IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 20, 2019, accepted October 29, 2019, date of publication November 4, 2019, date of current version November 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2951030

MSARN: A Deep Neural Network Based on an Adaptive Recalibration Mechanism for Multiscale and Arbitrary-Oriented SAR Ship Detection

CHEN CHEN[1], CHUAN HE[1,2], CHANGHUA HU[1], HONG PEI[1], AND LICHENG JIAO[2], (Fellow, IEEE)
[1]Xi'an Institute of High-Technology, Xi'an 710025, China
[2]School of Artificial Intelligence, Xidian University, Xi'an 710071, China

Corresponding author: Chuan He (hechuan8512@163.com)

This work was supported by the National Natural Science Foundation of China under Grant 61773389, Grant 61833016, and Grant 61573365.

ABSTRACT Ship detection plays an important role in synthetic aperture radar (SAR) image interpretation. However, there are still some difficulties in SAR ship detection. First, ships often have a large aspect ratio and arbitrary directionality in SAR images. Traditional detection algorithms can cause the detection area to be redundant, which makes it difficult to accurately locate the target in complex scenes. Second, ships in ports are often densely arranged, and the effective identification of densely arranged ships is complicated. Finally, ships in SAR images exist at a variety of scales due to the multiresolution imaging modes used and ship shape variations, which pose a considerable challenge for ship detection. To solve the above problems, we propose a multiscale adaptive recalibration network (MSARN) to detect multiscale and arbitrarily oriented ships in complex scenarios. The recalibration of the extracted multiscale features through global information increases the sensitivity of the network to the target angle, thereby increasing the accuracy of positioning. In particular, we designed a pyramid anchor and a loss function to match the rotated target. In addition, we modified the rotation non-maximum suppression (RNMS) method to solve the problem of the large overlap ratio of the detection box. The proposed model combines the positioning advantage of rotation detection with the speed advantage of a single-stage framework. Experiments show that based on the SAR rotation ship detection (SRSD) data set, the proposed algorithm has a faster detection speed and higher accuracy than some state-of-the-art methods.

INDEX TERMS Ship detection, synthetic aperture radar (SAR), adaptive recalibration, neural network.

I. INTRODUCTION

Synthetic aperture radar (SAR) has been widely used in civil remote sensing surveying and military reconnaissance due to its independence from solar illumination and ability to provide images in all-weather operating conditions [1]–[4]. In recent years, SARs such as TerraSAR-X, RADARSAT-2, and Sentinel-1 have rapidly developed, which has greatly promoted research on SAR image ship detection methods [5]–[7]. The constant false alarm rate (CFAR) and its various derivative algorithms are widely used in SAR image ship detection [8]–[10]. This type of algorithm is based on a statistical model of the contrast information, which can automatically adjust thresholds to suit different ocean backgrounds while maintaining the required performance. However, the algorithm modelling process is complicated, and ship detection in complex backgrounds cannot achieve the desired effect.

A deep convolutional neural network (DCNN) can automatically learn the structural features of a target and have been rapidly developed in the field of computer vision. A series of object detection algorithms based on DCNN was proposed. The detection algorithms based on DCNN mainly include two types: two-stage detection algorithms, such as the

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Di Martino.

Faster R-CNN [11], and single-stage detection algorithms, such as SSD [12], YOLO [13]–[15], and RFBNet [16]. The two-stage detection algorithms have high accuracy, and the single-stage detection algorithms have notable speed advantages. These methods have achieved good results for general detection data sets such as Pascal VOC [17] and COCO [18]. Researchers have made improvements to these detection methods and proposed some algorithms for SAR object detection. Kang *et al.* proposed a contextual region-based CNN with multilayer fusion for SAR ship detection, which is composed of a region proposal network (RPN) with high network resolution and an object detection network with contextual features. The network achieves an excellent performance on the Sentinel-1 data set [19]. Jiao *et al*. proposed a densely connected multiscale neural network based on the Faster R-CNN framework. The method utilizes a densely connected network as its framework to detect ship targets [20]. Li *et al*. established the SAR Ship Detection Data set (SSDD) and improved Faster R-CNN for SAR ship detection [21]. Kang *et al*. combined the traditional CAFR algorithm with Faster R-CNN [22]. Deng *et al*. proposed an effective approach to learn deep ship detector from scratch that can effectively solve the existing problems in ship detection. At the same time, the backbone network could be freely designed and effectively trained from scratch without using a large number of annotated samples [23]. Lin *et al.* innovatively integrated the squeeze-and-excitation mechanism into Faster_R-CNN and established a new network structure, which achieved a better performance than state-of-the-art methods at a SAR ship detection task [24]. Wang *et al.* proposed a spatial-spectral squeeze-and-excitation (SSSE) model that adaptively learns the weights of different spectral bands and different adjacent pixels. The advantage of the algorithm is that it can compress or activate features arbitrarily. In this manner, the noise is suppressed, and the performance for hyperspectral image classification is improved [25]. Zhang *et al.* applied a CNN to the task of high-resolution remote sensing image detection and proposed a ship detection method based on an improved version of Faster R-CNN. The proposed method has good positioning effects for small targets and ships with dense arrangements [26]. Chen *et al*. proposed an SAR ship detection network that integrates an attention mechanism, and this approach achieved satisfactory accuracy and speed performance [27]. However, there are still some obstacles in SAR ship detection. First, the above detection algorithms use a horizontal bounding box to locate the target. However, SAR uses a high-altitude overhead imaging mode in which the ship target has a large length-width ratio and arbitrary orientation. The traditional horizontal bounding box cannot reflect the true shape of the target. In addition, the horizontal bounding box introduces a large amount of background interference, which is not conducive to precise positioning of the target. Second, ships in ports are generally densely arranged, and the traditional detection algorithms cannot effectively distinguish among the densely arranged ship targets. When the horizontal bounding box is applied to densely arranged targets, the bounding boxes of different targets will have a large overlapping ratio, resulting in target misdetection. Third, SAR ship targets exhibit a broad diversity of scales due to the multiresolution imaging modes and the variety of ship shapes, making them difficult for existing algorithms to effectively detect and locate, especially small-scale ship targets. Finally, the above algorithms are based on a two-stage detection framework, and the resulting image processing efficiency is low.

Applications of object detection algorithms with arbitrary orientations begin with scene text detection. Similar to ship targets, scene text has a large aspect ratio and an arbitrary direction. The traditional horizontal detection algorithm is not applicable in these cases, so some detection algorithms for rotated targets have been proposed. Ma *et al*. proposed the rotation region proposal network (RRPN), which generates a candidate region with angle information text through the RRPN; then, through the rotation region-of-interest (RRoI) pooling layer, the proposal box is mapped to the feature map to obtain the final detection results [28]. Jiang *et al*. proposed the rotational region CNN (R^2CNN) to detect inclined text [29]. Unlike the RRPN, the R^2CNN uses multiscale RRoI pooling to accommodate large aspect ratios and retains a horizontal prediction box for effective detection. Inspired by scene text detection methods, researchers transferred these methods to remote sensing ship detection tasks. Liu *et al*. introduced the R^2CNN to high-resolution remote sensing image ship detection and achieved good results [30]. Li *et al*. proposed a full CNN based on rotated regions and applied it to high-resolution optical remote sensing images [31]. Yang *et al*. combined the rotation characteristics of targets with the dense feature pyramid networks (DFPNs) and proposed the rotational DFPN (R-DFPN), which achieved state-of-the-art performance in ship detection for remote sensing images from Google Earth [32].

The above method achieved good results in optical remote sensing images detection tasks, but SAR image detection does not yield the same effect, largely due to the essential difference between SAR images and optical remotely sensed images. First, an SAR image reflects the characteristics of the scattering of electromagnetic waves from the object. The optical image contains abundant visual scene information, whereas the SAR images are of low resolution, have a low signal-to-noise ratio and contain relatively monotonous information. Second, since SAR images are obtained in the forward direction from a lateral view, they are easily affected by the terrain, resulting in inverted top-bottom image distortion. Finally, there are a significant variety of scales for ships in SAR images due to the diversity of multiresolution imaging modes and ship shapes. In addition, the SAR images have strong speckled noise, which causes interference for object recognition and detection. In terms of the image processing efficiency, the above method is based on Faster R-CNN [11] and the two-stage detection framework with rotational region proposal and RRoI pooling for detecting rotated targets. However, the region proposal method additionally increases the

computational burden, resulting in a long algorithm run time, and the real-time detection of targets cannot be performed.

In the construction of CNN-based SAR image ship detection networks, it is necessary to fully consider the difference between optical and SAR images and design the CNN model in a targeted manner [33]. However, there are still few studies regarding efficient detection models for rotated SAR ship targets. In view of the existing problems related to current SAR ship detection tasks, this paper combines the precise positioning advantage of the rotated detection method with the speed advantage of the single-stage network and proposes a new network, the multiscale adaptive recalibration network (MSARN), for multiscale and arbitrarily oriented ship detection in complex scenarios. The network provides an accurate positioning capability for SAR ship detection and greatly reduces the time overhead. The main contributions of this paper are as follows.

1) A detection model for multiscale and arbitrarily oriented ship targets in SAR images is established. Unlike previous detection models, the newly proposed model combines the precise positioning advantages of rotation detection with the speed advantage of a single-stage framework. Compared with the models in reference [11], [15], [16], and [27], the proposed model improves the detection performance in different complex scenes, can effectively distinguish among densely arranged targets, and reduces redundant detection areas.
2) A multiscale adaptive recalibration module is proposed to calibrate the features extracted by the CNN through global information to improve the sensitivity of the network to changes in angles. At the same time, the module is lightweight and can serve as the basic unit of the established MSARN. In the established MSARN, the features of different levels are merged, which makes the merged network layer have both a robust feature representation capability and an accurate positioning capability.
3) Rotated anchor boxes and a loss function are designed to match targets of different scales. The rotation nonmaximum suppression (RNMS) algorithm is modified, which improves the detection of densely arranged ship targets. A modified labelling method is introduced to replace the method of labelling four consecutive points used in our previous paper, which reduces the error rate of manual labelling.
4) The model is based on a single-stage object detection framework, and near real-time detection can be achieved at a sufficient speed with satisfactory detection results. Compared with the rotation detection models in reference [28], [29], and [32], the proposed model has an absolute speed advantage.

The remainder of this paper is organized as follows. Section II illustrates the method and network structure proposed. Section III introduces the data sets used in the experiments and describes the experimental details and results. Section IV discusses the possibilities for future work. Section V presents the conclusions.

II. METHODS

This paper proposes an SAR image ship detection model based on an adaptive recalibration mechanism. The main flow of the model is as follows. First, after the original image is pre-processed, it is used as the input to the MSARN. Second, multilevel target mapping features are obtained by the network. Third, the mapping features are recalibrated with global information; then, the features expressed at different depths are fused via a feature fusion method. Based on the fused feature maps, the locations and confidence scores of the targets are predicted. Finally, the redundant predicted boxes are filtered via the modified RNMS, and the final detection results are obtained.

A. MULTISCALE ADAPTIVE RECALIBRATION MODULE

In computer vision, long-range dependencies can be captured to extract global information from visual scenes, and this approach can improve a wide range of recognition tasks, such as image classification, object detection and segmentation [34]–[36]. However, a CNN can establish pixel relationships only in local neighbourhoods [35]. Therefore, in a CNN, a large receptive field is usually formed by stacking a number of convolution modules to capture long-range dependencies. However, this approach is inefficient for capturing the dependencies, and as the network deepens, the transfer of information becomes difficult [37]. Inspired by the classical non-local means algorithm in computer vision, reference [35] proposed non-local neural networks (NNNs) to capture long-range dependencies. Compared to convolutional stacking, NNNs can more effectively capture the dependencies between pixels over a long distance to obtain rich global information. The core concept of NNNs is that when calculating the response for a certain position in an image, not only the neighbourhood information but rather all the location information in the image is aggregated to enhance the feature response for the current location. Inspired by the structure of the receptive fields (RFs) of the human visual system, reference [14] simulated the RFs by constructing dilated convolutions, thereby improving the ability of the network to recognize targets. Based on the structure of the visual RF, combined with the global concept of NNNs [35], a multiscale adaptive recalibration module is designed. The core concept of the module is that based on the fusion of the multiscale features extracted by the convolutional network, the features are recalibrated with the global information; thus, the network layer has both a robust feature representation ability and an accurate positioning capability.

The multiscale adaptive calibration module proposed mainly composed of three parts: multibranch convolution, global modelling, and bottleneck transform. The multibranch convolution structure allows ship features at different scales to be activated on different branches. The multibranch convolution structure is similar to that of Inception-ResNet [38],

but a dilated convolution [16] is added in the convolution of each branch to obtain a larger RF. Large RFs can include a wide range of information that is conducive to the separation of ship targets and complex backgrounds. In the multibranch convolution structure, 1∗1 convolution is first used to reduce the number of channels of the input features, thereby reducing the number of parameters included in the module; for the scale diversity of the ships, 3∗3 convolution, 3∗3 convolution combined with 3∗3 dilated convolution (ratio = 3), and 3∗3 convolution combined with 5∗5 dilated convolution (ratio = 3) are adopted for different convolution branches to achieve feature information extraction for three different scales (3∗3, 9∗9 and 15∗15, respectively). Then, the outputs of different branches are concatenated, and the features of different scales are fused. Finally, the number of output channels is adjusted via linear convolution. Setting $X \in \boldsymbol{R}^{W\times H\times C}$ as the input to the network, the multibranch convolution module can be represented by the following formula:

$$\boldsymbol{X'} = \mathrm{ReLU}\left\{\mathrm{BN}\left[f_{conv1\times1}\left(\boldsymbol{X}\right)\right]\right\} \tag{1}$$

$$\boldsymbol{X''} = \mathrm{Concatenate}\left[f_{br1}\left(\boldsymbol{X'}\right), f_{br2}\left(\boldsymbol{X'}\right), f_{br3}\left(\boldsymbol{X'}\right)\right] \tag{2}$$

$$\tilde{\boldsymbol{X}} = \mathrm{ReLU}\left[f_{CL}\left(\boldsymbol{X''}\right)\right], \quad \tilde{\boldsymbol{X}} \in \boldsymbol{R}^{W\times H\times C} \tag{3}$$

where $f_{conv1\times1}$ is the convolution function with a convolution kernel of 1∗1 and f_{br1}, f_{br2} and f_{br3} represent different branch convolutions. Additionally, f_{CL} represents a linear convolution. Batch normalization (BN) [39] layers and a rectified linear unit (ReLU) layer [40] are applied to the network to accelerate network convergence and avoid overfitting. Fig. 1 shows the multiple convolution structures with dilated convolution and their corresponding RFs. In Fig. 1(a), the first part of the module is the "previous layer", which represents the feature information extracted by the previous network. The last part of the module is "ReLU activation", which is used to increase the nonlinearity of the neural network and reduce the occurrence of overfitting problems. The data are transferred in the same direction as the arrow. The dilated convolution layer is behind the normal convolution layer. Notably, in Fig. 1(b), the RFs are significantly expanded after the adoption of dilated convolution.

For image data, traditional CNNs capture only the dependencies between pixels in their small spatial neighbourhood, and it is difficult to capture the dependence between long distance pixels. Although the multibranch convolution structure uses dilated convolution to obtain a large range of information, it still cannot obtain a global understanding of the scene information. Therefore, this paper introduces a non-local means algorithm [28] to model global context scenarios. We use feature map $\tilde{\boldsymbol{X}} = \left[x_1, x_2, \cdots x_i, \cdots, x_{N_p}\right]$ extracted by the multibranch convolution as the input of the non-local block, where $N_p = W \times H$. H and W represent the height and width of the input features, respectively. The representation of the non-local mean of the input features is as follows:

$$y_i = \frac{1}{C\left(x\right)} \sum\nolimits_{j=1}^{N_p} f\left(x_i, x_j\right) g\left(x_j\right) \tag{4}$$

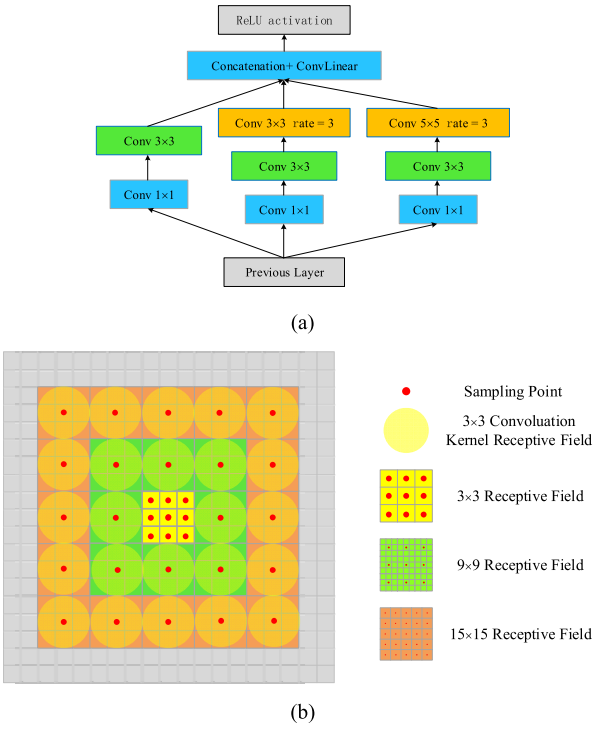


FIGURE 1. Multibranch convolution and the corresponding receptive fields. (a) Multibranch convolution. (b) The receptive fields corresponding to different convolution branches.

where $x_i \in \boldsymbol{R}^{1\times1\times C}$ is the input signal and $y_i \in \boldsymbol{R}^{1\times1\times C}$ is the output signal. Additionally, i is the index of the output position in the space, and j is the index that enumerates all possible positions correlated with i. $g\left(x_j\right)$ is a mapping function used to calculate the eigenvalue of the input signal at position j. $g\left(x_j\right)$ can be represented by a linear function:

$$g\left(x_j\right) = W_g x_j \tag{5}$$

where W_g is the weight matrix. $f\left(x_i, x_j\right)$ is the correlation coefficient between the positions of i and j, which can be expressed by an Embedded Gaussian function [35]. The correlation coefficient between two points in the embedded space can be presented as follows:

$$f\left(x_i, x_j\right) = e^{\theta(x_i)^T \phi(x_j)} \tag{6}$$

The embedded terms $\theta\left(x_i\right)$ and $\phi\left(x_j\right)$ can be expressed as follows:

$$\theta\left(x_i\right) = W_\theta x_i, \quad \phi\left(x_j\right) = W_\phi x_j \tag{7}$$

where W_θ and W_ϕ are the weight matrices. $C\left(x\right)$ is the normalization coefficient, which can be represented by $\sum_{j=1}^{N_p} f\left(x_i, x_j\right)$. Therefore, output signal y_i can be expressed as follows:

$$y_i = \frac{1}{\sum_{\forall j} e^{(W_\theta x_i)^T (W_\phi x_j)}} \sum\nolimits_{\forall j} e^{(W_\theta x_i)^T (W_\phi x_j)} \left(W_g x_j\right) \tag{8}$$

Then, we can obtain the following formula:

$$y = \mathrm{softmax}\left(x^T \cdot W_\theta^T \cdot W_\phi \cdot x\right) \cdot \left(W_g x_j\right) \tag{9}$$

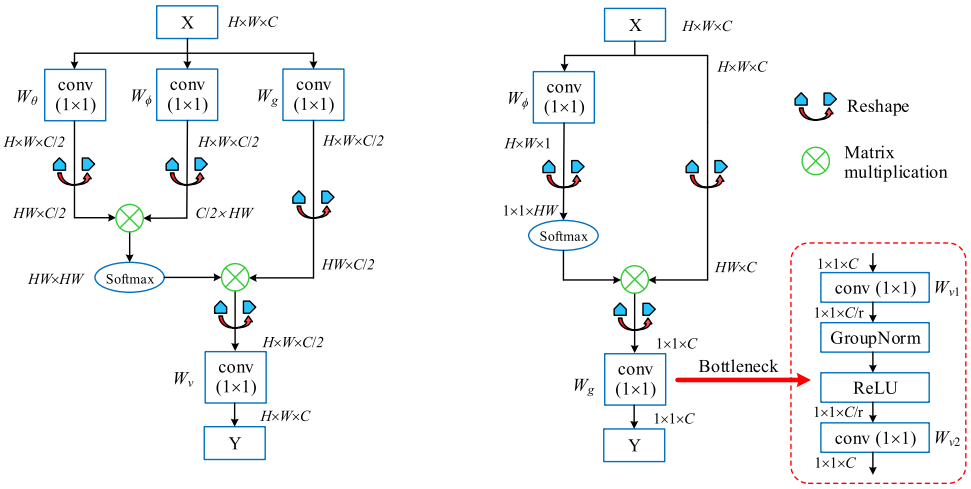


FIGURE 2. The non-local block and the simplified non-local block. (a) The non-local block. (b) The simplified non-local block.

To match the input dimension, we construct a linear function W_v to adjust the dimension, as shown in Eq. (10).

$$y = W_v \cdot \left[\text{softmax}\left(x^T \cdot W_\theta^T \cdot W_\phi \cdot x\right) \cdot \left(W_g x_j\right)\right] \quad (10)$$

In the embedding space, the weight matrix can be learned through a CNN. Then, the non-local block with a convolutional form is shown in Fig. 2(a).

However, the disadvantage of this block is that the number of parameters is large and not conducive to deployment in deep networks. Reference [37] found that after the network is trained, the global contextual features learned by the non-local network are almost the same at different locations, which indicates that the network learns global information without location dependence. However, the non-local block actually learns an independent attention map for each query location, which is a waste of computational resources for establishing pixel-level pairwise relationships. Therefore, the global features can be modelled directly as a weighted average of all locational features and then aggregated for the current query location. The non-local block is simplified by computing a global attention map and sharing it for all locations. Through the above analysis, the non-local block can be simplified to the following form:

$$y_i = \sum_{j=1}^{N_p} \frac{e^{W_\phi x_j}}{\sum_{m=1}^{N_p} e^{W_\phi x_m}} \left(W_g \cdot x_j\right) \quad (11)$$

Through the distributive property of multiplication, we can further obtain the following equation.

$$y_i = W_g \sum_{j=1}^{N_p} \frac{e^{W_\phi x_j}}{\sum_{m=1}^{N_p} e^{W_\phi x_m}} x_j \quad (12)$$

It can be inferred that unlike the traditional non-local block, x_j is independent of the query position i, which suggests that the global information is shared between all query positions after simplification. The network structure of the simplified non-local block is shown in Fig. 2(b). Although this method can effectively model the global information, the 1∗1 convolution represented by W_g still has C∗C parameters. In a deep network, the feature map contains many channels, and the number of parameters associated with 1∗1 convolution is enormous (for example, when the number of network channels is C = 1024). Therefore, the bottleneck structure proposed in [41] was introduced to optimize the network parameters. We used a bottleneck approach instead of the original 1∗1 convolution, as shown in Fig. 2(b). In the bottleneck structure, r represents the reduction ratio of the bottleneck. The advantages of this bottleneck are that the number of parameters is reduced (C∗C is reduced to 2C∗C/r) and that different proportions of convolutional layers increase the nonlinearity of the acquired information features. Introducing an additional two-layer convolution step increases the difficulty of optimization, so group normalization (GroupNorm) [42] is added to the bottleneck process (before the ReLU step), which simplifies the optimization process and provides regularization. Additionally, the impact of the batch size on network performance is reduced. The resulting output of the global information module can be expressed as follows:

$$y_i = W_{v2} \cdot \text{ReLU}\left[\text{GN}\left(W_{v1} \sum_{j=1}^{N_p} \frac{e^{W_\phi x_j}}{\sum_{m=1}^{N_p} e^{W_\phi x_m}} x_j\right)\right] \quad (13)$$

The extracted feature map can be globally recalibrated through the global information learned by the network itself, which enhances target information and suppresses non-target information. Therefore, the output of the adaptive recalibration module combined with global information can be expressed as follows:

$$z_i = x_i + x_i \cdot \text{Sigmoid} \times \left\{W_{v2} \cdot \text{ReLU}\left[\text{GN}\left(W_{v1} \sum_{j=1}^{N_p} \frac{e^{W_\phi x_j}}{\sum_{m=1}^{N_p} e^{W_\phi x_m}} x_j\right)\right]\right\} \quad (14)$$

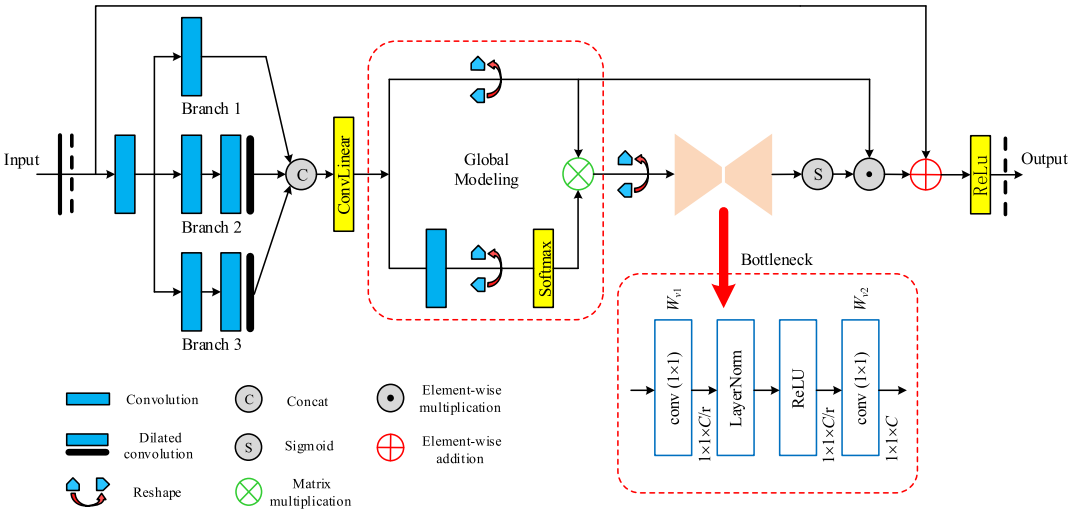


FIGURE 3. The network structure of multiscale adaptive recalibration module.

In (14), we use the sigmoid function to normalize the global information into [0,1], and use it as the weight factor to recalibrate the acquired features; we then construct a residual structure to obtain the final output of the module. This residual structure can enhance the feature information of the target through the global information on the basis of ensuring the original features. Finally, the network structure of the multiscale adaptive recalibration module can be defined as follows:

$$z = \tilde{x} + \tilde{x} \cdot \text{Sigmoid} \times \left\{ W_{v2} \cdot \text{ReLU}\left[\text{GN}\left(W_{v1} \cdot \text{softmax}\left(W_k \cdot \tilde{x} \right) \right) \right] \right\} \quad (15)$$

where z is the output of the module and the corresponding network structure is shown in Fig. 3. The multiscale features obtained by convolution are adaptively calibrated by the obtained global information. Similar to a self-attention mechanism, the target information is enhanced to suppress non-target information.

Compared with the attention module proposed in reference [27], the MSAR module proposed in this paper has the following advantages:

First, the MSAR module is more concise. The modules in these two papers have different implementations. In previous work, we constructed the attention model by superimposing convolution and deconvolution layers to obtain saliency feature maps, and we fused these saliency feature maps to make the network have a strong feature expression ability. In this paper, we calibrate the target features extracted by the convolution network through global information, and the acquisition of global information can be achieved by simply changing the network structure, without a need for superimposing convolutional layers to obtain a larger receptive field or establish a dependency between pixels, as shown in Fig. 3.

Second, the utilization of the convolution layer is higher. The attention model in previous papers adopted a parallel structure, which acquires masks via the convolution layer and deconvolution layer. However, the mask branch is independent of the convolution branch, and the utilization of the convolution layer is insufficient. The MSAR module proposed in this paper uses a serial structure to calibrate the target information on the basis of extracting features from the convolution layer, which improves the utilization efficiency of the convolution network.

Third, the module is more lightweight. The proposed MSAR block reduces the number of parameters and makes the network lighter by optimizing the network structure and using a bottleneck structure. For rotation detection, more target information must be predicted, so the lightweight module is necessary. In the model proposed in the previous paper, the horizontal detection algorithm needs only 9 anchor boxes to match the target, while in rotation detection, the number of anchor boxes increases to 108 because of the need to predict additional information of scale and angle, which makes the computation increase exponentially. Therefore, if the attention model proposed in the previous article is used for rotated target detection, the detection time will be greatly increased. The model proposed in this paper has the advantage of being lightweight, which can make the network maintain a faster detection speed. In the experimental part, we present an experimental comparison with the previous attention model to illustrate the advantages of our algorithm more specifically.

Fourth, fast detection of rotated SAR ship targets is realized. More importantly, the proposed detection algorithm is very effective for ship targets that are densely arranged. In previous work, a horizontal detection algorithm was used to detect ship targets in various scenarios. In this paper, the proposed algorithm combines the precise positioning advantages of rotation detection and the speed advantage of a single-stage detection framework, effectively overcoming issues related to the complexity of application scenarios, the difficulty of dense target detection, the redundancy of detection areas and the diversity of target scales in SAR ship detection. In addition, the proposed algorithm can estimate the movement trend of a sea surface target through the predicted

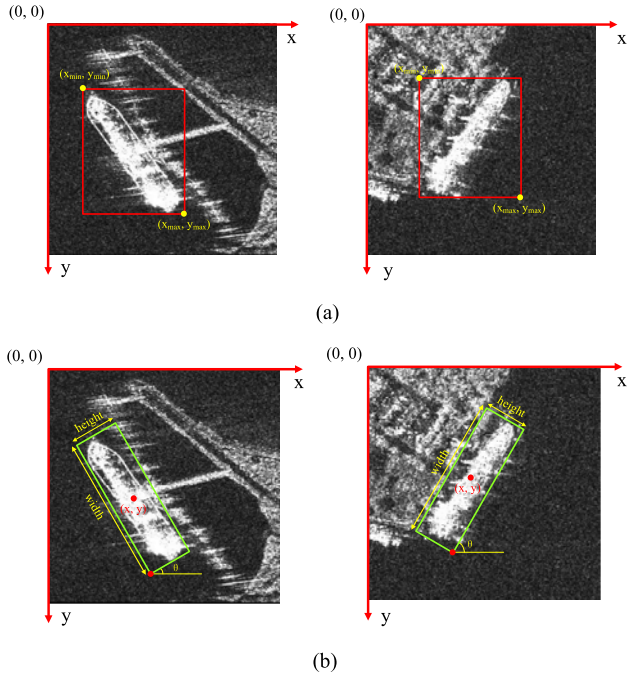


FIGURE 4. Comparison of horizontal bounding and rotated bounding boxes. (a) Horizontal bounding boxes. (b) Rotated bounding boxes.

angle, and it is not necessary to design the relevant algorithm separately. In particular, we propose a multiscale adaptive calibration network, MSARN, to calibrate the multiscale features extracted from the network through global information, such that the network has both a robust expression ability and an accurate positioning ability. Furthermore, the proposed MSARN is lightweight, which enables the algorithm to achieve near-real-time detection. In addition, we design many new methods for the model, including a pyramid anchor box, a loss function and the Soft-RNMS algorithm for rotation detection, and these methods further improve the efficiency of detecting rotated SAR ship targets.

B. ROTATED BOUNDING BOX REPRESENTATION

Traditional detection methods use a horizontal rectangular box to mark the detected object and determine the position of the object based on the coordinates of the upper-left corner (xmin, ymin) and lower-right corner (xmax, ymax) of the rectangular box [11], [13], [115], as shown in Fig. 4(a). However, a ship target is different from a conventional object, usually with a large aspect ratio and a certain angle of rotation. It is inappropriate to represent a ship target with a traditional horizontal bounding box. Additionally, the bounding box contains a large amount of non-target information, which causes some interference associated with the precise positioning of the target. Moreover, the width and height of the horizontal bounding box have a large variation from the true width and height of the target and cannot express the true size information of the target. Therefore, we have introduced a rotated bounding box to improve the positioning effect of the detector, as shown in Fig. 4(b). Different from the rectangular bounding box, the rotated bounding box is mainly represented by five parameters (x, y, w, h, θ), where (x, y) are the coordinates of the centre point of the bounding box and w and h represent the width and height of the bounding box, respectively. The rotation angle θ represents the angle at which the horizontal axis (x-axis) is rotated counterclockwise to the first edge of the encountered rectangle, and we define this edge as the width and the other edge as the height [28]. The range of the rotation angle is $[-90, 0]$. Rotating the bounding box can not only effectively reduce the non-target information in the detection area but also display the true aspect ratio of the object. In addition, the rotated bounding box can estimate the movement trend of a sea surface target through the predicted angle, and it is not necessary to design the relevant algorithm separately.

C. PYRAMID ANCHOR BOX

The anchor box is based on a priori knowledge obtained by statistical analysis of training samples. In the detection task, prediction boxes of different sizes are usually generated with reference to the anchor box [11], [15]. A reasonable selection of the size of the anchor box can effectively improve the ability of the model to detect objects of unknown size and shape. In an SAR image, a ship target has a large aspect ratio and an arbitrary direction. In the conventional detection method, the horizontal anchor box is not well matched with the target. In addition, for densely arranged ship targets, when using non-maximum suppression (NMS) to filter the prediction, the horizontal anchor box is likely to cause the missed detection of targets. Therefore, we have redesigned the rotated pyramid anchor box based on the characteristics ship targets. A rotated pyramid anchor box consists of three parameters, including the scale of the target, the aspect ratio, and the angle. To design a reasonable anchor box, we first analyse the prior information of the data set, as shown in Fig. 5.

It can be inferred that the distribution of the length-width ratio of ship targets is relatively uneven and requires an elaborate anchor box design. According to observations, the aspect ratio of targets has a certain relationship with the target scale. The ships in SAR images exhibit significant scale diversity due to the different imaging resolutions of the images and the scales of the target ships. When the resolution of an SAR image is low, the scale of a ship target is small, and the aspect ratio is not obvious. Therefore, a small target does not produce a large aspect ratio. Conversely, when the resolution of an SAR image is high, the scale of the ship target is large, and the corresponding aspect ratio is also large. Therefore, in the allocation of anchor box scales, we adopt a pyramid structure. In this approach, the feature map of the shallow network output has a high resolution, and the smaller the RF is, the more sensitive the detection will be to small targets; therefore, an anchor box with a small size and aspect ratio is preferred. In a deep network, the resolution of the feature map is low, the RF is large, and the layer is sensitive to large targets, so a large anchor box and a large ratio are preferred. For the selection of the best angle, through Fig. 5(b), a ship

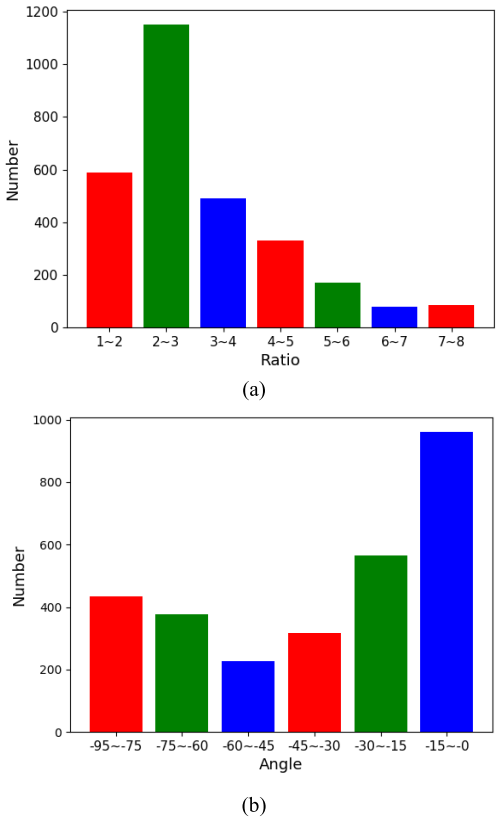


FIGURE 5. Statistical results for the length-width ratio and angle of samples in SRSD. (a) The statistical results for the length-width ratio of the sample. (b) The statistical results for the sample rotation angle.

target is most inclined at $-15\sim0$ degrees, and the distributions in other directions are unconcentrated. Considering the cost of calculations, we choose six angles to include as many ship directions as possible.

In this paper, all parameters of the anchors are selected according to statistics of the data set and a large number of experiments. The statistical information of the data set provides a reference for the setting of the initial values of anchor frame parameters, but a large number of experiments are needed to optimize the initial value in order to set the parameters more accurately. On the basis of statistical information, this paper determines the parameters through a large number of experiments. Similarly, the angle of the anchor is determined via quantitative experiments. Since the distribution of the rotation angle of the target is not concentrated, selecting the angle parameters via experiment is an effective method. The scope of the rotation angle defined in this paper is $[-90, 0]$, and we choose six angles separated by equal intervals to cover all ship directions as much as possible. For different scales, we assign different angle parameters to them through a large number of experiments. On one hand, using different angles for different scales does not increase the amount of computation (the number of anchors per point is still 36); on the other hand, this setting increases the diversity of angles, which can yield a positive detection effect.

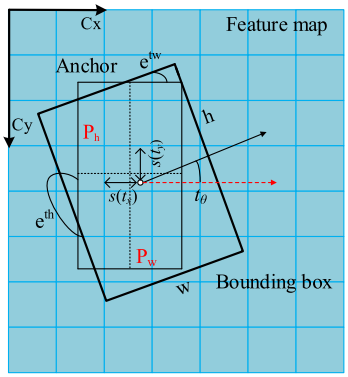


FIGURE 6. The location relationship between the anchor box and bounding box.

Based on the above analysis, we design three anchor boxes of different scales to match the different network outputs. The corresponding scale, angle and ratio of the anchor boxes are set as reported in Table 1. Each point on the feature map will generate 36 anchors boxes of different angles and sizes.

D. LOSS FUNCTION

For the detection task, if the width and height of the bounding box are directly predicted through the network, the stability of the gradient will be affected [14]. Therefore, the output of the network defined in this paper is the positional offset between the predicted bounding box and the anchor box, as represented by t_x, t_y, t_w, t_h and t_θ. The network output has the following mapping relationship with the predicted bounding box:

$$x = S(t_x) + c_x, \quad y = S(t_y) + c_y \tag{16}$$

$$w = e^{t_w} \cdot P_w, \quad h = e^{t_h} \cdot P_h \tag{17}$$

where $S(\cdot)$ represents the sigmoid function; (x, y) are the centre point coordinates of the predicted bounding box; (c_x, c_y) are the upper-left coordinates of the grid where (x, y) is located; P_w and P_h represent the width and height of the anchor box, respectively; and w and h are the width and height of the predicted bounding box, respectively. The position relationship between the anchor box and the predicted bounding box is shown in Fig. 6.

For angle prediction, this paper defines the network prediction as the angle deviation between the prediction box and the anchor box, rather than directly predicting the value of θ, as shown in Eq. (18):

$$t_\theta = \theta - \theta_a + k\pi/2, \quad t_\theta^* = \theta^* - \theta_a + k\pi/2 \tag{18}$$

where θ and θ^* are the angle of the predicted bounding box and the actual angle, respectively, and θ_a is the angle of the anchor box. The parameter $k \in Z$ keeps θ in the range of $[-90°, 0)$. For the bounding box to be in the same position, when k is odd, w and h must be swapped. After introducing the angle information, the rotated bounding box can accurately locate the target.

According to the network output, this paper designs a multitask loss function based on a regression method to

TABLE 1. Setting of anchor box parameters.

Outputs/Anchor parameters	Scales	Angles	Ratios (w:h)
Output 1 (14*14)	80	-10, -25, -40, -55, -70, -85	1:7, 1:6, 1:5, 5:1, 6:1, 7:1
Output 2 (28*28)	50	-15, -30, -45, -60, -75, -90	1:6, 1:5, 1:4, 4:1, 5:1, 6:1
Output 3 (56*56)	30	-5, -20, -35, -50, -65, -80	1:4, 1:3, 1:2, 2:1, 3:1, 4:1

optimize the detection model. The loss function consists of three components: positioning loss, confidence score loss, and angle loss, as shown in Eq. (19).

$$\begin{aligned} &L\left(t_x, t_y, t_w, t_h, C, t_\theta\right) \\ &= \lambda_{pos} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{obj} \mu \left[f\left(x_i, x_i^*\right) + f\left(y_i, y_i^*\right)\right] \\ &\quad + \lambda_{pos} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{obj} \mu \left[\left(w_i - w_i^*\right)^2 + \left(h_i - h_i^*\right)^2\right] \\ &\quad + \lambda_1 \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{obj} f\left(C_i, C_i^*\right) + \lambda_2 \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{noobj} f\left(C_i, C_i^*\right) \\ &\quad + \lambda_{reg} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{obj} f\left(\theta_i, \theta_i^*\right) \end{aligned} \tag{19}$$

where S denotes the number of grids that feature maps are divided into, B is the number of anchor boxes contained in each grid, l_{ij}^{obj} represents the predicted bounding box containing targets, and l_{ij}^{noobj} represents the predicted bounding box without targets. λ_{pos} is the weight of positioning loss, which is set to 5. λ_1 and λ_2 are the weight of the confidence score and penalty term, respectively, and are set to 1 and 0.5. λ_{reg} is the weight of the angle and is set to 2. Moreover, x, y, w, and h are the position parameters of the prediction bounding box, C is the confidence score, and θ is the bounding box angle. The corresponding label information is represented by x^*, y^*, w^*, h^*, C^* and θ^*. To increase the weight of a small target in the loss function, we introduce the balance factor μ related to the position loss, which is defined as follows:

$$\mu = [2 - (w * h) / (w_{in} * h_{in})] \tag{20}$$

where w_{in} and h_{in} are the width and height of the network input, respectively. In the loss function, the cross-entropy function, shown in Eq. (21), is adopted to calculate the loss of the centre point coordinates, confidence score and angle.

$$f\left(x, x^*\right) = x^* \log\left(x\right) + \left(1 - x\right) \log\left(1 - x^*\right). \tag{21}$$

where x is the predicted value and x^* is the true label.

E. MODIFIED ROTATION NON-MAXIMUM SUPPRESSION

NMS is a result processing module in the object detection framework that can effectively remove highly redundant bounding boxes and obtain the final detection result. An important step in NMS is the calculation of the IoU (intersection over union). As noted in Section II.D, the direction of the rotated bounding box is arbitrary. However, the traditional IoU is designed for a horizontal bounding box and is not applicable to a rotated bounding box. Therefore, we introduced the skew IoU [28] to calculate the IoU of the rotated bounding box, as shown in Fig. 7. The internal vertices A and G and the intersections I, J, K, L of two rotated bounding boxes may constitute a convex polygon AIJGKL. Then, the intersection area S_I of the two rotated bounding boxes can be calculated via triangular decomposition [46], as shown in Eq. (22).

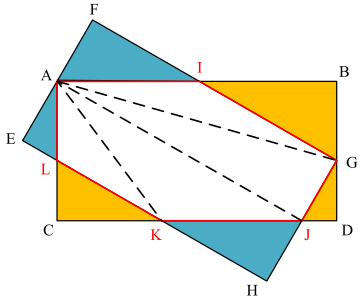


FIGURE 7. Illustration of the skew IoU.

$$\begin{aligned} S_I &= S_{AIGJKL} \\ &= S_{\Delta AIG} + S_{\Delta AGJ} + S_{\Delta AJK} + S_{\Delta AKL} \end{aligned} \tag{22}$$

Next, the skew IoU of two arbitrary rotated bounding boxes A and B can be defined as

$$SIoU\left(a, b\right) = \frac{S_I}{S_1 + S_2 - S_I} \tag{23}$$

where S_1 and S_2 are the areas of rectangles a and b, respectively.

The standard NMS algorithm needs to consider only the influence of the IoU, but this is insufficient for the rotated bounding box, and an angle constraint must be added. On one hand, the angle constraint can make the retained prediction box better match the ground truth data. On the other hand, pre-screening can be performed based on the angle constraint to reduce the number of calculations. We set the angle constraint to 15° (the minimum change in the angle of the anchor box); that is, the IoU is calculated only when the absolute value of the angle difference between the detection box with the highest score and the candidate box is less than 15°; otherwise, the IoU is set to zero.

In NMS and RNMS [28], it is determined whether the bounding box is retained by comparing the score of the bounding box with a threshold. However, when the overlap

ratio of the detection box is large, such a hard threshold setting will result in missed detection. This paper introduces a soft mechanism that uses a re-scoring method to optimize this suppression mode and preserve the detection box with a large overlap rate, as shown in Eq. (24):

$$S_i = \begin{cases} S_i & SIoU\,(B_M, B_i) < N_t \\ S_i e^{-\frac{f(\theta_M,\theta_i)^2}{\sigma}} & SIoU\,(B_M, B_i) > N_t \end{cases} \tag{24}$$

where $e^{-\frac{f(\theta_M,\theta_i)^2}{\sigma}}$ is the penalty function and σ is a hyperparameter that is selected through experiment. The core concept of the soft mechanism is to attenuate the scores of detection boxes with large overlap rates with a penalty function rather than setting the scores of these detection boxes to zero. Unlike the punishment method proposed by [47], we define the penalty factor as follows:

$$f\,(\theta_M, \theta_i) = 1 - \frac{|\theta_M - \theta_i|}{\theta_{\max}} \tag{25}$$

where θ_M is the angle of the detection box with the highest score, θ_i is the angle of the candidate box, and $\theta_{\max} = \max\,(|\theta_M|\,, |\theta_i|)$. When the IoU is the same, the angle can effectively reflect the degree of coincidence of the two rotated bounding boxes. Soft-RNMS can make the remaining detection boxes largely match the ground truth data, effectively removing the redundant detection boxes. When the ships are densely arranged, the detection boxes that may contain targets can be preserved by reducing the scores, which avoids the missed detection of ships that are densely arranged. The algorithm flow is shown in Fig. 8. The subsequent experiments demonstrate the effectiveness of this approach.

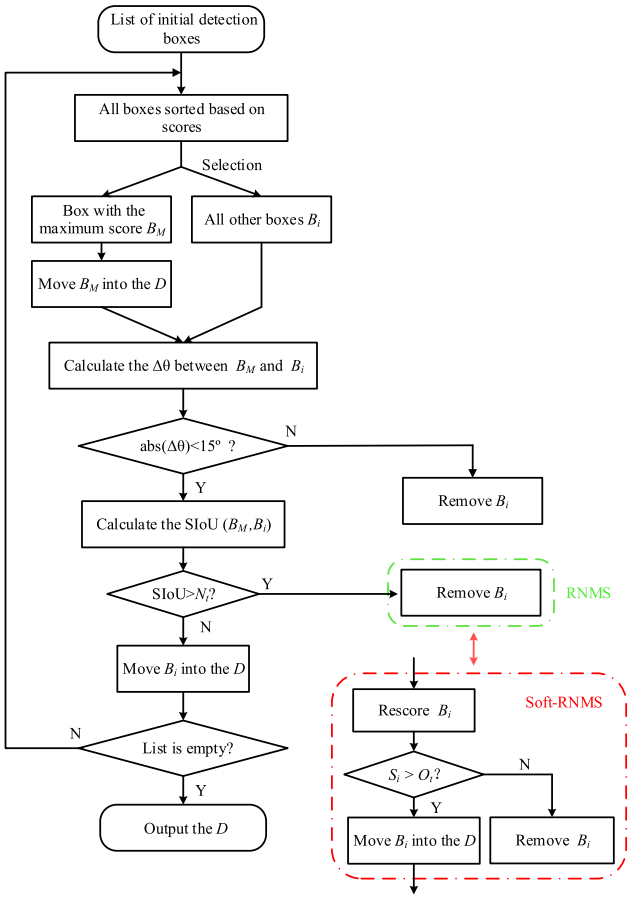


FIGURE 8. The algorithm flow of Soft-RNMS. B_M is the detection box with the highest score, B_i is the candidate box, S is the score of the detection box, $\Delta\theta$ is the angle difference between different detection boxes, D is the final result list, N_t is the NMS threshold, and O_t is the evaluation threshold. The redundant detection boxes are filtered by Soft-RNMS, and the final detection result is obtained.

F. OBJECT DETECTION NETWORK BASED ON AN ADAPTIVE RECALIBRATION MECHANISM

Ships in SAR images have complex backgrounds and large scale variations, which are represented in feature maps at different depths. Therefore, global contextual modelling of features at different depths can effectively distinguish between targets and the background. The traditional non-local block can obtain global information, but it is not suitable for applications in deep networks. In these networks, global modelling using a non-local block will increase the number of parameters exponentially, which will consume extensive computational resources. However, if global modelling is performed only in a shallow network, a single global model can calibrate a feature only once. Once the calibration of the feature is inaccurate, it is difficult to correct this issue in a deep network, which contradicts the idea of using global information to calibrate the target features. The multiscale adaptive recalibration module proposed in this paper has the advantages of being lightweight and easily embedded in the framework of a network of any depth. Therefore, the module can be constructed as a basic unit to form an MSARN, and the MSARN can model features of any depth with only a small increase in the number of parameters. In addition, the MSARN combines a feature pyramid network (FPN) structure [43] to fuse the position information of the shallow features with the semantic information of the deep features. This approach not only retains sufficient semantic information but also ensures the accuracy of the location information. The overall structure of the network is shown in Fig. 9.

In the network, the dimensions of the input image are first adjusted with a 7∗7 convolutional layer and then downsampled with the maxpooling layer. The feature extraction network consists of four stages, each of which uses the multiscale adaptive recalibration module as a basic unit to construct a feature pyramid. A residual block [44] with a step size of 2 is used for downsampling between different stages, and the convolutional layer in the residual block can change the dimension and connect the features of different stages. In each stage, by connecting several MSAR modules in series, a series of feature maps with decreasing spatial resolution and increasing RF is obtained. We fuse these feature maps that express different meanings, highlighting features that are advantageous for positioning. The entire network does not adopt MSAR modules with the same structure because the module includes dilated convolution, and

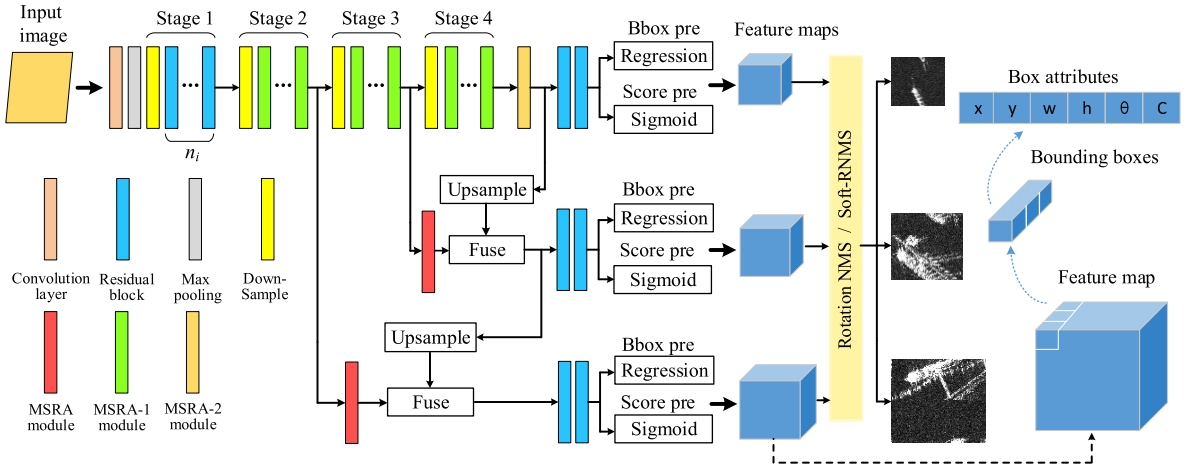


FIGURE 9. Structure of the proposed multiscale adaptive recalibration network.

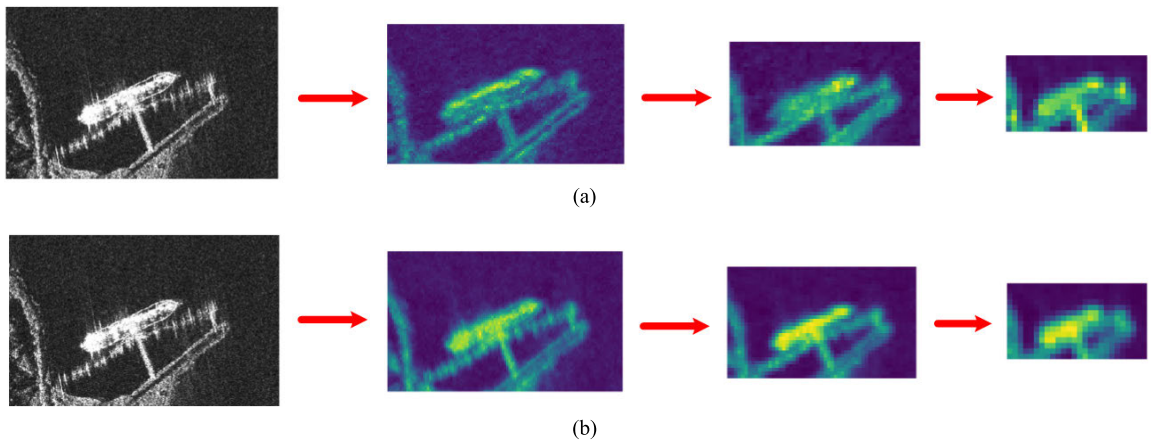


FIGURE 10. Comparison of feature maps at Stages 2, 3 and 4. (a) Standard convolution. (b) The MSAR module.

continuous dilated convolution will disrupt the continuity of spatial information. Therefore, we use MSAR module-1 in the backbone structure. The main difference between MSAR module-1 and the original MSAR module is that dilated convolution in the former is replaced by standard convolution, and MSAR module-1 retains only one convolutional branch considering the complexity of the network. The MSAR module is used as an independent branch to aggregate multiscale features. In addition, since the original image is downsampled 32 times, the output feature map has a small size. We adjust the MSAR module structure accordingly, leaving only 3∗3 convolution and 3∗3 convolution combined with 3∗3 dilated convolution for these two branches as MSAR module-2. This module is applied after Stage 4.

The outputs of the network are three different scale feature map tensors, each of which was divided into multiple grids. Each grid contains the position attributes, angles, and confidence scores of the predicted targets. The network output is filtered by a confidence threshold and a Soft-RNMS algorithm to obtain a final result.

To illustrate the advantages of the multiscale adaptive recalibration module, feature maps at different stages of the network are compared, as shown in Fig. 10. It can be observed that when the network structure is the same, compared with the standard convolution block, adopting the MSAR module as the basic unit of the network can generate a stronger response to the feature information of the target. Notably, the proposed module can adaptively capture target features and is more sensitive to the direction of the rotated target.

III. EXPERIMENTS

In this section, we describe the experiments conducted in this study, including the network training process, experimental details, and analysis of the experimental results.

A. THE EXPERIMENTAL PLATFORM AND DATA SET

The network proposed adopts TensorFlow [48] as the basic framework. Network training and testing were performed on a workstation with an Intel(R) Xeon Silver 4114@2.20 Hz×40 CPU, an NVIDIA GTX TITAN-XP GPU and 128 GB memory. The input image was rescaled to a

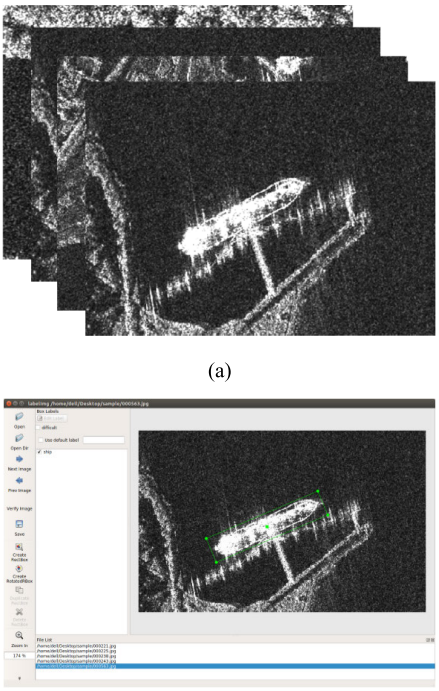
size of 448∗448 pixels. Different data augmentation strategies, such as random flipping, adding noise, and random cropping, are used in the network training process to make the trained model more robust [49], [50]. Stochastic gradient descent (SGD) is used as the optimization algorithm of the network. The weight attenuation coefficient is 0.0005, and the momentum parameter is 0.9. An early stopping mechanism was adopted in the training process [40]. The learning rate is attenuated as the network loss decreases, and the initial learning rate is set to 0.01. To avoid gradient explosion, a warm-up step [49] was introduced in the initial training stage, and the corresponding number of epochs was 3.

We validated the proposed model based on the SAR rotation ship detection (SRSD) data set. The SRSD data set is based on SSDD [21]. These SAR images were collected from RadarSat-2, Sentinel-1 and TerraSAR-X. They include ship targets at different resolutions (1 m to 15 m) and different sizes of ships for different scenarios (nearshore and offshore). The diversity of sample scenarios ensures that the trained model has strong generalization ability. In addition, since the ships are too small to be detected in low-resolution images, only targets with more than three pixels are marked. In summary, the data set contains 1,160 ship target images of different scenes. Different from SSDD, in SRSD, we introduce a modified labelling method and label the actual length and width of the target and the angle of rotation of the target relative to the horizontal axis to replace the original horizontal bounding box. The method was implemented in the LabelImg software package, as shown in Fig. 11. Compared with the method of labelling four consecutive points used in [28], [32], and [45], the method adopted in this paper reduces the error rate of manual labelling. In addition, to improve the generalization ability of the trained model, we extend the data set according to the above labelling method. Specifically, 14 SAR images containing the rotated ship targets were cut into small slices and labelled in the PASCAL VOC format [17]. Finally, the number of images contained in the data set was increased to 1,442. We divided the data set into a training set, a validation set, and a test set at a ratio of 7:1:2.

The ships in SAR images exist at a variety of scales due to multiresolution imaging modes and the variety of ship shapes. In the established SRSD, original SAR images were cut into small slices which containing the rotated ship targets. These small SAR images are labelled and fed into the neural network for training and testing. Therefore, whether the resolution is different or the actual size of the ship target is different, the target in SAR image will exhibit different scales. In SAR images, different target scales can be defined as different numbers of pixels.

B. EVALUATION METRICS

In this paper, the average precision (AP) and precision-recall curve (PRC) are used to as the evaluation metrics, which reflect the comprehensive performance of the



(a)

(b)

FIGURE 11. Image rotation annotation process. (a) Original SAR images. (b) Modified LabelImg software.

algorithm [15] [16], [27], as shown in Eq. (26):

$$AP = \sum_{k=1}^{n} precision\,(k) \times \Delta recall\,(k) \tag{26}$$

where n is the total number of images in the data set, $precision\,(k)$ is the precision at a cutoff point of k images, and $\Delta recall\,(k)$ is the difference in $recall$ between cutoff point $k-1$ and cutoff point k. Precision and recall can be calculated as follows:

$$precision = \frac{N_{tp}}{N_{tp} + N_{fp}} \tag{27}$$

$$recall = \frac{N_{tp}}{N_{tp} + N_{fn}} \tag{28}$$

where N_{tp} is the number of correctly detected samples, N_{fp} is the number of falsely detected samples, and N_{fn} is the number of missed samples [51]. Unlike those used in the horizontal bounding box method, this paper uses rotating evaluation metrics to determine whether the target is detected correctly. Ships generally have a large length-width ratio and a certain rotation angle, which makes them sensitive to changes in the IoU. In one extreme case, when the length-width ratio of the ship target is 1:7 and the angle offset is 15°, the IoU of the predicted bounding box and the ground truth is only 0.38, but in this case, the target can be defined as correctly detected [28], [45]. If the IoU threshold is set to 0.5, the bounding box will be deleted, even though the object actually exists. To avoid this situation, this paper sets the IoU threshold to 0.35, which is defined as correct detection when

TABLE 2. The details of the SAR rotation ship detection (SRSD) data set.

Sensors	Polarization	Resolution	Scenario	Number of images	Number of ships
RadarSat-2 Sentinel-1 TerraSAR-X	HH, VV HV, VH	1 m-15 m	nearshore offshore	1442	3691

FIGURE 12. A comparison of trends in loss curves. (a) Experiments on the training set. (b) Experiments on the verification set.

the IoU between the predicted bounding box and the actual target is greater than 0.35.

C. EXPERIMENTAL DETAILS

To evaluate the proposed network, we conducted an in-depth study of the design details of the algorithm through experimental methods. All experiments were performed on the test set of the SRSD. Differing from the conventional evaluation criteria, this paper adds a rotation evaluation criterion to comprehensively evaluate the performance of the algorithm. Therefore, the evaluation index of the algorithm includes four parts: the rotation metric $AP_{0.35}$, PASCAL VOC metric $AP_{0.50}$, strictness metric $AP_{0.75}$, and inference time.

1) EXPERIMENTS ON THE VERIFICATION SET

In order to clearly illustrate that the proposed model does not have an overfitting problem, we conducted the experimental analysis on the validation set, including the loss variation

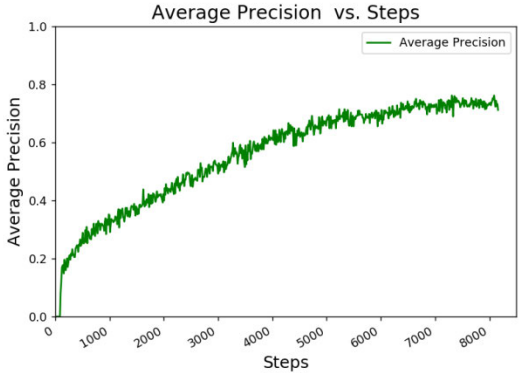


FIGURE 13. The trend of AP with the number of training steps.

curve and the AP variation curve, as shown in Fig. 12 and 13. As can be observed from Fig. 12, the loss of the verification set is convergent, and its change trend is consistent with the change trend of the loss of the training set, indicating that there is no overfitting problem in the model. At the same time, we adopted the callback to test the verification set, as shown in Fig. 13. It can be observed that with the increase of training steps, the AP increases, and the final AP is 75.64%.

2) EFFECT OF THE MSAR MODULE ON THE DETECTION PERFORMANCE

The MSARN is mainly composed of four stages, each of which contains a given number of MSAR modules (3, 4, 6, and 4). Table 3 presents the results of integrating the MSAR modules into different stages. We use ResNet-50 [44] as the baseline. The reduction ratio of the bottleneck in the MSAR module is set to 8 by default. Compared with the results for the baseline model using standard residual blocks, all the stages benefit from integrating the new module. Embedding MSAR modules in Stage 3 and Stage 4 can achieve better performance than embedding in Stage 2, indicating that the most obvious semantic features can enable MSAR modules to effectively calibrate the target information. In addition, integrating MSAR modules into all stages can achieve higher AP than embedding in a certain stage alone, indicating that the gains of the MSAR modules in different stages are complementary and can be effectively combined to further improve network performance. The PRCs of the MSAR modules embedded in different stages are shown in Fig. 14. It is apparent that the MSAR modules can be embedded in all stages to achieve the maximum performance improvement.

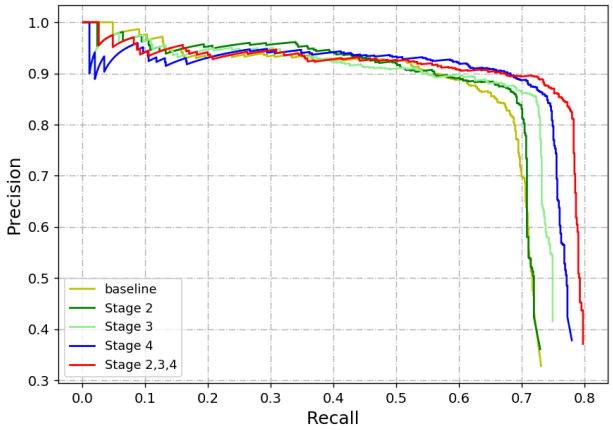


FIGURE 14. Precision-recall curves corresponding to embedding MSAR modules in different stages.

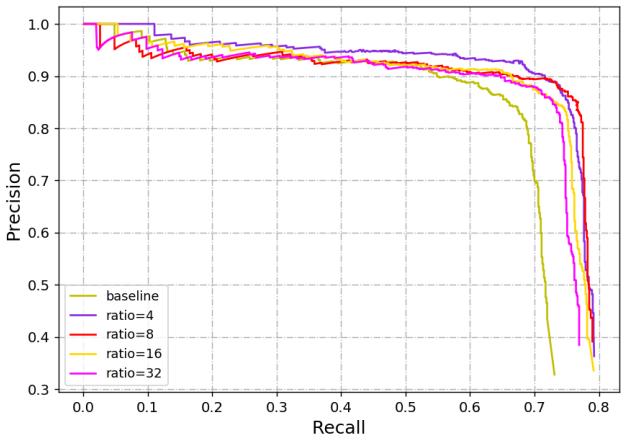


FIGURE 15. Precision-recall curves corresponding to different reduction ratios.

TABLE 3. Effects of embedding MSAR modules at different stages on the detector performance.

Stages	$AP_{0.35}$	$AP_{0.50}$	$AP_{0.75}$	Time (ms)
Baseline	80.11	72.46	31.64	31.42
Stage 2	81.08	73.07	32.14	31.67
Stage 3	81.86	73.95	32.56	32.44
Stage 4	82.10	74.29	33.17	33.69
Stages 2, 3, and 4	**82.91**	**75.33**	**33.80**	35.38

TABLE 4. Effects of different reduction ratios on the detector performance.

Stages	$AP_{0.35}$	$AP_{0.50}$	$AP_{0.75}$	Time (ms)
Baseline	80.11	72.46	31.64	31.42
Ratio 4	**83.30**	**75.78**	**34.20**	35.41
Ratio 8	82.91	75.33	33.80	35.38
Ratio 16	82.85	75.25	33.72	35.26
Ratio 32	82.78	75.10	33.50	34.62

3) BOTTLENECK SETTING

A bottleneck was introduced into the MSAR modules to reduce redundant network parameters. By adjusting the reduction ratio r of the bottleneck, the network can achieve a trade-off between performance and the inference time [37], [41]. Table 4 reports the results for different reduction ratios. As ratio r decreases, AP continues to increase. When $r = 4$, AP achieves the maximum increase, and the inference time increases only a small amount. When $r = 32$, the performance of the model is still greatly improved compared to that of the baseline, indicating that the performance of the network is robust for various reduction ratios. The PRCs corresponding to different ratios r are shown in Fig. 15. The best results are obtained when $r = 4$. Considering the AP and inference time, this paper sets the reduction ratio to 4. It should be noted that all MSARN modules in the network adopt the same reduction ratio. However, there are differences among networks at different levels, and the use of the same ratio may cause the network to be non-optimal. Therefore, on the basis of a given structure, fine tuning r can further improve the performance of the network.

4) COMPARISON OF RNMS AND SOFT-RNMS

The AP at different evaluation thresholds can reflect the best performance of the detector [40]. To verify the effectiveness of Soft-RNMS for ship detection, we compared the detection performance of the original RNMS method and Soft-RNMS at different thresholds, as reported in Table 5.

The left and right sides of the table correspond to the multiple evaluation thresholds O_t (0.35-0.75) and the AP values of RNMS and Soft-RNMS at multiple NMS thresholds N_t (0.3-0.8). A horizontal comparison indicates that when N_t is the same, the APs of Soft-RNMS at multiple O_t values have a certain degree of improvement compared with those of RNMS. Notably, Soft-RNMS attenuates the scores of detection boxes that have large overlap with the detection box with the highest score based on a penalty function instead of directly deleting these boxes, as in RNMS. In this manner, the detection boxes with high overlap rates are preserved, which improves the detection rate of ship targets that are densely arranged. Through vertical comparison, it can be found that with the continuous increase of N_t, the AP decline is more obvious for RNMS than for Soft-RNMS because the excessive threshold reduces the filtering effect for the repeated detection boxes. However, for Soft-RNMS, the effect of the improvement becomes increasingly obvious because when the IoU between the highest-scored detection box and the candidate box is large, the candidate box has a high probability of repeated detection, and the greater the penalty weight is. When the score of a candidate box after being punished is lower than the set threshold, it will still be deleted, which guarantees the effective filtering of the repeated detection box. Therefore, when the NMS threshold is large, compared with RNMS, Soft-RNMS delays the decline in AP and exhibits strong robustness. Through this set of experiments, we can clearly compare the difference between RNMS and Soft-RNMS based on the detection

TABLE 5. AP comparison across multiple NMS thresholds N_t and values of the parameter σ for RNMS and Soft-RNMS. The best performance for each evaluation threshold O_t is marked in bold for each method.

N_t	$AP_{0.35}$	$AP_{0.50}$	$AP_{0.65}$	$AP_{0.75}$	σ	$AP_{0.35}$	$AP_{0.50}$	$AP_{0.65}$	$AP_{0.75}$
0.3	0.8313	0.7536	0.5268	0.3402	0.1	0.8369	0.7580	0.5326	0.3452
0.4	**0.8330**	**0.7578**	0.5295	0.3420	0.3	**0.8373**	**0.7624**	0.5345	0.3480
0.5	0.8317	0.7576	**0.5318**	**0.3432**	0.5	0.8360	0.7610	**0.5383**	0.3484
0.6	0.8297	0.7561	0.5295	0.3406	0.7	0.8327	0.7597	0.5353	**0.3492**
0.7	0.8247	0.7484	0.5233	0.3336	0.9	0.8284	0.7552	0.5305	0.3437
0.8	0.7915	0.7169	0.4846	0.3157	1.1	0.7986	0.7229	0.5083	0.3362

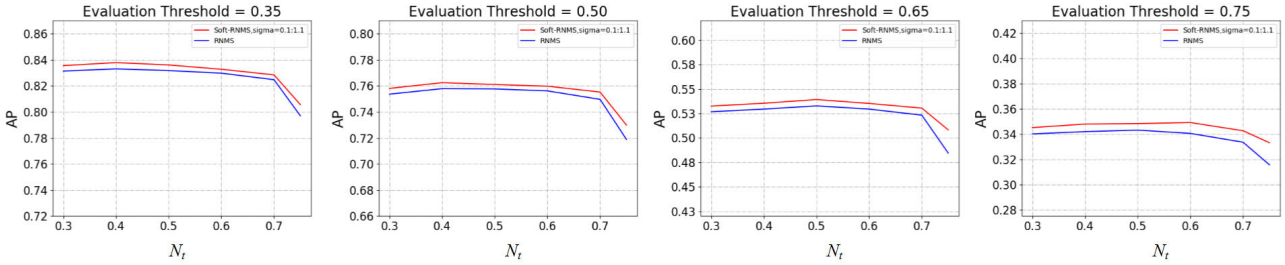


FIGURE 16. Comparison of AP values for RNMS and Soft-RNMS at multiple evaluation thresholds.

results and the influence of the parameter σ in Soft-RNMS to achieve a reasonable selection of σ under different conditions. Fig. 16 is a comparison of RNMS and Soft-RNMS under different evaluation thresholds. Notably, Soft-RNMS provides a certain improvement over RNMS under all threshold conditions.

D. EXPERIMENTAL RESULTS UNDER DIFFERENT SCENARIOS

The complex scenarios in this paper mainly include three aspects: ship targets in the port, ship targets in dense arrangements and ship targets with different scales. These scenarios are common and indeed existing problems in current SAR ship detection. Therefore, to verify the validity of the network model, the ship detection results under different scenarios in the extended SRSD are analysed, as shown in Fig. 17. The first line is the result for ships in a port. This type of ship is characterized by arbitrary directionality and a diversity of scales. In addition, the complex background of the buildings on the shore interferes with the detection process. The proposed algorithm achieves good performance for ships with arbitrary directions and multiple scales. Additionally, the use of global information allows the target to be better distinguished from the complex onshore background. The second line shows the detection result for densely arranged ships. Since the algorithm predicts the angle information of the target, it has the ability to distinguish among densely arranged targets. At the same time, rotating detection solves the problem of missed detection due to the large overlap rates of different detection boxes. The third line is the detection result for inland river ships. The problem in this scenario is that the reef is the same size and shape as the ship target, which results in the false detection of targets. In contrast with the ships, the reefs have no rotation feature. The proposed algorithm learns the rotation feature of the target through a CNN, which can distinguish the ship and the reef. Therefore, the algorithm exhibits good performance in this scenario. The fourth line is the detection result for small objects characterized by a sparse distribution. Since the algorithm integrates shallow location information with deep semantic information and aggregates multiscale features by MSAR modules, the algorithm exhibits satisfactory detection performance for sparse small targets.

E. COMPARISON OF HORIZONTAL DETECTION ALGORITHM AND ROTATION DETECTION ALGORITHM

To intuitively illustrate the advantages of the rotation detection algorithm over the horizontal detection algorithm, we selected the single-stage detection algorithm YOLO v3 [15] as a representative horizontal detection algorithm and compared it with the proposed algorithm. Both algorithms have similar detection frameworks and detection times.

1) DETECTION AREA COMPARISON

The ships in the SAR image have an arbitrary rotation angle. The traditional horizontal detection algorithm cannot express the real scale information of the targets, and the mismatched bounding box results in redundant detection areas. The detection results of the horizontal YOLO v3 detection algorithm, rotation YOLO v3 detection algorithm and the proposed detection algorithm are shown in Fig. 18. The first line reports the ground truth, and the second line lists the detection results of the horizontal YOLO v3 detection algorithm. The third and fourth lines present the detection results of the rotation YOLO v3 detection algorithm and proposed detection algorithm,

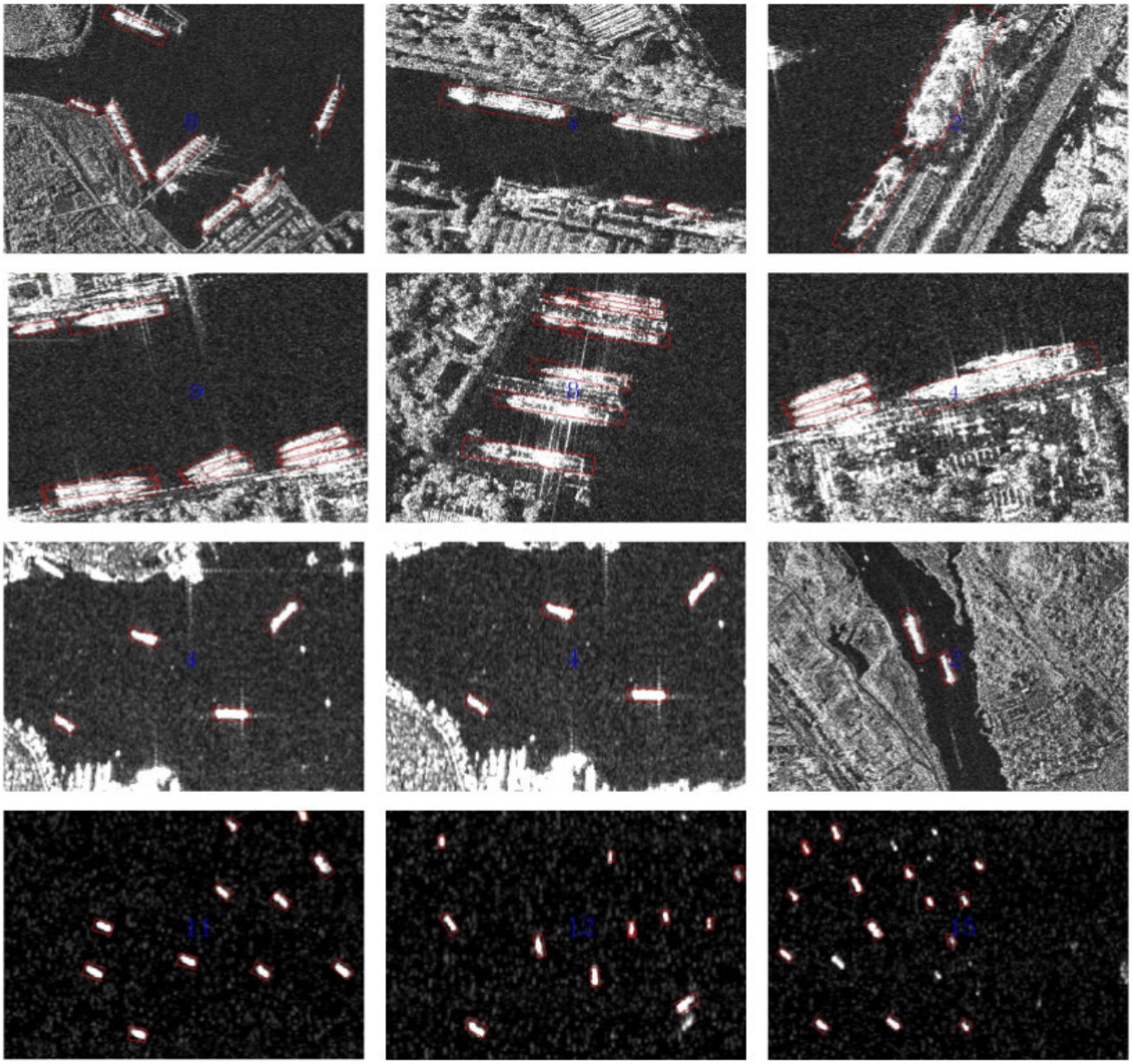


FIGURE 17. Experimental results.

respectively. The bounding box predicted by the proposed rotation detection algorithm best matches the real target, and the redundant information is greatly reduced. Moreover, the proposed algorithm can express the actual size of the target more clearly than the traditional algorithm.

2) DENSELY ARRANGED TARGETS

The ships near the port are densely arranged. When the horizontal bounding box predicted by the traditional algorithm is applied to the densely arranged targets, it will be suppressed due to the high overlap ratio between different detection boxes, thus causing targets to be missed. Fig. 19 compares the detection effects of the three algorithms for densely arranged targets. The first line presents the ground truth, and the second line reports the detection results of the horizontal YOLO v3 detection algorithm. The third and fourth lines show the detection results of the rotation YOLO v3 detection algorithm and proposed detection algorithm, respectively. The horizontal detection algorithm cannot effectively distinguish among the densely arranged ships, resulting in the missed detection of targets (multiple densely arranged targets are detected as one target). Conversely, the proposed rotated detection algorithm can effectively distinguish among the densely arranged targets by assigning a well-designed a priori box and modelling the global information. At the same time, the use of Soft-RNMS reduces the suppression effect on the detection box with a high overlap rate. Therefore, the detection effect of the proposed algorithm is obviously improved for ship targets in a dense array. It can be observed that compared with the proposed algorithm, the rotating YOLO v3 algorithm can detect the target, but the positioning accuracy of the densely arranged is poor.

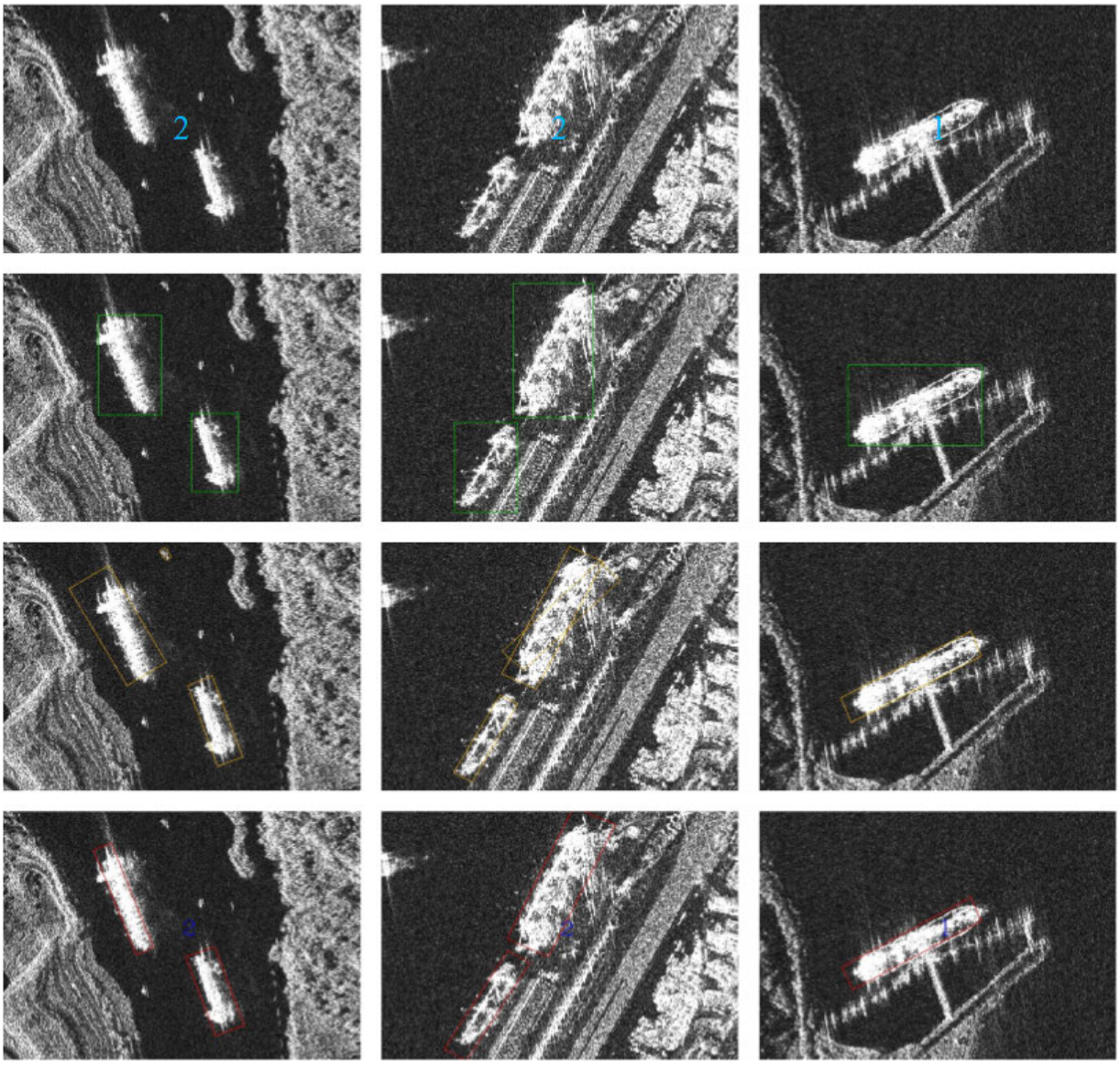


FIGURE 18. Detection area comparison.

3) SMALL TARGETS

The high rate of missed detection of small targets is always one of the main factors that affects the performance of the detector. Since a neural network loses position information during continuous downsampling, it causes the missed detection of small targets. In low-resolution images, the size of ship targets on the sea surface is small, and their distribution is sparse. The missed detection rate of traditional algorithms for this type of target is high. This paper introduces the MSAR module and optimizes the loss function to improve the detection of small-sized ship targets. A comparison between the horizontal detection algorithm and the proposed algorithm for small-sized ship targets is shown in Fig. 20, where (a) is the ground truth and (b), (c) and (d) represent the detection effect of the horizontal YOLO v3 detection algorithm, the rotation YOLO v3 detection algorithm and the proposed detection algorithm, respectively. The number indicates the number of correctly detected ships in the image. The proposed algorithm improves the detection rate of small-sized ships, and the positioning accuracy of ships is significantly improved. In addition, the angle information predicted by the algorithm provides a reference for assessing the trend of the ship targets.

F. QUANTITATIVE COMPARISON

To verify the effectiveness of the proposed method in SAR ship target detection tasks, we compared a variety of influential detection algorithms and detection frameworks, including the Faster R-CNN [11], YOLO v3 [15], RFB Net [16], Attention-ResNet [27], RRPN [28], R2CNN [29], and R-DFPN [32], as reported in Table 6. Compared with that for the horizontal bounding box, the performance of the detection model using the rotated bounding box is generally higher, which proves the effectiveness of the rotating method at ship detection tasks. However, the rotation detection models

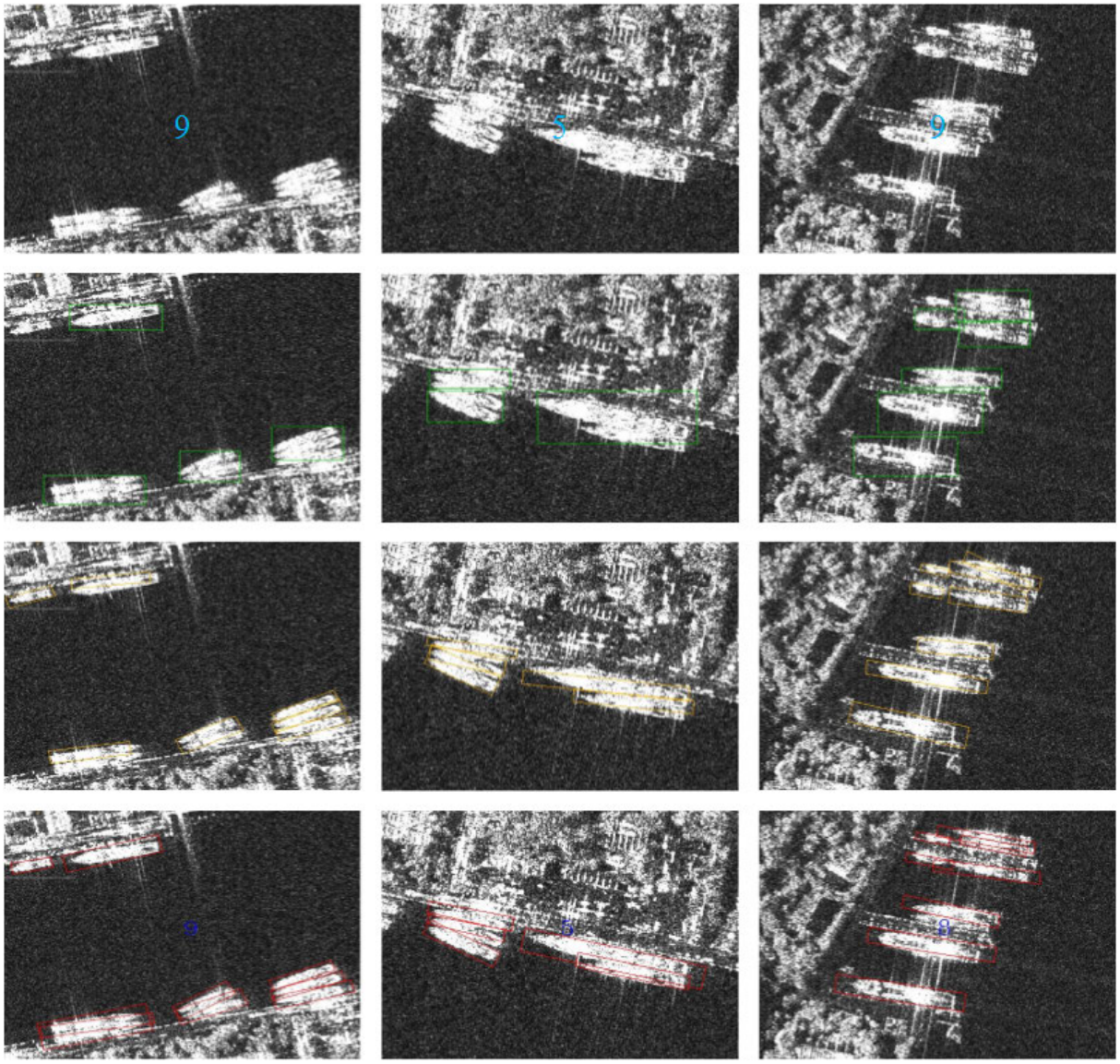


FIGURE 19. Comparison of the detection effect for densely arranged targets.

RRPN, R2CNN, and R-DFPN are based on the two-stage detection framework Faster R-CNN. The detection takes a long time, and the processing efficiency is low. In this paper, the proposed algorithm combines the positioning advantage of the rotating method with the speed advantage of the single-stage framework. Compared with some single stage detection algorithms, such as YOLOv3 with horizontal bounding box and YOLOv3 with rotated bounding box, the average precision is effectively improved and the inference time only increases slightly for the proposed method. Compared with the two-stage algorithm, the proposed algorithm has an absolute speed advantage. The inference time of a single image is 35.4 ms, which is only 1/6 that of R2CNN and 1/11 that of R-DFPN, and it is superior to RRPN based on the AP and inference time. At the same time, we compare the performance of the proposed algorithm with that of the previous Attention-ResNet. The two have similar average accuracy on SRSD, but the algorithm proposed in this paper has a faster detection speed. In addition, we note that under the same network structure, Faster R-CNN combined with FPN has a large difference in terms of AP compared to that of the original Faster R-CNN because the size of most SAR ships in the data set is small. The original Faster R-CNN did not fuse shallow location information with deep semantic information, resulting in the missed detection of small-sized targets. This result verifies the importance of the FPN network structure for SAR ship detection.

G. ABLATION STUDY

To analyse the impact of each of the basic roles in the algorithm on the detection performance, we conducted a step-by-step experiment for the extended public

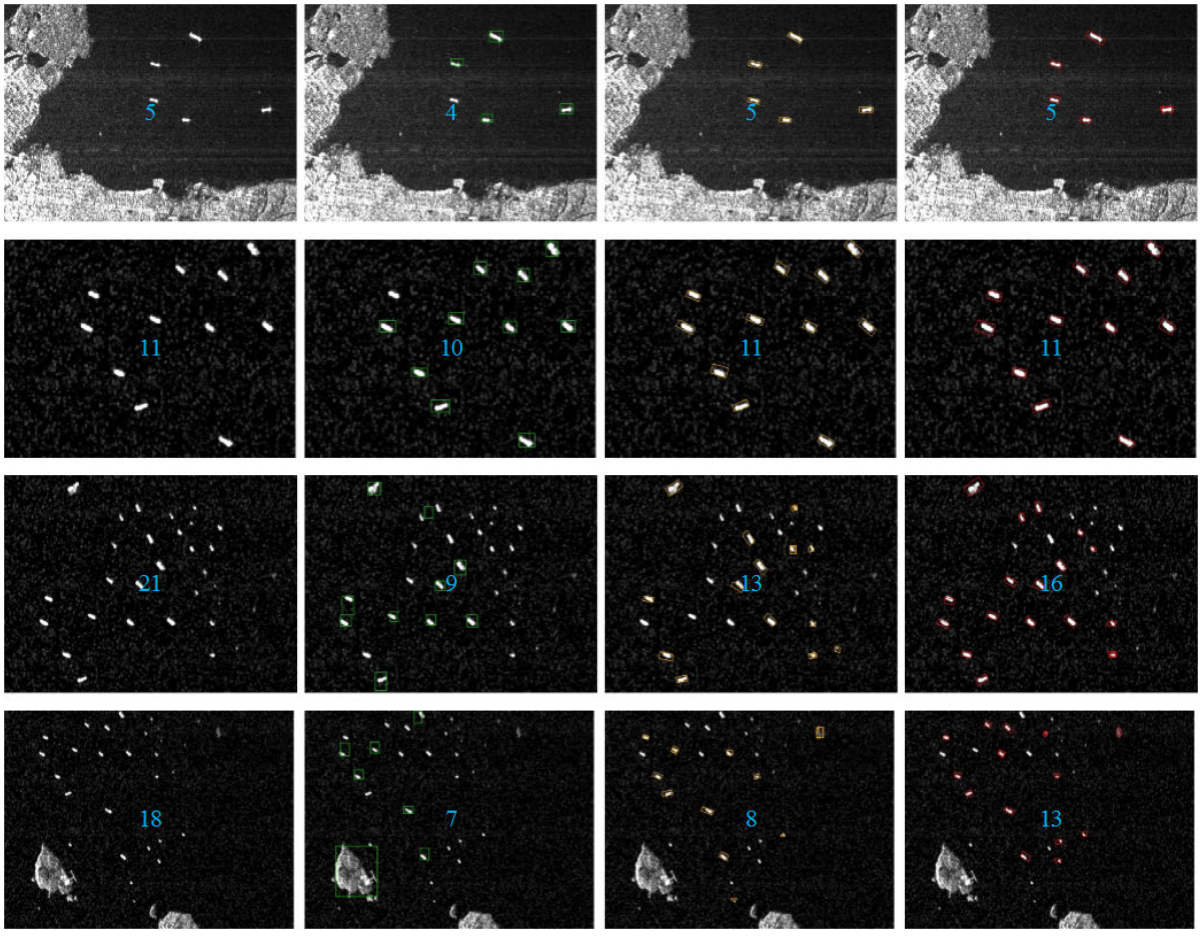


FIGURE 20. Comparison of the detection effect for small targets.

SRSD data set. The experimental results are reported in Table 7.

The basic model uses ResNet-50 as the backbone network. Data augmentation strategies such as random scaling, cropping, and adding noise are applied during the training process. By constructing the rotation anchor and designing the corresponding loss function to detect ship targets, the AP is increased to 72.46 (up 1.66%), which suggests that the rotation detection is effective for the SAR ship detection tasks. Adding the MSAR module separately in the algorithm can increase the accuracy to 73.66 (up 2.86%). However, combining the MSAR modules with rotation anchors yields a better improvement effect of 75.78 (up 4.98%). This finding indicates that the MSAR module has better adaptability to rotation features than do standard convolution blocks. Using Soft-RNMS in the algorithm improved only the AP of the model by 0.46%. The reason for the low AP improvement effect is that there are few ship samples with dense arrangements in the data set, and Soft-RNMS yields a significant improvement in the detection of ship targets with high overlap rates, so the improvement for these data is small. If the number of densely arranged ship targets in the data set increases, the algorithm can achieve better results. By merging the methods mentioned above, the final AP of the model is increased to 76.24%.

IV. DISCUSSION

The performance at detecting SAR ship targets is affected by many factors, including the image resolution, polarimetry, sea surface conditions, wind speed, ship size and ship direction [52]. In this paper, the ship size, complex background and ship direction are mainly considered. However, high wind speeds and poor ocean conditions can create turbulent water and produce volumetric scatter, which complicate the environment around ship targets [53]. In future work, we will try to combine the target information with complex environmental information to achieve ship target detection in some special scenarios. In addition, by combining continuous multi-frame images, the motion state of the target as the reference information for identification will also be considered in future research.

This paper has conducted some research on training from scratch. Reference [54] noted that although pretraining can accelerate convergence, the same effect can be achieved by

TABLE 6. The detection performances of different methods.

Detection model	Bounding box	Framework	AP (%)	Time (ms)
RFB Net	horizontal	one stage	71.22	28.3
YOLO v3	horizontal	one stage	71.90	27.6
YOLO v3 +Soft-NMS	horizontal	one stage	72.30	27.6
Attention-ResNet	horizontal	one stage	75.20	31.3
Faster R-CNN	horizontal	two stages	70.35	73.2
Faster R-CNN+FPN	horizontal	two stages	80.06	94.3
YOLO v3	rotation	one stage	73.15	34.2
Attention-ResNet	rotation	one stage	76.40	39.6
RRPN	rotation	two stages	74.82	316.0
MSARN	rotation	one stage	**76.24**	**35.4**
R^2CNN	rotation	two stages	80.26	210.8
R-DFPN	rotation	two stages	83.44	370.5

TABLE 7. The impact of each base role on performance.

Method	Rotation anchor	MSAR module	Soft-RNMS	Data augmentation	AP (%)
Base model	√	√	√	√	76.24
	√	√		√	75.78
		√		√	73.66
	√			√	72.46
				√	70.80

adopting a proper normalization method and a sufficient number of iterations. Therefore, this paper performed training under the condition that the network parameters are randomly initialized. Batch normalization and Leak ReLU are added after each convolutional layer to accelerate network convergence and prevent overfitting. We compared the training effects of the ImageNet pretraining weights and random initialization weights and found that using the ImageNet pretraining weights could not achieve the same performance as ordinary optical image detection in SAR image object detection tasks. This limitation may be due to the sensitivity of SAR ship targets to spatial location information. In addition, training from scratch has application potential in cross-domain scenarios, such as medical and multispectral imaging [55], [56]. Similarly, for SAR images, after carefully designing the network, training from scratch may achieve better results.

V. CONCLUSION

This paper proposes a detection model for multiscale and arbitrary-oriented SAR ship in complex scenarios. The model combines the precise positioning advantages of rotation detection and the speed advantage of a single-stage detection framework, effectively overcoming issues related to the complexity of application scenarios, the difficulty of dense target detection, the redundancy of detection areas and the diversity of target scales in SAR ship detection. In particular, we propose a multiscale adaptive calibration network, MSARN, to calibrate the multiscale features extracted from the network through global information, such that the network has both robust expression ability and accurate positioning ability. In addition, we design many new methods for the model, including a pyramid anchor box, loss function for rotation detection, and Soft-RNMS algorithm, and verify the effectiveness of these methods at SAR ship detection. Compared with other rotation detection models, the proposed detection method has an absolute speed advantage, and the inference time of a single image is only 35 ms, which is close to real-time detection. The continuous development of SAR technology will enable us to obtain more high-quality data, which will strongly promote the application of deep learning algorithms in the field of SAR image processing.

REFERENCES

[1] L. Zhai, Y. Li, and Y. Su, "Inshore ship detection via saliency and context information in high-resolution SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1870–1874, Dec. 2016.

[2] 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.

[3] Y. Liu, M.-H. Zhang, P. Xu, and Z.-W. Guo, "SAR ship detection using sea-land segmentation-based convolutional neural network," in *Proc. IEEE Int. Workshop Remote Sens. Intell. Process.*, Shanghai, China, May 2017, pp. 1–4.

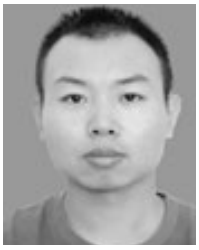
[4] X. Leng, K. Ji, X. Xing, S. Zhou, and H. Zou, "Area ratio invariant feature group for ship detection in SAR imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2388–2736, Jul. 2017.

[5] S. Wang, M. Wang, S. Yang, and L. Jiao, "New hierarchical saliency filtering for fast ship detection in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 351–362, Jan. 2017.

[6] N. Liu, Z. Cao, Z. Cui, Y. Pi, and S. Dang, "Multi-scale proposal generation for ship detection in SAR images," *Remote Sens.*, vol. 11, no. 5, pp. 526–546, Mar. 2019.

[7] D. Xiang, T. Tang, Y. Ban, and Y. Su, "Man-made target detection from polarimetric SAR data via nonstationarity and asymmetry," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 4, pp. 1459–1469, Apr. 2016.
[8] 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.
[9] 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.
[10] 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.
[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.
[13] 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. (CVPR)*, Boston, MA, USA, Jun. 2016, pp. 779–788.
[14] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525.
[15] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767
[16] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 385–400.
[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.
[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 740–755.
[19] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, pp. 860–874, 2017.
[20] J. Jiao, Y. Zhang, H. Sun, X. Yang, X. Gao, W. Hong, K. Fu, and X. Sun, "A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection," *IEEE Access*, vol. 6, pp. 20881–20892, 2018.
[21] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. BIGSARDATA*, Beijing, China, Nov. 2017, pp. 1–6.
[22] 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.*, Shanghai, China, May 2017, pp. 1–4.
[23] Z. Deng, H. Sun, S. Zhou, and J. Zhao, "Learning deep ship detector in SAR images from scratch," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 4021–4039, Jun. 2019.
[24] 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.
[25] L. Wang, J. Peng, and W. Sun, "Spatial–spectral squeeze-and-excitation residual network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 7, pp. 884–900, Apr. 2019.
[26] S. Zhang, R. Wu, K. Xu, J. Wang, and W. Sun, "R-CNN-based ship detection from high resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 6, pp. 631–645, Mar. 2019.
[27] 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, doi: 10.1109/ACCESS.2019.2930939.
[28] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.
[29] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2CNN: Rotational region CNN for orientation robust scene text detection,"Jun. 2017, *arXiv:1706.09579*. [Online]. Available: https://arxiv.org/abs/1706.09579
[30] Z. Liu, J. Hu, L. Weng, and Y. Yang, "Rotated region based CNN for ship detection," in *Proc. IEEE Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 900–904.
[31] M. Li, W. Guo, Z. Zhang, W. Yu, and T. Zhang, "Rotated region based fully convolutional network for ship detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Valencia, Spain, Jul. 2018, pp. 673–676.
[32] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, "Automatic ship detection in remote sensing images from Google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, pp. 132–145, Jan. 2018.
[33] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," 2019, *arXiv:1809.02165*. [Online]. Available: https://arxiv.org/abs/1809.02165
[34] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 3588–3597.
[35] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7794–7803.
[36] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7151–7160.
[37] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu Apr, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," 2019, *arXiv:1904.11492*. [Online]. Available: https://arxiv.org/abs/1904.11492
[38] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, pp. 4278–4284.
[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Lille, France, Jul. 2015, pp. 448–456.
[40] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 7132–7141.
[42] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 3–19.
[43] 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. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 936–944.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Amsterdam, The Netherlands, Oct. 2016, pp. 770–778.
[45] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *IEEE Access*, vol. 6, pp. 50839–50849, 2018.
[46] D. A. Plaisted and J. Hong, "A heuristic triangulation algorithm," *J. Algorithms*, vol. 8, no. 3, pp. 405–437, 1987.
[47] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–improving object detection with one line of code," in *Proc. Int. IEEE Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5561–5569.
[48] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," Mar. 2016, *arXiv:1603.04467*. [Online]. Available: https://arxiv.org/abs/1603.04467
[49] Z. Zhang, T. He, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of freebies for training object detection neural networks," 2019, *arXiv:1902.04103*. [Online]. Available: https://arxiv.org/abs/1902.04103
[50] Z. Huang, Y. Zhang, Q. Li, T. Zhang, N. Sang, and H. Hong, "Progressive dual-domain filter for enhancing and denoising optical remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 759–763, May 2018.
[51] Z. Huang, L. Chen, Y. Zhang, Z. Yu, H. Fang, and T. Zhang, "Robust contact-point detection from pantograph-catenary infrared images by employing horizontal-vertical enhancement operator," *Infr. Phys. Technol.*, vol. 101, pp. 146–155, Sep. 2019.
[52] B. Tings, C. Bentes, D. Velotto, and S. Voinov, "Modelling ship detectability depending on TerraSAR-X-derived metocean parameters," *CEAS Space J.*, vol. 11, no. 1, pp. 81–94, Oct. 2018.

[53] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sens.*, vol. 11, no. 7, pp. 765–778, Mar. 2019.
[54] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," 2018, *arXiv:1811.08883*. [Online]. Available: https://arxiv.org/abs/1811.08883
[55] Y. Li, J. Li, W. Lin, and J. Li, "Tiny-DSOD: Lightweight object detection for resource-restricted usages," 2018, *arXiv:1807.11013*. [Online]. Available: https://arxiv.org/abs/1807.11013
[56] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," in *Proc. Int. IEEE Conf. Comput. Vis. (ICCV)*, Venice, Italy, Jun. 2017, pp. 1919–1927.



CHEN CHEN received the B.E. degree in measurement and control technology and instrumentation from the School of Measurement and Control Technology and Communication Engineering, Harbin University of Science and Technology, Harbin, China, in 2017. He is currently pursuing the M.S. degree with the Department of Automation, Xi'an Institute of High-Tech, Xi'an, China. His research interests include synthetic aperture radar image interpretation, pattern recognition, and deep learning.

CHUAN HE received the B.S., M.S., and Ph.D. degrees from the High-tech Institute of Xi'an, China, in 2008, 2010, and 2015, respectively. He is currently an Associate Professor with High-tech Institute of Xi'an. He has authored or coauthored two books and more than ten articles. His research interests include image processing, videometrics, machine learning, synthetic aperture radar image interpretation, pattern recognition, and deep learning.

CHANGHUA HU received the B.S. degree in control science and engineering and the master's degree in instrument science and technology from the High-Tech Institute of Xi'an, Xi'an, China, in 1987 and 1990, respectively, and the Ph.D. degree from Northwestern Polytechnical University, Xi'an, in 1996. He was a Visiting Scholar with the University of Duisburg, Duisburg, Germany, from September 2008 to December 2008. He is currently a Cheung Kong Professor with the Xi'an Institute of High-Tech. He has authored or coauthored two books and about 100 articles. His research interests include fault diagnosis and prediction, life prognosis, and fault tolerant control.

HONG PEI received the B.S. degree from the Hefei University of Technology, in 2014, and the master's degree from the Xi'an Institute of High Technology, in 2016, where he is currently pursuing the Ph.D. degree. His research interests include prognostics and health management, predictive maintenance, and lifetime estimation.

LICHENG JIAO (SM'89–F'17) received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively. From 1990 to 1991, he was a Postdoctoral Fellow with the National Key Laboratory for Radar Signal Processing, Xidian University, Xi'an. Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China. His research interests include image processing, natural computation, machine learning, and intelligent information processing. He has led 40 major scientific research projects, and published more than 20 monographs and a hundred articles in international journals and a hundred articles in international journals and conferences. Dr. Jiao is a member of IEEE Xi'an Section Executive Committee and the Chairman of Awards and Recognition Committee, vice board chairperson of Chinese Association of Artificial Intelligence, councilor of Chinese Institute of Electronics, committee member of Chinese Committee of Neural Networks, and expert of Academic Degrees Committee of the State Council.