Ships Detection in SAR Images Based on Anchor-Free Model With Mask Guidance Features

Haicheng Qu, *Member, IEEE*, Lei Shen, Wei Guo, and Junkai Wang

***Abstract*—Ship targets in synthetic aperture radar (SAR) images have various scales. The detection model based on anchor boxes requires manual design of candidate boxes, which are fixed and cannot completely match all kinds of targets. Instead, large of anchor boxes with different sizes also result in large amounts of computing resources being consumed. Another potential issue comes from complex background information of near-coast scenes, which leads to ship targets being unrecognized because the background contains similar appearing objects. Therefore, this article proposes an anchor-free detection model based on mask guidance features, which achieves detection mainly through three modifications. First, feature maps of multiple scales are fused to obtain high-resolution feature maps containing rich semantic information. Second, a transformer encoder module is introduced to focus on the context relationship between the target object and the global image and to enhance the dependence between ship targets. Third, the mask guide feature is used to highlight the positions of the target in the feature map, and a loss function in the mask guide mechanism is designed to optimize the mask feature map to reduce false detections and missed detections. Testing the model on the public dataset SAR ship detection dataset, the model's detection accuracy reached 96.17%, with its accuracy on small-size ships reaching 96.11% and 97.84% on large ships.**

***Index Terms*—Anchor-free, mask guide, position enhancement, ship detection, transformer.**

I. Introduction

SYNTHETIC aperture radar (SAR) is an active microwave sensor that can obtain high-resolution images like optical images in environments with extremely poor visibility. SAR imaging is widely used for guiding missiles and observation in the military field [1], [2], as it can ensure all-weather monitoring. Ship target detection has strategic significance in this field; therefore, accurate target detection using SAR imaging is currently a research hotspot. SAR imaging mechanisms differ from those of optical imaging, resulting in a different style of targets. SAR image targets are shown as blurred shapes composed of bright spots, with coherent speckle noise [3]. The scattering of objects with similar target shapes, such as islands, reefs, and land buildings, also complicate the background of SAR images, making interpretation more difficult.

Traditional SAR image target detection algorithms are composed mostly of constant false alarm rate (CFAR [4]) algorithms and derived methods [5], [6]. CFAR conducts statistical modeling of marine clutter and disturbance factors to determine a threshold. From this, CFAR determines the presence of a target through its contrast from the background. The truncated statistics log normal CFAR (TS-LNCFAR) algorithm proposed by Ai *et al.*, [7] utilizes adaptive clutter truncation statistics to construct more accurate parameter estimations for truncated clutter; this is done by removing high-strength heterogeneous points in local sliding windows through adaptive thresholds. However, traditional SAR image detection algorithms rely heavily on hand-designed anchor boxes and have weak generalization abilities.

The application of deep learning technology [8] in the target detection field promotes the development of SAR based detection. Li *et al.* [9] constructed the first publicly available SAR image data set [SAR ship detection dataset (SSDD)], and applied Faster R-CNN to SAR image ship detection through feature fusion and migration learning. Chen *et al.* [10] used the GAN (Generative Adversarial Networks) to generate enough training samples, and used the YOLO-v2 model to improve the detection effect of small-scale ships. Lin Ty *et al.* [11] proposed a single-stage detection model Retinanet, which can obtain faster detection speed by simplifying the model. At the same time, in order to solve the problem of uneven distribution of positive and negative samples, focal loss [11] is proposed. The detection model based on the anchor frame has higher requirements for the performance of the machine. The design of the anchor frame is for a specific data set, and the generalization ability is low. Sun *et al.* [12] proposed an improved model based on FCOS [13], using the category-position module to optimize the location and regression branch features in the network. At the same time, the classification method of the target and the bounding box regression method were redesigned to solve the influence of the fuzzy area. In 2017, Google proposed the transformer model [14], which simplified the training process by using encoding and decoding methods, with fewer training parameters. In the transformer encoder module, feature map generated by fusion is transformed into feature sequence with length $H \times W$ and width C as the input of transformer encoder module, which can solve the problem of inflexible segmentation of input feature map and makes the input image more flexible. The self-attention module ensures that the transformer encoder always maintains the global

Manuscript received September 6, 2021; revised November 21, 2021; accepted December 12, 2021. Date of publication December 21, 2021; date of current version January 7, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42071351 and in part by the Foundation of Liaoning Educational Committee under Grant LJ2019JL010. *(Corresponding author: Lei Shen.)*

The authors are with the School of Software, Liaoning Technical University, Huludao 125105, China (e-mail: quhaicheng@lntu.edu.cn; shenlei95821@163.com; guow9966@163.com; wangjunkailntu@163.com).

Digital Object Identifier 10.1109/JSTARS.2021.3137390

receptive field, while adding a context encoding module to capture global context information. The transformer encoder module generates feature sequences with global receptive fields and context information and converts them into feature maps (length:H, width:W, height:C) as the input of the next stage.

Ship targets comprise a minimal part of the whole SAR image and have varying scales. The ship detection models based on anchor boxes rely overly on artificially designed bounding boxes. Additionally, hyperparameters, such as the number of anchor frames and size rate, consume a lot of computing resources. Another problem comes from the complexity of image backgrounds, which can cause target-like objects to influence the model's learning, resulting in a high false alarm rate. In response to the above problems, this article proposes a transformer encoding module and a mask guidance module. The coding module enables the model to learn the dependencies between ship targets and allows it to act on the feature maps that contain rich dependencies after feature fusion. To solve the issue of a high false alarm rate caused by complex backgrounds, this article designs a truth-value mask guidance mechanism. The mask module generates a mask feature map ranging from 0 to 1 and updates the feature map at each pixel by using the weight of the mask. This method can improve the weight of the target area, reduce the weight of the non-target area, and better distinguish the target area from the nontarget area. Finally, the prediction mask and truth mask are optimized by mask loss. The public remote sensing data set SSDD verifies that the method achieves ideal detection results in multi-scale targets and complex backgrounds.

II. Related Works

In recent years, anchor-free detection models such as FCOS [13], CornerNet [15], CenterNet [16], and ExtremeNet [17] have become popular. These predict corners and center points pixel by pixel and then group them into candidate frames. The detection model without anchor frame has also achieved good performance in the field of remote sensing. Our team performed contextual information fusion in the FCOS model, fusing semantic information and spatial location information into the feature map. This method combined with the context information, refined the ship target features in the feature map, and achieved better detection results.

Wang proposed a mask to guide attention maps [18], which performs well in the instance segmentation field. Masks are also occasionally used to enhance ship position information in the ship detection field and to eliminate the influence of complex backgrounds and ocean clutter noise. To accomplish this, a mask feature map is used to enhance the foreground and background information. The mask feature map is also used to optimize the prediction mask feature map through cross-entropy loss with the truth mask. As it is difficult to train a classifier on a small dataset like the SAR image dataset, the mask is used to guide the high-resolution feature map that integrates semantic information.

The use of the transformer model in machine translation improves the RNN [19] training speed, enhances the transfer of information between the encoding and decoding modules, and obtains richer context information. Srinivas *et al.* [20] proposed a transformer-based backbone, which generates feature maps with low resolution from large scale images. Additionally, this model uses information from the global self-attention aggregation feature map to achieve good results in many fields, such as target detection [21], instance segmentation [22], image super-resolution [23], and image generation [24]. Zhu proposed the detection transformer (DETR [25]), which through self-attention in encoding and decoding, can focus on all targets in the world frame and achieve end-to-end detection. Facebook has also applied the transformer in the target detection field.

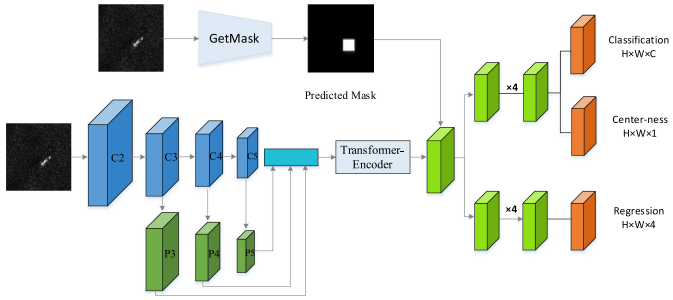

Fig. 1. Network architecture of our method.

III. Method

To highlight the target's position information in a complex background and reduce the amount of calculation in the bounding boxes generation process, this article proposes an anchor-free detection model with mask guidance features. As seen in Fig. 1, the proposed detection model is divided into three stages. The process starts after Resnet101 [26] has extracted the features, with each stage containing multiple residual blocks. Each stage activates the feature output of the last residual block as the current feature map. The output feature map is then sent to the encoder module, where the multihead attention learns the position information for each target while using multiple iterations to enhance the dependence between targets. Following this, the bounding boxes are generated pixel by pixel in the feature map, and the feature maps are then guided by the mask. The target position information is enhanced twice, and the mask feature maps are optimized through the loss function. Finally, the feature map containing rich semantic information [27] and location information [28] is inputted into the detection module to obtain the final detection result.

A. Anchor-Free Boxes Detection Model

Ship targets constitute a small percentage of the whole SAR image, and the detection model with anchor boxes relies on artificially designed candidate frames. Unsuitable bounding boxes directly affect the model's detection abilities. Additionally, the generation of anchor boxes requires a lot of computing resources. As shown in Fig. 2, in the FCOS model, any position (x, y) in the feature map is mapped to the receptive area of the original image $(\lfloor \frac{s}{2} \rfloor + sx, \lfloor \frac{s}{2} \rfloor + sy)$, and all pixels mapped to the bounding boxes are regarded as positive samples $(x_{\min}, y_{\min})$

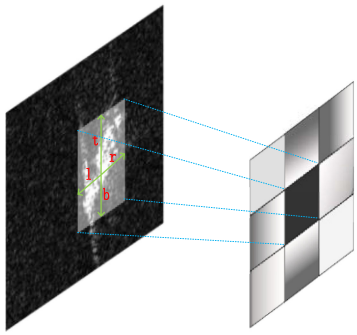

Fig. 2. Location mapping.

and $(x_{\max}, y_{\max})$ are the coordinates of the upper left corner and the lower right corner of the candidate box, respectively. The offset formula is calculated as follows:

$$\begin{aligned} l &= x - x_{\min} \\ r &= x_{\max} - x \\ t &= y - y_{\min} \\ b &= y_{\max} - y \end{aligned} \tag{1}$$

where (l, r, t, b) are the offsets from the center point (x, y) to the candidate frame's edges. Since there is no anchor box, there is also no need to compare the intersection over union (IOU) of the candidate box and the truth box to determine the positive samples. Instead, only demand regression on each determined pixel and offset is required. In summary, the detection model without anchor boxes has fewer hyperparameters and computational resource consumption.

B. Transformer Encoder

The transformer encoder module provides long-distance and location-dependent feature extraction on the context, improving the dependence between ship targets. Using the image sequence as an input, the encoder generates the feature map with rich location-dependent features through the following components: a multihead attention mechanism [29], a regularization module, and a forward propagation module. The whole process is shown in Fig. 3.

As the encoder input is a sequence, the input feature map $f \in \Re^{H\times \mathrm{W}\times \mathrm{C}}$ must first be decomposed into a 1-D sequence C×HW. Following this, the position code and feature sequence obtained by sin, cos linear transformations are sent to the multihead attention module. The expression for position coding is as follows:

$$\mathrm{Posencoding} = \begin{cases} \sin\left(\mathrm{pos}/10000^{2i/d_{\mathrm{model}}}\right) \\ \cos\left(\mathrm{pos}/10000^{2i+1/d_{\mathrm{model}}}\right) \end{cases} \tag{2}$$

where i is the feature sequence vector andpos is the feature sequence position. Each dimension of the positional encoding corresponds to a sinusoid. The wavelengths form a geometric progression from 2π to 10000 2π [25]. Cos and sin functions have regular periodicity, so $PE_{(pos+k)}$ at any position can be represented by $PE_{(pos)}$. The characteristics of trigonometric

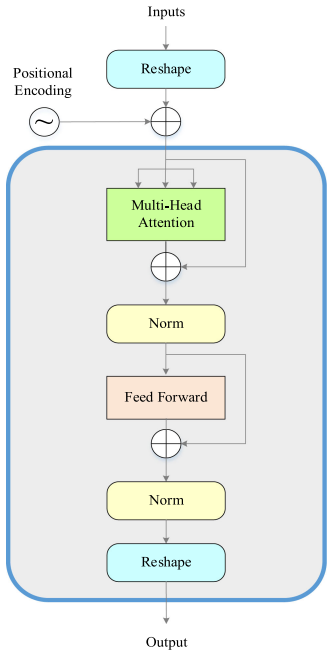

Fig. 3. Network architecture of the transformer encoder.

functions are as follows:

$$\cos(A+B) = \cos(A)\cdot\cos(\mathrm{B}) + \sin(A)\cdot\sin(B) \tag{3}$$

$$\sin(A+B) = \sin(A)\cdot\cos(B) + \cos(A)\cdot\sin(B)\,. \tag{4}$$

The multihead attention obtains the correlation between the targets and allows each target to contain the vector information of the others. In each iteration of the network, the position information between targets is continuously learned to increase its accuracy. The multihead attention expression is as follows:

$$\mathrm{Attention} = \mathrm{softmax}\left(\frac{QK}{\sqrt{d_k}}\right)V \tag{5}$$

$$Q, W, V = X\cdot W_Q, X\cdot W_W, X\cdot W_V \tag{6}$$

where W_Q, W_W, W_V are weight parameters, which are continuously learned during model training. The feature sequence generated by the multi-head attention is normalized to speed up the model's convergence. Finally, the feature sequence is inputted into the forward propagation network, where the dimension is extended to the input time scale.

C. Mask Guide Features

Because target detection often imitates human vision in the sense that obvious location information leads to simpler target recognition, objects with a similar shape to the target usually get a higher response in complex backgrounds. By contrast, the target in a complex background has a relatively poor response, resulting in a high false alarm rate and poor model detection. Additionally, feature coupling between coherent speckle noise and the ship targets can result in the boundary blur being targeted. However, the feature map guided by the mask effectively suppresses the background information, thereby highlighting

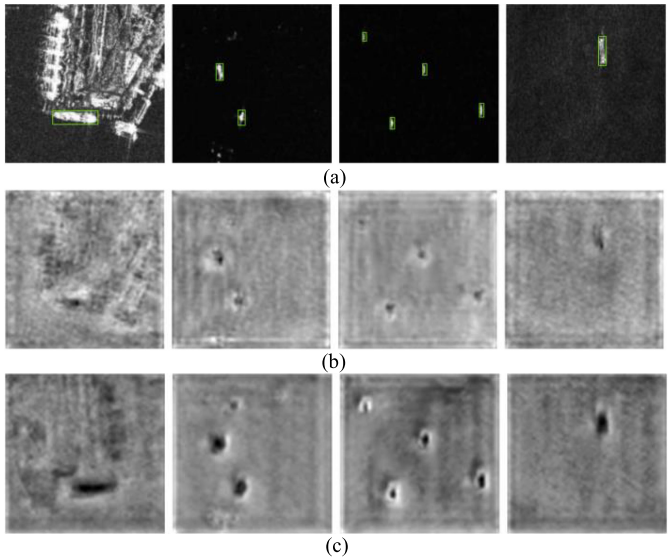

Fig. 4. Feature visualization.

the ship targets and improving the detection effect of the ship targets. The mask prediction process is divided into three steps. First, the receptive field of the input feature map is expanded [30]; second, the network nonlinearity is increased through a 1×1 convolution; third, the mask output is obtained through two parallel 1×1 convolutions, resulting in the obtained prediction mask and the truth mask of the labeled bounding boxes being optimized by mask loss. The optimization process is as follows:

$$\mathrm{L}_{\mathrm{mask}} = \mathrm{SCE}(m^*_{i,j}, m_{i,j}) \tag{7}$$

where $m^*_{i,j}$ is the predicted target mask, and $m_{i,j}$ is the true value target mask.

Fig. 4 shows the visualization of $\{P_3, P_4, P_5\}$ feature maps fused to P_3 scale in TensorBoard, and compares the feature maps before and after mask guidance. It is evident that the location feature after mask guidance is more recognizable and that the target is more prominent against the complex background. Obviously, the feature map guided by the mask is beneficial to model detection and positioning of ship targets, thereby reducing the probability of false detection of background information as ship targets.

D. Loss Function

Small-scale ships in SAR images occupy a minimal area in the feature map, with the number of target areas and background areas differing heavily. Most of the bounding box consists of a background, making it likely for an imbalanced distribution of positive and negative samples to occur. To solve this problem, the focal loss was used to optimize the classification loss. Focal loss is defined as follows:

$$L_{\mathrm{cls}} = \begin{cases} -\alpha(1-p_{i,j})^{\gamma}\log(p_{i,j}) & p_{i,j} = 1 \\ -(1-\alpha)p_{i,j}^{\gamma}\log(1-p_{i,j}) & p_{i,j} \neq 1 \end{cases} \tag{8}$$

where $p_{(i,j)}$ is the category predicted at (i, j), γ is the adjustment coefficient used to reduce easily classifiable samples, and α is the balance factor between positive and negative samples. In accordance with Faster R-CNN, this model sets $\gamma = 2$ and $\alpha = 2.5$. The larger the deviation between the prediction box and the truth box, the less effective the model detection effect will be. To obtain a more ideal candidate frame, this article chooses IoU Loss to adjust the candidate frame. The regression loss is defined as follows:

$$L_{\mathrm{reg}} = -\log\left(\max\left(\frac{\mathrm{Inter}\left(b_{i,j}, b^*_{i,j}\right)}{\mathrm{Union}\left(b_{i,j}, b^*_{i,j}\right)}\right)\right) \tag{9}$$

where $b_{i,j}$ is the prediction candidate box, and $b^*_{i,j}$ is the truth box. Therefore, the total loss in this method can be represented as

$$L_{\mathrm{total}} = \frac{1}{N_{\mathrm{all}}}\sum_{i,j} L_{\mathrm{cls}}(p_{i,j}) + \frac{1}{N_{\mathrm{pos}}}\sum_{i,j\in\mathrm{pos}} L_{\mathrm{reg}}(b_{i,j}, b^*_{i,j}) + \frac{1}{N_{\mathrm{all}}}\sum_{i,j} L_{\mathrm{mask}}(m_{i,j}, m^*_{i,j}). \tag{10}$$

TABLE I
BASIC INFORMATION FOR THE SHIP DATASETS

Dataset	Sensors	Resolution	Polarization	Image size
SSDD	RadarSat-2、TerraSAR-X、Sentinel-1	1m-15m	VV、HH、VH、HV	Min:390×205 Max:600×500
SAR-Ship Dataset	GF-3、Sentinel-1	3m、5m、8m、10m	VV、HH、VH、HV	256×256

IV. DATASET

For model training and data testing, the public data set SSDD [31] was adopted. The data in the SSDD dataset are mainly acquired by RadarSat-2, TerraSAR-X, and Sentinel-1 sensors, with a resolution of 1–15 m. In addition, this article also uses the SAR-Ship-Dataset data set to verify the effectiveness and robustness of the model in this article. Table I shows the specific parameters of the dataset. The SSDD dataset contains 1160 images, including 2456 ship targets, with near-coast, far-sea, and ocean clutter and noise scenes. Ship sizes range from 7×7 pixels to 211×298 pixels. In the experiment, all the images are resized into 500×500 pixels. The SAR-Ship-dataset has 43 819 images in total, where the dataset includes 59 535 ship targets. The data in the SAR-Ship dataset are mainly acquired by GF-3 and Sentinel-1 satellites. The incident angle is minimum 19° and maximum 50°, and the swath range is from 30 to 250 km. In this article, the data sets is divided into the training set and test set at a ratio of 7:3.

The number of training samples in the SSDD data set is insufficient. In order to improve the robustness and generalization ability of the model, the SSDD data set was enhanced by adding Gaussian noise, changing image brightness and flipping images. The final training set has 3248 images. In the image flipping process, we choose horizontal flipping, through which we obtain the new image, and then update the data of the annotation bounding box to the new XML file. When we get an image with Gaussian noise or changes brightness, we do not need to update

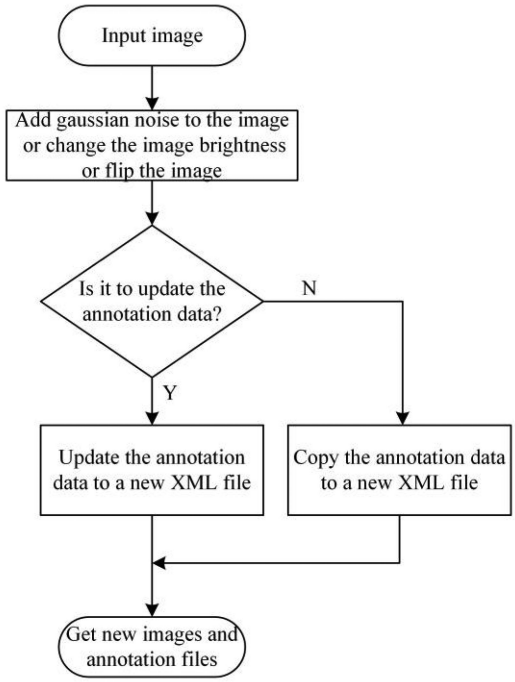

Fig. 5. Data enhancement flowchart.

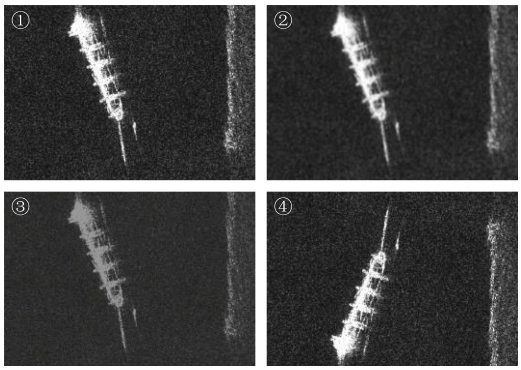

Fig. 6. Image enhancement.

the bounding box data, just copy the data to a new XML file. Fig. 5 shows the process of image enhancement. Fig. 6 shows the enhancement effect of images data. ① The original image. ② Add Gaussian noise to the image. ③ Change the brightness of the image. ④ Flip the image.

V. Experiments

In order to ensure the fairness and consistency of the experiment, all experiments were completed under the same environment configuration. Table II shows the hardware configuration and deep learning environment of the experiment in this article. The following variables are also specified: the learning rate of the model is 0.0005(Every 40K iterations, the learning rate is reduced to 1/10 of the original), the decay weight is set 0.00001, and the momentum is set 0.9. The Resnet101 model is selected as the feature extraction network and the pre-trained parameters from the ImageNet training set is also migrated to this model.

TABLE II
Experimental Environment Configuration

Project	Model/Parameter
System	Ubuntu 16.04 LTS
CPU	i7-7700
GPU	NVIDIA GeForce GTX 1080Ti 11G
Code	Python2.7
Framework	CUDA8.0/cudnn5.0/tensorflow-gpu1.4.0

TABLE III
Comparison of Different Backbone Networks

-	Resnet50	Resnet101	Resnet152
conv1	7×7,64		
	3×3,max pool		
conv2_x	$\begin{Bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{Bmatrix}\times3$	$\begin{Bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{Bmatrix}\times3$	$\begin{Bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{Bmatrix}\times3$
conv3_x	$\begin{Bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{Bmatrix}\times4$	$\begin{Bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{Bmatrix}\times4$	$\begin{Bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{Bmatrix}\times8$
conv4_x	$\begin{Bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{Bmatrix}\times6$	$\begin{Bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{Bmatrix}\times23$	$\begin{Bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{Bmatrix}\times36$
conv5_x	$\begin{Bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{Bmatrix}\times3$	$\begin{Bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{Bmatrix}\times3$	$\begin{Bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{Bmatrix}\times3$
FLOPs	3.8×10^9	7.6×10^9	11.3×10^9
params	25,610,269	44,654,608	60,344,387
Model size	98MB	170MB	230MB

A. Evaluation Criteria

To objectively evaluate the effectiveness of the model in SAR image ship target detection, the evaluation indicators recall (R), precision (P), and average precision (AP) were used

$$R = \frac{TP}{TP + FN} \tag{11}$$

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$\mathrm{AP} = \int_0^1 P(R)d(R) \tag{13}$$

where *TP* is the number of ships correctly detected, *FN* is the number of ships missed to be detected, and *FP* is the number of ship targets that are falsely detected.

B. Evaluation of Different Backbone Networks

As shown in Table III, the Resnet50 [32] backbone network contains four layers {conv2_x, conv3_x, conv4_x, conv5_x},

TABLE IV
DETECTION PERFORMANCE OF DIFFERENT BACKBONE NETWORKS (%)

Backbone	Recall	Precision	Average Precision
Resnet50	96.43	90.33	95.85
Resnet101	**97**	91.26	**96.17**
Resnet152	94.27	**91.92**	92.10

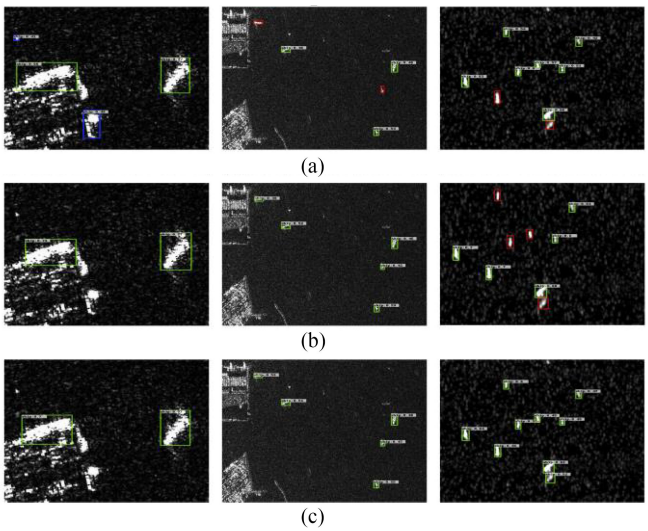

Fig. 7. Test results for different backbone networks. The green rectangles are the true detection result. The blue and red rectangles represent the false detection and missing ships (a) is the Resnet50 detection effect diagram, (b) is the Resnet152 detection effect diagram, and (c) is the Resnet101 detection effect diagram.

which respectively contain (3, 4, 6, 3) residual blocks, and a total of 50 layers of convolution. He *et al.* [33] proposes that reasonably deepening the network depth can achieve better detection results. Compared with Resnet 50, Resnet101 [26] contains 101 layers of convolution, which can extract richer features of semantic information and avoid the suppression of candidate boxes with high IOU scores and low classification confidence. However, [34] and [35] come to the same conclusion that on the premise of deep network convergence, the accuracy rate will become saturated or even decrease with the increase of network depth. This is because feature maps tend to lose the location information of small targets in the process of convolution, resulting in the model missing targets. Although Resnet50 has fewer parameters and model sizes, it is prone to misdetection when inputting feature maps that lack semantic information into the detection model. The small target location information extracted by Resnet152 is seriously lost, and too many parameters and floating-point operations per second reduce the model detection speed. Therefore, in order to obtain better detection effect, Resnet101 is selected as the backbone network in this article. It can be seen from Table IV that using Resnet101 as the backbone network, the average accuracy of the detection model reaches 96.17%. Compared with Resnet50 and Resnet152, Resnet101 has better detection effect.

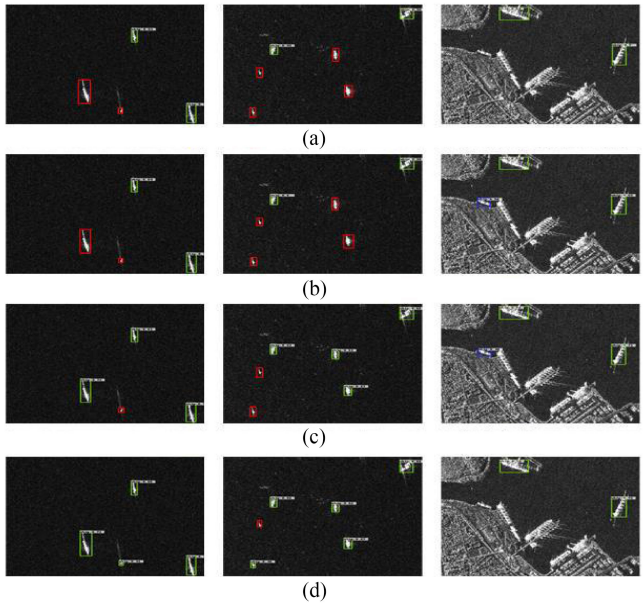

Fig. 8. Test results of different scales. The green rectangles are the true detection result. The blue and red rectangles represent the false detection and missing ships. (a) and (b) are the detection effects for $P4$ and the $P4$ fused feature maps, respectively; and (c) and (d) are the detection effects for $P3$ and the $P3$ fused feature maps, respectively.

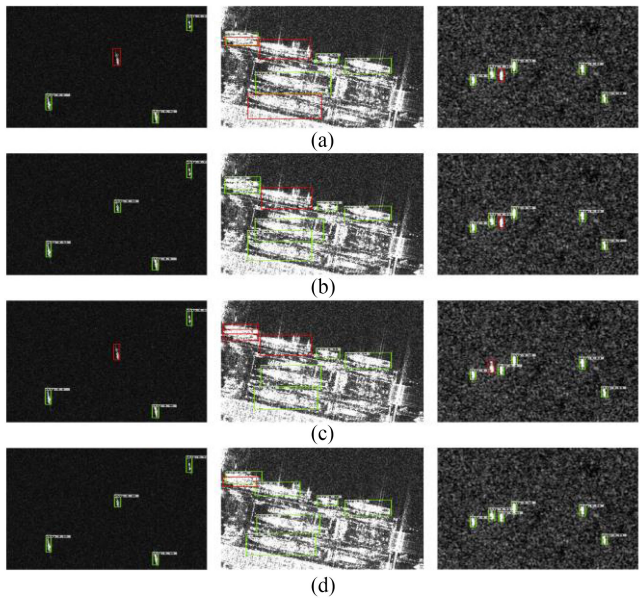

Fig. 9. Test results of different modules. (a) is the baseline, (b) contains the mask guide feature map, (c) contains the coding module, and (d) is the final model.

Fig. 7 shows the detection effect diagrams of different backbones, where (a) is the Resnet50 detection effect diagram, (b) is the Resnet152 detection effect diagram, and (c) is the Resnet101 detection effect diagram. To a certain extent, deepening the network can result in richer features being extracted, along with richer semantic and spatial location information from the feature map after fusion.

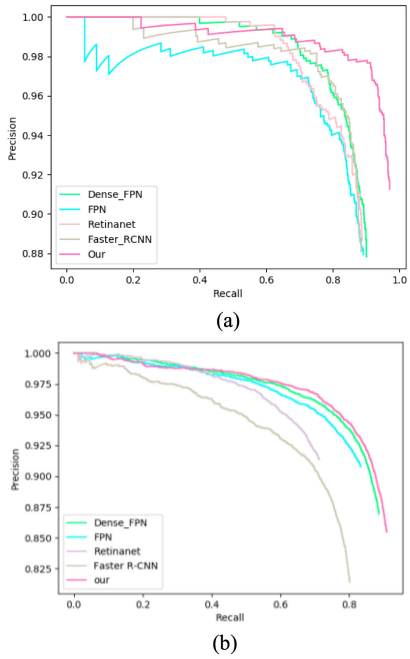

Fig. 10. P-R diagrams of different methods.

TABLE V
DETECTION PERFORMANCE OF DIFFERENT SCALES (%)

Method	Recall	Precision	Average Precision
P_4	89.93	79.06	87.92
Fusion to P_4 size	89.94	86.73	88.61
P_3	96.82	**94.18**	96.14
Fusion to P_3 size	**97**	91.26	**96.17**

C. Evaluation of Different Scales

All feature maps contain different features, with high-level maps owning rich semantic information, and low-level maps owning rich spatial location information. After passing through a layer of the feature extraction network, the resolution of the feature map becomes half the size of the original, resulting in small-scale ships losing their position after multiple feature extractions. This model fuses the $\{P_3, P_4, P_5\}$ scale feature maps, resulting in more balanced semantic and spatial location information for the generated feature maps. This section compares the detection effect of the $\{P_3, P_4\}$ scales on their own and when fused with a feature map. Table V shows the test results.

As shown in Fig. 8(a) and (b) are the detection effects for P_4 and the P_4 fused feature maps, respectively; and (c) and (d) are the detection effects for P_3 and the P_3 fused feature maps, respectively. Fig 8 shows that small-scale targets are often lost in low-resolution maps during the detection process and that the false alarm rate of fused feature maps is lower than that of unfused maps.

TABLE VI
DETECTION PERFORMANCE OF DIFFERENT MODELS (%)

Mask	Encoder	Recall	Precision	Average Precision
×	×	89.94	**94.26**	89.23
√	×	96.05	92.86	95.33
×	√	94.52	91.15	93.49
√	√	**97.00**	91.26	**96.17**

D. Evaluation of Different Models

As mentioned, small-scale targets in the feature map may experience information loss after multilayer convolution processing, and its position in a complex background can go undetected. To mitigate this, the mask guide feature can selectively activate the target information while reducing the activation of background information, making the target information more prominent in the feature map and increasing its contrast with the background information to make it more obvious in the detection process. In image interpretation, there are higher requirements for this relationship. The transformer increases dependency between targets through the multi-head attention module. Through model iteration, the transformer learns the position code for each target and enhances the contextual relationship between the target and the feature map. It can be seen in Table VI that the feature map generated by the mask guidance provides more prominent position information, improves the model detection effect, and increases the average accuracy by 6.1% compared to the original model. Through the encoding module in the transformer, the correlation between the targets is increased, and the average accuracy is improved by 5.09% compared to the original model. Under the simultaneous actions of the two modules, the model effect recall rate reached 97.00%, the precision rate reached 91.26%, and the average precision reached 96.17%.

Fig. 9 shows the performance of each model, where (a) is the baseline, (b) contains the mask guide feature map, (c) contains the coding module, and (d) is the final model. It can be seen from (b) that the feature map sent to the detector provides rich semantic and spatial position information to obtain a better detection effect. The learned position-coding information is also important, as it guides the detector to the target position, as seen in (c).

E. Evaluation of Different Methods

To validate the effectiveness of the proposed method, Faster R-CNN [36], Retinanet [11], FPN [37], Dense_FPN [38] and FBR-Net [39] were applied to the SSDD. To summarize the methods, the Faster R-CNN model obtains its results through two candidate box optimizations; the Retinanet model discards the rough adjustment stage and streamlines the model; FPN uses a top-down feature fusion mechanism and unifies the number of channels to achieve multi-scale target detection; and Dense_FPN merges the features of each layer, and its feature maps contain features from other layer. It can be seen from Table VII that under the average accuracy index, the model in this article is

TABLE VII
DETECTION PERFORMANCE OF DIFFERENT METHODS ON SSDD

Method	Recall	Precision	Average Precision
Faster R-CNN	88.54	88.09	87.26
Retinanet	89.17	87.94	86.91
FPN	89.17	86.41	87.92
Dense-FPN	90.19	87.84	89.11
FBR-Net [39]	92.79	**94.01**	94.10
Our	**97.00**	91.26	**96.17**

TABLE VIII
DETECTION PERFORMANCE OF DIFFERENT METHODS ON SAR-SHIP-DATASET

Method	Recall	Precision	Average Precision
Faster R-CNN	80.24	81.39	76.35
Retinanet	72.17	90.62	71.03
FPN	83.40	**90.80**	81.38
Dense-FPN	88.71	86.95	86.50
Our	**92.35**	81.98	**89.79**

8.91% higher than Faster R-CNN, 9.26% higher than Retinanet, 8.25% higher than FPN, and 7.06% higher than Dense-FPN.

In order to prove the generalization and robustness of the model in this article, the model in this article and other methods are applied to the SAR-Ship-Dataset data set. The experimental results are shown in Table VIII.

Fig. 10 shows the different models' detection effects more intuitively by displaying their P-R diagrams. Fig. 11 shows the detection effect diagrams of the different models in the SSDD dataset, where (a) is Faster R-CNN, (b) is Retinanet, (c) is FPN, (d) is Dense-FPN, and (e) is the proposed method. Compared to the others, the proposed model produces the best detection effect in SAR image ship detection with multiple scales and backgrounds.

F. Evaluation of Different Scenes

Complex SAR image backgrounds can result in land buildings or ports affecting the detection of near-coast ship targets and the model marking objects that look similar to ship targets increase the false alarm rate. Furthermore, ship targets are easily obscured in a complex background, thus increasing the false detection rate. The scale of ship targets is changeable, although small-scale ships can often be lost in the feature extraction process. Large-scale ships, on the other hand, are prone to missing and inaccurately positioned bounding boxes. The test set is divided into four sections to account for different scenarios: near-coast, far-sea, and large-scale, small-scale targets within a 60×60 range. Table IX shows that the model has greatly mitigated missed detections and false alarms in the near-coast scene, and after the position-coding stage, it has a better detection effect on large-scale ship targets. Under the combined effects of coding and mask guidance, the model has greatly improved the detection of ship targets in different scales and scenes.

Fig. 12 shows the detection effects on the tested scenarios. The first line is test results of baseline. The second line is the test results of our method, where (a) is near-coast, (b) is far sea, (c) is large-scale targets, and (d) is small-scale targets. Compared with the original model, the proposed model can highlight target information in complex backgrounds, including ship targets obscured by near-coast objects. The detection of small-scale ship targets is also significantly improved.

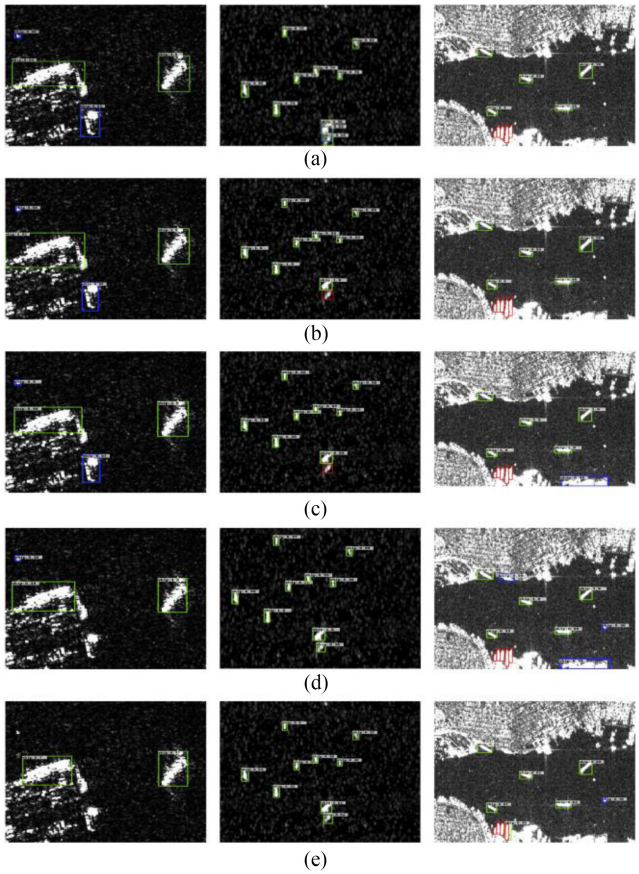

Fig. 11. Test results of different methods. The green rectangles are the true detection result. The blue and red rectangles represent the false detection and missing ships. (a) is Faster R-CNN, (b) is Retinanet, (c) is FPN, (d) is Dense-FPN, and (e) is the proposed method.

TABLE IX
DETECTION PERFORMANCE OF DIFFERENT SCENES (%)

Scenes	Method	Recall	Precision	Average Precision
inshore	Baseline	79.01	**87.07**	77.79
	Our	**87.65**	80.68	**85.55**
offshore	Baseline	92.78	96.01	92.24
	Our	**98.07**	**96.68**	**97.55**
Small scale target	Baseline	90.79	90.78	90.66
	Our	**96.9**	**91.85**	**96.11**
Large-scale target	Baseline	89.84	**94.65**	89.13
	Our	**98.68**	86.21	**97.84**

VI. CONCLUSION

An anchorless frame detection model based on mask-guided feature maps is proposed in this article to improve the ship detection performance in SAR images. The new method shows

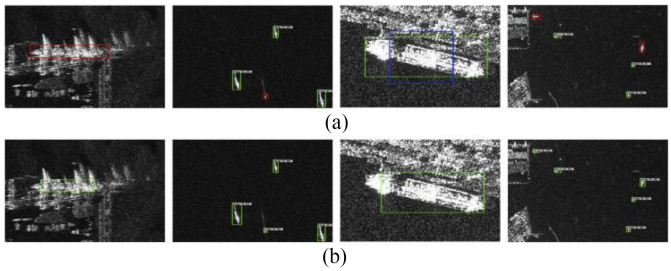

Fig. 12. Test results for different scenes. The green rectangles are the true detection result. The blue and red rectangles represent the false detection and missing ships. The first line is test results of baseline. The second line is the test results of our method. (a) is near-coast, (b) is far sea, (c) is large-scale targets, and (d) is small-scale targets.

ideal detection effects in different scenes and ship targets with different scales compared with the current popular detection models, including Faster R-CNN, Retinanet, FPN, Dense-FPN, and FBR-Net. The mask guide feature mechanism highlights the ships' target position in the feature maps affected by coastal object information and uses the loss function to optimize the accuracy of the target position information of the mask feature map. The coding module continuously iterates through the model to learn the position information of the ship targets, enhancing the dependency between targets, and resulting in better performance in large-scale ship targets detection. After secondary position information enhancement, the performance in small-scale ship targets detection is also improved. Additionally, the anchor-free detector removes unnecessary hyperparameters, avoids using inaccurate artificially designed bounding boxes, and shows better detection effects compared to models based on anchor boxes. The model in this article aims to optimize target detection results, and while missed detections and false detections are unavoidable, it has shown the ability to mitigate such outcomes. In summary, the proposed detection model can achieve better detection accuracy in a variety of backgrounds, such as near-shore, far sea, and among marine clutter. However, areas dense with ship targets can suppress candidate frames, resulting in some containing multiple targets. Horizontal candidate frames contain not only ship targets but also an excess of background information. In order to eliminate the influence of background information, we will apply rotating candidate boxes to the detection model in future works.

REFERENCES

[1] C. Belloni *et al.*, "SAR image dataset of military ground targets with multiple poses for ATR, target and background signatures III," *Int. Soc. Opt. Photon.*, vol. 10432, 2017, Art. no. 104320N.

[2] T. Zhang *et al.*, "Depthwise separable convolution neural network for high-speed SAR ship detection," *Remote Sens.*, vol. 11, no. 21, pp. 2483–2520, 2019.

[3] H. Yu, J. Gao, and A. Li, "Probability-based non-local means filter for speckle noise suppression in optical coherence tomography images," *Opt. Lett.*, vol. 41, no. 5, pp. 994–997, 2016.

[4] A. Avi and D. Roee, "CFAR detection algorithm for objects in sonar images," *IET Radar, Sonar Navigation*, vol. 14, no. 11, pp. 1757–1766, 2020.

[5] B. Magaz, A. Belouchrani, and M. Hamadouche, "Automatic threshold selection in OS-CFAR radar detection using information theoretic criteria," *Prog. Electromagnetics Res. B*, vol. 30, pp. 157–175, 2011.

[6] 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.

[7] J. Ai, X. Yang, J. Song, Z. Dong, L. Jia, and F. Zhou, "An adaptively truncated clutter-statistics-based two-parameter CFAR detector in SAR imagery," *IEEE J. Ocean. Eng.*, vol. 43, no. 1, pp. 267–279, Jan. 2018.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

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

[10] Z. Chen *et al.*, "Deep learning for autonomous ship-oriented small ship detection," *Saf. Sci.*, vol. 130, pp. 104812–104821, 2020.

[11] T. Y. Ross and G. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Venice, Italy, 2017, pp. 2980–2988.

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

[13] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 9627–9636.

[14] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, pp. 5998–6008, 2017.

[15] H. Law and D. J. Cornernet, "Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.

[16] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 6569–6578.

[17] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 850–859.

[18] J. Wang, X. Yu, and Y. Gao, "Mask guided attention for fine-grained patchy image classification," 2021, *arXiv:2102.02771*.

[19] H. I. Liu and W. L Chen, "Re-transformer: A self-attention based model for machine translation," *Procedia Comput. Sci.*, vol. 189, pp. 3–10, 2021.

[20] A. Srinivas, T. Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, 2021, pp. 16519–16529.

[21] L. Wang, J. Tang, and Q. Liao, "A study on radar target detection based on deep neural networks," *IEEE Sensors Lett.*, vol. 3, no. 3, Mar. 2019, Art. no. 7000504.

[22] Y. Wang *et al.*, "End-to-end video instance segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, 2021, pp. 8741–8750.

[23] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 5791–5800.

[24] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 1219–1228.

[25] X. Zhu *et al.*, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[26] Z. Xu, K. Sun, and J. Mao, "Research on ResNet101 network chemical reagent label image classification based on transfer learning," in *Proc. IEEE 2nd Int. Conf. Civil Aviation Saf. Inf. Technol.*, Weihai, China, 2020, pp. 354–358.

[27] H. Luo, P. Wang, H. Chen, and M. Xu, "Object detection method based on shallow feature fusion and semantic information enhancement," *IEEE Sensors J.*, vol. 21, no. 19, pp. 21839–21851, Oct. 2021.

[28] Y. Zeng, P. Zhang, Z. Lin, J. Zhang, and H. Lu, "Towards high-resolution salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 7234–7243.

[29] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," 2021, *arXiv:2102.12122*.

[30] F. Yu, V. Koltun, and T Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 472–480.

[31] Y. Wang *et al.*, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 765.

[32] X. Tian and C. Chen, "Modulation pattern recognition based on ResNet50 neural network," in *Proc. IEEE 2nd Int. Conf. Inf. Commun. Signal Process.*, Weihai, China, 2019, pp. 34–38.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.
[34] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards AI," *Large-Scale Kernel Mach.*, vol. 34, no. 5, pp. 1–41, 2007.
[35] K. Zhou *et al.*, "Understanding and resolving performance degradation in graph convolutional networks," 2020, *arXiv:2006.07107*.
[36] S. Ren *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 91–99, 2015.
[37] 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.*, Honolulu, HI, USA, 2017, pp. 2117–2125.
[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 4700–4708.
[39] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1331–1344, Feb. 2021.

Haicheng Qu (Member, IEEE) received the B.S. degree in computer science from Qingdao University of Technology, Qingdao, China, in 2005, the M.S. degree in computer application technology from Liaoning Technical University, Fuxin, China, in 2008, and the Ph.D. degree in information and communication engineering from Harbin Institute of Technology, Harbin, China, in 2016.

He is currently an Associate Professor with the School of Software, Liaoning Technical University. His research interests include remote sensing image rapid processing and intelligent big data processing.

Lei Shen received the B.S. degree in software engineering from Linyi University, Linyi, China, in 2019. He is currently working toward the M.S. degree in software engineering with Liaoning Technical University, Huludao, China.

His research interests include computer vision, pattern recognition, and SAR image processing, especially on object detection.

Wei Guo received the B.S. degree in computers and applications from Fuxin Mining Institute, Fuxin, China, in 1992, and the M.S. degree in computer application technology from Liaoning Technical University, Huludao, China, in 2005.

She is currently an Associate Professor with the School of Software, Liaoning Technical University. Her research interests include image and visual information computing and intelligent data processing.

Junkai Wang received the B.S. degree in computer science from the Northeastern University, Shenyang, China, in 2014. He is currently working toward the M.S. degree in software engineering with the School of Software, Liaoning Technical University, Huludao, China.

His research interests include deep learning, image processing, and object detection, especially on 3-D reconstruction.