Momentum	0.9
Weight decay	0.0001

B. Evaluation Metrics

To objectively evaluate the performance of the method, we adopt some quantitative metrics in this section, such as the most widely used recall (r) and precision (p). The precision-recall curve (PRC) can show precision and recall, and can describe their relationship specifically. Therefore, we introduce average precision (AP), AP_{50}, and AP_{75} from the evaluation metrics of COCO [43] to quantify the PRC for a more comprehensive evaluation instead of single precision. In addition, small ships have always been a nodus for SAR ship detection, so we adopted the AP_{s} and $AP_{50_{\mathrm{s}}}$ to evaluate the detection performance of small ships. The AP_{s} and $AP_{50_{\mathrm{s}}}$ are the AP and AP_{50} of small ships, respectively. The recall and precision are defined by

$$\mathrm{Recall} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \tag{15}$$

$$\mathrm{Precision} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} \tag{16}$$

where N_{TP}, N_{FN}, and N_{FP} are number of true positives (TP), false positives (FP), and false negatives (FN), respectively. In SAR ship detection, the TP is the correctly detected ship and the IoU between the bounding box and GT box is higher than 0.5. The FP is a false alarm or the IoU between the bounding box predicted for the ship and GT box is lower than 0.5, and the FN stands for missed ships

$$AP = \int_0^1 \mathrm{Precision}(r)\, dr. \tag{17}$$

The AP represents the area under the PRC. In addition, calculating the AP needs to set the IoU threshold between the GT box and the bounding box predicted for the ship to determine the TP. The AP_{50} is the area under the PRC curve with an IoU threshold of 0.5, and AP_{75} has an IoU threshold of 0.75 like AP_{50}. It is worth noting that the AP in the following content is the average

TABLE III
EXPERIMENTAL RESULTS OF HYPERPARAMETERS FOR ε IN THE GAM(%)

ε	AP	AP_{50}	AP_{75}
0	**65.6**	95.2	**76.9**
0.005	64.5	95.1	76.5
0.01	64.8	**95.7**	76.6
0.015	64.2	95.5	75.3
0.02	64.2	95.4	75.5

TABLE IV
EXPERIMENTAL RESULTS OF HYPERPARAMETERS FOR σ_1 AND σ_2 IN THE GAM(%)

σ_1	σ_2	AP	AP_{50}	AP_{75}
0.1	0.5	63.7	95.1	74.2
0.2	0.5	64.8	**95.7**	**76.6**
0.3	0.5	64.3	94.8	74.9
0.4	0.5	64.8	95.6	76.2
0.1	0.6	63.4	94.8	71.7
0.2	0.6	63.6	94.9	74.3
0.3	0.6	**65.2**	95.1	74.9
0.4	0.6	64.8	95.6	75.9
0.5	0.6	64.3	94.9	75.8

$AP_{\varepsilon_{\text{IoU}}}$ value of the IoU threshold $\varepsilon_{\text{IoU}} \in [0.5{:}0.05{:}0.95]$. Our definition of small ships follows the COCO, i.e., the ships whose GT box is area $< 32^2$ pixels are considered as small ships.

C. Hyperparameter Experiment

To maximize the effectiveness of the designed modules in our method, we conduct hyperparameter experiments for ε in GAM and step size of dilated rate (α) in RFEM. Except for the different values of the hyperparameters studied in each group of experiments, the rest of the network parameters, training settings and datasets used are the same, and the performance evaluation indicators are AP, AP_{50}, and AP_{75}.

1) ε in the GAM: We take $\varepsilon \in [0{:}0.005{:}0.02]$ to carry out comparative experiments. The ε of 0 means that all locations of the feature map are not filtered and set anchors. As shown in Table III, the highest AP and AP_{75} are obtained by setting ε as 0, and AP and AP_{75} are at least 0.8% and 0.3% higher than other ε, but AP_{50} is 0.5% lower than the highest value. The highest AP_{50} was obtained at ε of 0.01, which is at least 0.2% higher than other ε, but AP and AP_{75} are 0.8% and 0.3% lower than the highest values, respectively. In the target detection task, when the IoU of the bounding box and the GT box is higher than 0.5, the prediction of the bounding box is considered as correct. The AP_{50} commonly used in the evaluation of Pascal VOC is also calculated when the IoU threshold is 0.5. Although the highest AP and AP_{75} are obtained when ε of 0, the lower AP_{50} indicates that fewer ships are detected. In addition, ε of 0 will generate a large number of redundant negative samples and increase the computational cost, so we choose 0.01 as ε in the GAM.

2) σ_1 and σ_2 in the GAM: Since the docking direction of ships is arbitrary, the proportion of ships in the manually annotated horizontal bounding box is random. To obtain the hard negative samples, we set σ_2 to 0.5 or 0.6 for hyperparameter experiments [35], [44]. We set $\sigma_1 \in [0.1{:}0.1{:}\sigma_2\text{-}0.1]$, and the experimental results are shown in the Table IV. The highest AP_{50} and AP_{75} are obtained by setting $\sigma_1 = 0.2$ and $\sigma_2 = 0.5$, and the AP_{50} and AP_{75} are at least 0.1% and 0.4% higher than other combinations of σ_1 and σ_2, but the AP is 0.4% lower than the highest value. The highest AP and AP_{50} of 95.1% are obtained when $\sigma_1 = 0.3$ and $\sigma_2 = 0.6$. The AP_{50} is 0.6% lower than the highest value, which means that fewer ships are detected than $\sigma_1 = 0.2$ and $\sigma_2 = 0.5$. Therefore, the σ_1 and σ_2 in our method are set as 0.2 and 0.5, respectively.

TABLE V
EXPERIMENTAL RESULTS OF HYPERPARAMETERS FOR α IN THE RFEM(%)

α	AP	AP_{50}	AP_{75}
2	**64.8**	95.7	**76.6**
3	64.4	**95.8**	76.3
4	64.4	95.5	74.2
5	64.2	95.6	74.1

TABLE VI
PERFORMANCE OF ASPECT RATIO LOSS OF $\varepsilon = 0.01$ IN THE GAM(%)

	AP	AP_{50}	AP_{75}
×	62.8	94.8	73.5
✓	64.3	94.7	75.1

3) α in the RFEM: In the SAR ship detection, small ships make up a very small proportion of the SAR image, so we choose $\alpha \in [2, 3, 4, 5]$ to carry out comparative experiments. In the Table V, we can observe that the highest AP and AP_{75} are obtained at α of 2, while AP_{50} is only 0.1% lower than that at α of 3. The network gets the highest AP_{50} when α is set to 3, however, its AP and AP_{75} are 0.4% and 0.3% lower compared with α of 2, respectively. From the Table V, with the increase of α, AP and AP_{75} show a decreasing trend. To sum up, setting α as 2 can get more accurate bounding boxes, and can also detect more ships, so we set the α in the RFEM as 2.

D. Ablation Experiment

In order to fairly verify the effectiveness of the two components in our method, we conduct ablation experiments under the same experimental setup and data configuration. The baseline is faster R-CNN, and five indicators are used in this experiment. The AP, AP_{50}, and AP_{75} are used to evaluate the improvement of the component on the overall dataset. The AP_{s} and $AP_{50_{\text{s}}}$ can verify the improvement of the component's detection ability for small ships.

1) Effect of GAM: We first investigate the effect of adding the aspect ratio loss in loss function of the GAM. The results are shown in Table VI. After adding the aspect ratio loss, the AP and AP_{75} gain 1.5% and 1.6% improvement, respectively, and the AP_{50} decreases by 0.1%. These indicate that the network benefited from the aspect ratio loss predicts more accurate bounding box. This means that the aspect ratio loss supervises the GAM to generate higher quality anchors. We then evaluate the performance of adding GAM to the baseline. The Fig. 5 shows the distribution of proposals predicted on the input SAR image by the RPN of baseline with or without the GAM, respectively. We can observe that after inserting the GAM, the overall number of proposals is greatly reduced compared

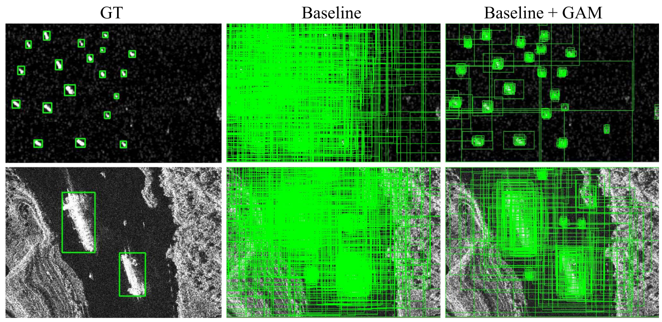


Fig. 5. Visualization results of proposals from the RPN.

TABLE VII
INFLUENCE OF EACH COMPONENT IN THE PROPOSED METHOD(%)

GAM	RFEM	AP	AP_{50}	AP_{75}	AP_s	AP_{50_s}
×	×	61.3	91.5	73.2	61.2	91
×	✓	62.7	92.4	74.1	62	92
✓	×	64.3	94.7	75.1	63.8	93.9
✓	✓	**64.8**	**95.7**	**76.6**	**64.5**	**95.1**

to the baseline, and most proposals are concentrated on the ships. The distribution of proposals in marine area is sparse, and the shape of proposals on the ships is closer to GT boxes. If without GAM, too many redundant proposals will be fed into the head of network resulting in extra computation. The results in Table VII show that the GAM increases AP, AP_{50}, and AP_{75} by 3.0%, 3.2% and 1.9%, respectively. This means that the high-quality anchors generated by the GAM can adapt to longer or wider ships, allowing the network head to regress more refined bounding boxes. The GAM improves AP_s by 2.6% and AP_{50_s} by 2.9% on the baseline, because the GAM can predict anchors more matching than the preset anchors on position and shape for small ships, as shown in the top row of Fig. 5.

2) Effect of RFEM: We explore the impact of the RFEM by adding it to the baseline. From the Table VII, the RFEM achieves improvement of AP, AP_{50}, and AP_{75} by 1.4%, 0.9%, and 0.9%, respectively. The RFEM expands the receptive field of each element in the feature map, so the information around the ships is collected and assists the network to detect the ships. The top row of Fig. 7 is a near-shore scenario, the baseline produces false positives in the area close to the shore. The RFEM uses information from the ocean and coastal areas around the ships to eliminate false positives and facilitates the network to predict more accurate bounding boxes. The RFEM increases AP_s by 0.8% and AP_{50_s} by 1% on the baseline, which shows that the RFEM improves the sensitivity of the network to small ships. The bottom row of Fig. 7 is a scene that small ships park densely. After the RFEM is inserted into baseline, the semantic information of the feature map becomes more abundant, which reduces the number of missed ships of the baseline by three, so the RFEM can further reduce the missing rate for small ships.

The above content verifies and analyzes the effectiveness of the two components, the GAM and the RFEM, respectively. As shown in Table VII, all metrics used in combination with the two components are further improved compared to using the components alone. Compared with the baseline, our method gains improvement of AP, AP_{50}, and AP_{75} by 3.5%, 4.2%, and 3.4%, respectively. The corresponding PRC in Fig. 6 displays more comprehensive results. These mean our method can detect more ships and predict more accurate bounding box of ships. In addition, the AP_s and AP_{50_s} increasing by 3.3% and 4.1% are obtained, which means that our components greatly improves the detection ability of small ships.

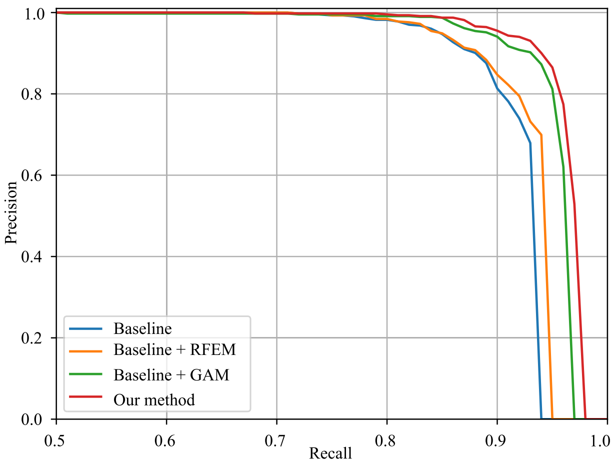


Fig. 6. PRC of different improvements.

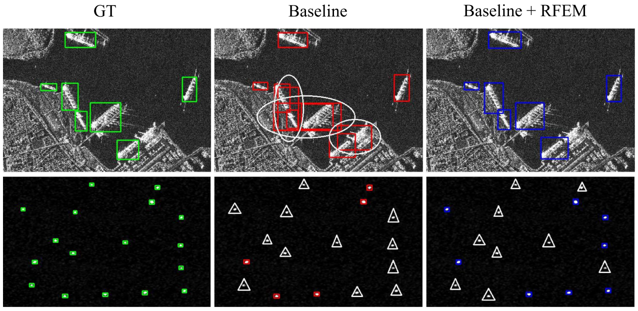


Fig. 7. Visualization results of the RFEM. The rectangles are the detection results. The triangles and circles represent the missing ships and the false alarm, respectively.

E. Comparison and Discussion

In this experiment, we fairly compare the performance of our method with some CNN-based methods, and the experimental settings and data configuration of the comparison experiments are exactly the same. The CNN-based methods we compare are divided into anchor-based methods and anchor-free methods. The anchor-based methods are RetinaNet [15], faster R-CNN [19], cascade R-CNN [45], Libra R-CNN [46], GA R-CNN [35], and HR-SDNet [41], the last five are developed based on the two-stage network framework like our method. The anchor-free methods selects FCOS [23], CP-FCOS [44], Autoassign [47], and ATSS [48]. The last two anchor-free methods optimize the strategy of sample assignment. The HR-SDNet and CP-FCOS are specifically designed for ship detection in SAR images.

In order to compare the performance of each method more comprehensively, we add AP, AP_{50}, and AP_{75} of inshore ships

TABLE VIII
COMPARISON WITH THE TYPICAL DETECTION METHODS ON SSDD(%)

Method		Entire						Inshore			Offshore		
		$Recall$	AP	AP_{50}	AP_{75}	AP_s	AP_{50_s}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Anchor-based	RetinaNet	93	44.4	77.8	47.8	44.6	77.9	24.7	51.8	23.3	54.1	89.3	60.5
	Faster R-CNN	93.6	61.3	91.5	73.2	61.2	91	44.1	76.8	46	68.6	97.6	85
	Cascade R-CNN	93.8	65.6	92.5	77.8	64.9	91.6	51.1	79	57.5	72.1	97.9	87.9
	Libra R-CNN	94	62.1	91.3	73.9	62.1	91.6	43.7	75.7	46.3	69.6	97.6	85.7
	GA R-CNN	96.2	60.3	94	69.9	59.9	93.6	44.9	79.8	45.7	66.4	97.8	80.5
	HR-SDNet	93.6	**67.5**	92.6	**79.8**	**66.6**	91.7	**52.8**	79.9	**58.2**	**73.9**	97.9	**89.3**
Anchor-free	FCOS	93	57.1	89.1	66.8	57.9	89.7	34.7	65.4	32.2	64.5	97.9	78.6
	CP-FCOS	92.5	57.3	89.3	65.4	57.5	89.1	37.7	71.3	34.8	66.1	97	80
	Autoassign	96	59.3	92.2	69	60.9	92.4	41.2	73.8	39.2	66.8	98.8	81.6
	ATSS	94.3	60.6	91.3	69.4	61.1	91.9	40.3	72.4	40	69	98.1	81.8
	Our method	**97.1**	64.8	**95.7**	77.6	64.5	**95.1**	52.5	**83.6**	57.1	69.9	**99.7**	84.9

TABLE IX
COMPARISON WITH THE TYPICAL DETECTION METHODS ON HRSID(%)

Method		Entire						Inshore			Offshore		
		$Recall$	AP	AP_{50}	AP_{75}	AP_s	AP_{50_s}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Anchor-based	RetinaNet	88.9	56.3	81.8	62.7	57.4	81.5	37	63.6	37.2	77.5	98.2	90.9
	Faster R-CNN	81.9	60	79.6	63.8	60.5	78.3	49.5	68.4	55.8	83.3	97	94.7
	Cascade R-CNN	84.4	64	82.4	72.9	64.4	81.1	55.3	72.9	61.6	88.1	98	95.8
	Libra R-CNN	83	60.9	80.4	69.9	61.6	79.3	49	66.7	55.7	84.2	97	94.7
	GA R-CNN	94.5	65.5	90.7	74.9	66.7	91.2	59.7	87.3	67.3	83.2	**98.9**	94.7
	HR-SDNet	86.2	65.2	83	74.5	65.6	83	55.8	73.6	63.1	**88.5**	98	**95.9**
Anchor-free	FCOS	87.9	53.7	82.7	59.4	55.1	82.7	37.7	69.4	36	75.6	97.9	88.8
	CP-FCOS	88.1	55.6	82.9	62.3	56.9	82.8	49.3	74	52.1	81.5	97.9	90.5
	Autoassign	90.5	59.8	86.1	66.7	61.7	86	46.5	77.1	48.1	80.2	97.9	92.8
	ATSS	89.5	60.1	84.6	65.7	61.4	84.7	46	73.8	47.9	81.7	97.9	93.1
	Our method	**95.2**	**66.5**	**91.1**	**76.2**	**68.4**	**92**	**60.9**	**87.7**	**69.1**	83.1	**98.9**	95.6

TABLE X
COMPARISON WITH THE TYPICAL DETECTION METHODS ON MODEL SIZE AND RUNNING TIME

Method		Parameters (M)	FPS (SSDD)	FPS (HRSID)
Anchor-based	RetinaNet	36.1	25.8	22.5
	Faster R-CNN	43.24	28.9	24.9
	Cascade R-CNN	68.93	24.1	21.2
	Libra R-CNN	41.71	28.6	24.7
	GA R-CNN	41.71	22.8	20.2
	HR-SDNet	77.16	7.7	7.4
Anchor-free	FCOS	**31.84**	**29.7**	**26.5**
	CP-FCOS	35.97	26.0	23.1
	Autoassign	35.97	26.1	22.7
	ATSS	31.89	26.2	22.7
	Our method	43.42	20.0	16.4

and offshore ships on the basis of the five indicators in ablation experiment. In addition, to verify the generalization ability of our method, comparative experiments are carried out on SSDD and HRSID datasets, respectively. The experimental results are shown in Table VIII–X. Our method achieves certain advantages on AP_{50} of the entire, inshore, and offshore ships on both datasets. The recall and AP_{50_s} of our method are the highest on both datasets compared with the other methods. These mean that our method can detect more ships in complex scenes and is more capable of detecting small ships than these CNN-based methods. Although both RetinaNet and our method can solve the problem of imbalance between positive and negative samples, the RetinaNet is inferior to our method on all metrics due to the lack of high-quality anchors. Compared with these anchor-based and anchor-free networks, our method also has advantages on AP and AP_{75}, indicating that the receptive field improvement by the RFEM and high-quality anchors generated by the GAM in our method can help the network regress more accurate bounding box of ships. It is worth noting that some AP and AP_{75} of our method on SSDD are slightly lower than cascade R-CNN. That is due to the lack of multilevel refinement of bounding boxes in our method, compared to the cascade R-CNN with more parameters and longer inferential time. The cascade R-CNN can obtain more accurate bounding boxes of easily detectable ships, however, our method is much higher than cascade R-CNN on AP_{50} of both two datasets. And our method obtains better performance of AP and AP_{75} on entire and inshore samples in HRSID with larger number of samples, so our method is more practical for ship detection in SAR images.

As shown in Figs. 8 and 9, we visualize the detection results of some methods in four types of complex scene. We selected four representative methods to more highlight the strengths and weaknesses of our method. The selected methods are ATSS, faster R-CNN, and cascade R-CNN. The ATSS with the highest AP among the anchor-free methods in Tables VIII and IX. Comparing with the baseline network (faster R-CNN) of our method can fully show the improvement effect of our designed GAM and REFM on ship detection. The multistage detection network cascade R-CNN has more parameters and it uses the faster R-CNN as the baseline network like our method. Comparing with it can show that our method improves detection performance without increasing the number of parameters too much. We can observe the small ships with dense distribution, as shown in the first row of Fig. 8. The ATSS detects all targets but gives false alarms. Both faster R-CNN and cascade R-CNN have

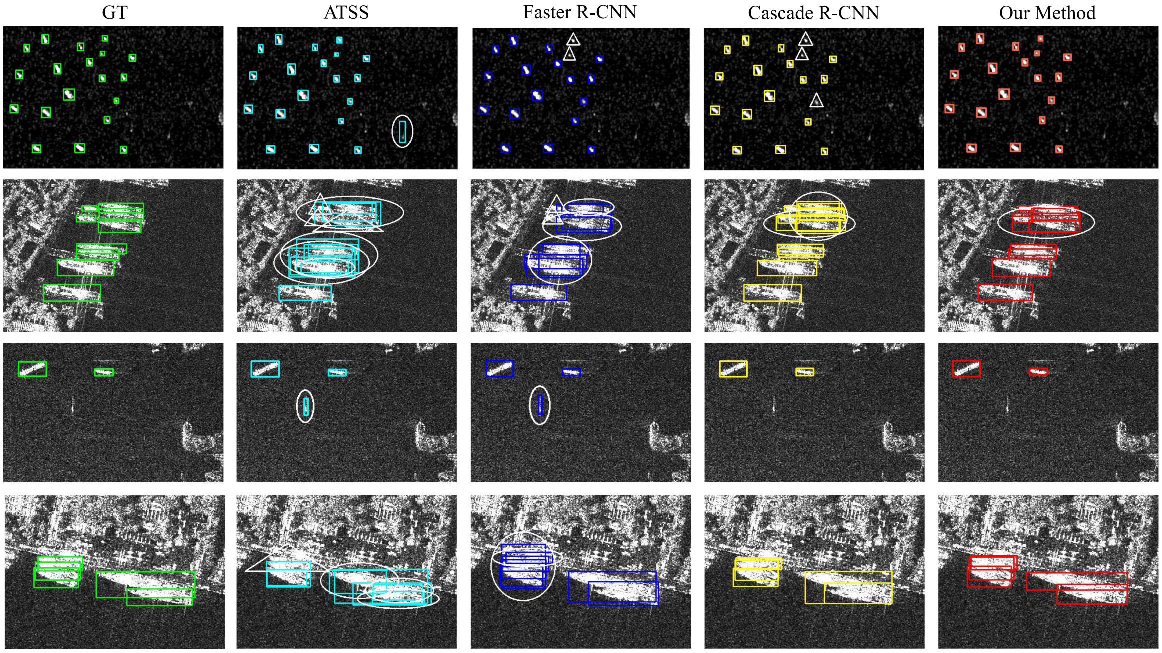


Fig. 8. Visualization results of four different methods in complex scenarios of the SSDD. The rectangles are the detection results. The triangles and circles represent the missing ships and the false alarm, respectively.

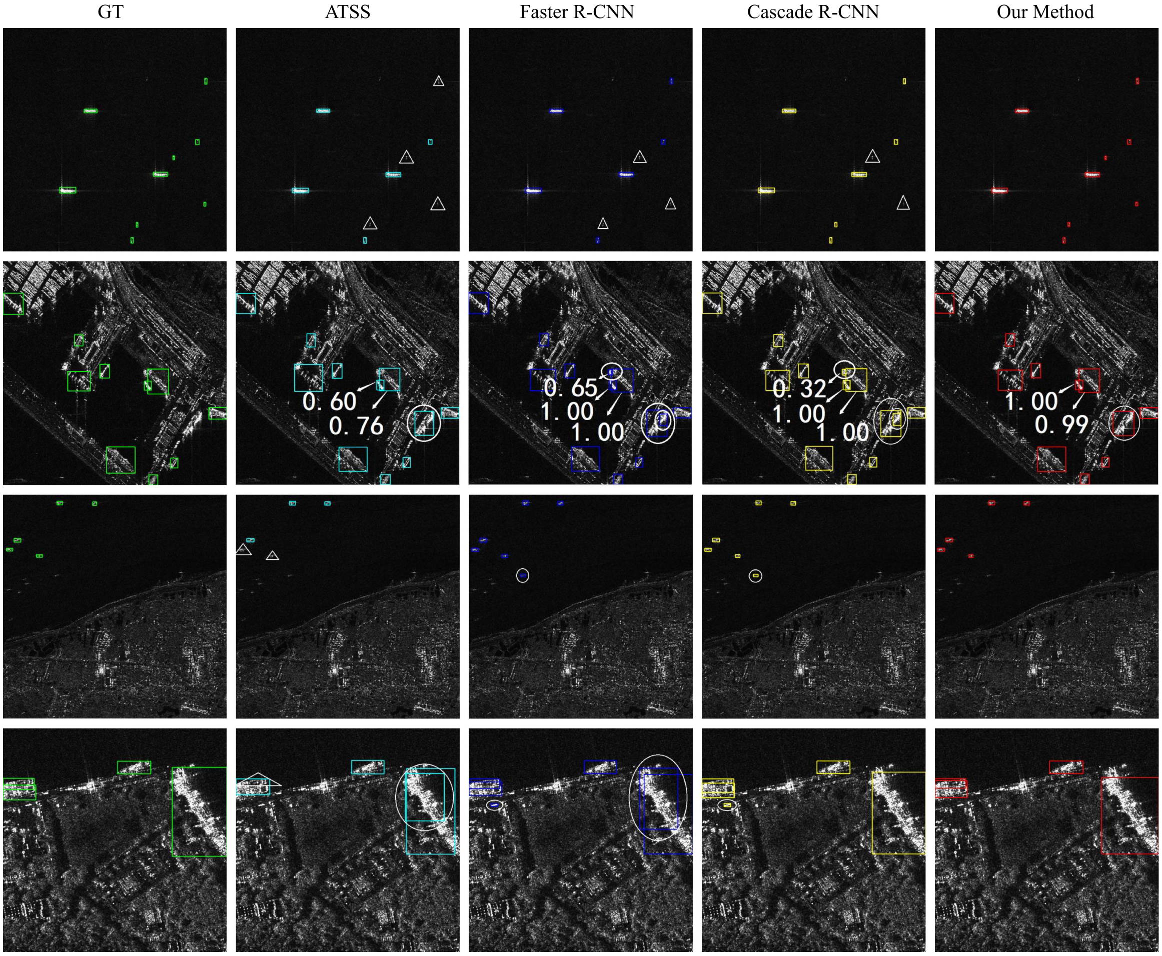


Fig. 9. Visualization results of four different methods in complex scenarios of the HRSID. The rectangles are the detection results, and the number next to the rectangle is the confidence of the corresponding predicted ship. The triangles and circles represent the missing ships and the false alarm, respectively.

missing ships, and our method detects all small ships without false alarms simultaneously. In the same scenario in the first row of Fig. 9, our method detects all small ships, while the other three methods produce missing ships. This means that our method has stronger detection ability for small targets, which is consistent with the results in Tables VIII and IX. In the scene where ships are parked adjacently (the second row in Figs. 8 and 9), both ATSS and baseline have missing ships and false alarms on the SSDD. And on the HRSID, our method detects adjacent ships targets with confidence close to one and no false alarms. In the scene where the ships appears near the island (the third row in Figs. 8 and 9), our method overcomes the interference caused by the island and detects all ships accurately. The ATSS and faster R-CNN regard the small island as a ship on the SSDD. And on the HRSID, the ATSS misses small ships near the small islands, while faster R-CNN and cascade R-CNN mispredict the island as a ship. The detection of inshore ships is very prone to false alarms and missing alarms. The fourth row of Figs. 8 and 9 display that our method detects all ships and predicts the bounding box of ships most accurately without false alarm compared to the other three CNN-based methods. In general, our method can achieve better detection effects for small targets. In complex scenes, our method can detect more ships and can predict more accurate bounding boxes of ships without adding stages of refinement boxes.

IV. Conclusion

In this article, we propose a two-stage network that can generate anchors by the network for detecting small ships in SAR image. We introduce the GAM and purposely add aspect ratio loss to its loss function for capturing ships. The redesigned GAM can generate higher quality anchors, which is more conducive to regress bounding box of ships. In addition, we propose a RFEM and embed it into FPN. The RFEM sets atrous convolutions with different dilation rates for feature maps of different resolutions, which expands the receptive field of elements in the feature map and enriches their semantic information. The information about the region around the ships is collected to help the network improve the accuracy of the ship's location. The experimental results show the effectiveness of our designed components. And compared with some CNN-based methods, our method can detect more ships, and the detection ability for small ships of our method is stronger than the state-of-the-art networks, which show the superiority of our method.

References

[1] C. Wang, S. Jiang, H. Zhang, F. Wu, and B. Zhang, "Ship detection for high-resolution SAR images based on feature analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 119–123, Jan. 2014.

[2] 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.

[3] T. Liu, J. Zhang, G. Gao, J. Yang, and A. Marino, "CFAR ship detection in polarimetric synthetic aperture radar images based on whitening filter," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 58–81, Jan. 2020.

[4] S.-I. Hwang and K. Ouchi, "On a novel approach using MLCC and CFAR for the improvement of ship detection by synthetic aperture radar," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 391–395, Apr. 2010.

[5] X. Leng, K. Ji, K. Yang, and H. Zou, "A bilateral CFAR algorithm for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 7, pp. 1536–1540, Jul. 2015.

[6] C. Wang, F. Bi, W. Zhang, and L. Chen, "An intensity-space domain CFAR method for ship detection in HR SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 529–533, Apr. 2017.

[7] X. Leng, K. Ji, S. Zhou, and X. Xing, "Ship detection based on complex signal kurtosis in single-channel SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6447–6461, Sep. 2019.

[8] X. Wang, G. Li, X.-P. Zhang, and Y. He, "Ship detection in SAR images via local contrast of fisher vectors," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6467–6479, Sep. 2020.

[9] Z. Cui, X. Wang, N. Liu, Z. Cao, and J. Yang, "Ship detection in large-scale SAR images via spatial shuffle-group enhance attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 379–391, Jan. 2021.

[10] G. Gao, "An improved scheme for target discrimination in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 277–294, Jan. 2011.

[11] Z. Wang, L. Du, and H. Su, "Target detection via Bayesian-morphological saliency in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5455–5466, Oct. 2017.

[12] J. Ai, Y. Mao, Q. Luo, L. Jia, and M. Xing, "SAR target classification using the multikernel-size feature fusion-based convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

[13] Z. Ren, B. Hou, Z. Wen, and L. Jiao, "Patch-sorted deep feature learning for high resolution SAR image classification," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 11, no. 9, pp. 3113–3126, Sep. 2018.

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

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

[16] W. Liu et al., "SSD: Single Shot Multibox Detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[17] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[18] S. Gao, J. Liu, Y. Miao, and Z. He, "A high-effective implementation of ship detector for SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art no. 401900.

[19] 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.

[20] B. Hou, Z. Ren, W. Zhao, Q. Wu, and L. Jiao, "Object detection in high-resolution panchromatic images using deep models and spatial template matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 956–970, Feb. 2020.

[21] T. Zhang et al., "Balance learning for ship detection from synthetic aperture radar remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 182, pp. 190–207, 2021.

[22] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.

[23] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2020, pp. 9626–9635.

[24] Q. Hu, S. Hu, and S. Liu, "BANet: A balance attention network for anchor-free ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, Art no. 5222212.

[25] F. Gao, Y. He, J. Wang, A. Hussain, and H. Zhou, "Anchor-free convolutional network with dense attention feature aggregation for ship detection in SAR images," *Remote Sens.*, vol. 12, no. 16, 2020, Art. no. 2619.

[26] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1331–1344, Feb. 2020.

[27] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 3–22, 2018.

[28] 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.

[29] Z. Wang, B. Wang, and N. Xu, "SAR ship detection in complex background based on multi-feature fusion and non-local channel attention mechanism," *Int. J. Remote Sens.*, vol. 42, no. 19, pp. 7519–7550, 2021.

[30] X. Yang, X. Zhang, N. Wang, and X. Gao, “A robust one-stage detector for multiscale ship detection with complex background in massive SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art no. 5217712.

[31] P. Chen, H. Zhou, Y. Li, B. Liu, and P. Liu, “Shape similarity intersection-over-union loss hybrid model for detection of synthetic aperture radar small ship objects in complex scenes,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9518–9529, 2021.

[32] N. Su, J. He, Y. Yan, C. Zhao, and X. Xing, “SII-Net: Spatial information integration network for small target detection in SAR images,” *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 442.

[33] S. Wei, X. Zeng, H. Zhang, Z. Zhou, J. Shi, and X. Zhang, “LFG-Net: Low-level feature guided network for precise ship instance segmentation in SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022, Art no. 5231017.

[34] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, *arXiv:1804.02767*.

[35] J. Wang, K. Chen, S. Yang, C. L. Chen, and D. Lin, “Region proposal by guided anchoring,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2960–2969.

[36] T. Zhang et al., “SAR ship detection dataset (SSDD): Official release and comprehensive data analysis,” *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3690.

[37] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, “HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation,” *IEEE Access*, vol. 8, pp. 120234–120254, 2020.

[38] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[39] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, *arXiv:1706.05587*.

[40] K. Chen et al., “MMDetection: Open MMLab detection toolbox and benchmark,” 2019, *arXiv:1906.07155*.

[41] S. Wei, W. Ming, and Zhang, “Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet,” *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 167.

[42] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[43] T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

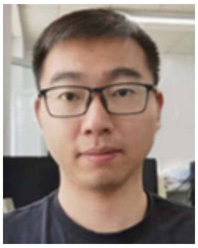
[44] Z. Sun et al., “An anchor-free detection method for ship targets in high-resolution SAR images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7799–7816, 2021.

[45] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.

[46] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, “Libra R-CNN: Towards balanced learning for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 821–830.

[47] B. Zhu et al., “AutoAssign: Differentiable label assignment for dense object detection,” 2020, *arXiv:2007.03496*.

[48] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, “Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.



Tingxuan Yue received the B.S. degree from the China University of Mining and Technology, Xuzhou, China in 2018. He is currently working toward the Ph.D. degree with the School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China.

His research interests include computer vision, deep learning, and image processing in remote sensing.

Yanmei Zhang received the B.S. degree in electromechanical engineering, the M.S. degree in mechanical and electrical automation, and the Ph.D. in electromechanical engineering from the Beijing Institute of Technology, Beijing, China, in 1989, 1995, and 2010, respectively.

She is currently with the School of Integrated Circuits and Electronics, Beijing Institute of Technology, as a Professor and a Ph.D. Supervisor. Her research interests include signal detection, signal processing technology, photoelectric detection, and imaging processing technology.

Pengyun Liu received the B.S. degree from Qingdao University of Technology, Qingdao, China, in 2020. He is currently working toward the M.S. degree with the School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China.

His research interests include high-speed signal processing, photoelectric detection, and image processing in remote sensing.

Yanbing Xu received the B.S. degree from the Hunan University, Changsha, China in 2017. He is currently working toward the Ph.D. degree with the School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China.

His research interests include computer vision, deep learning, and high-dimensional image processing in remote sensing.

Chengcheng Yu received the B.S. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2016. He is currently working toward the Ph.D. degree with the School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China.

His research interests include pattern recognition and image processing in remote sensing.