IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 24 October 2022, accepted 10 November 2022, date of publication 14 November 2022,
date of current version 14 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3222364

RESEARCH ARTICLE

A Decoupled Head and Coordinate Attention Detection Method for Ship Targets in SAR Images

QINZUO LI[1,2], DENGJUN XIAO[1], AND FANGYING SHI[3,4]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[4] University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Dengjun Xiao (xiaodj@aircas.ac.cn)

This work was supported by the National Natural Science Foundation of China Youth Fund under Grant 61901442.

ABSTRACT Currently, deep learning-based synthetic aperture radar (SAR) image ship target detection methods have been widely used in the field of SAR image ship detection. However, these methods suffer from high model complexity and poor performance when detecting small dense targets. To address this problem, this paper proposes a ship target detection algorithm based on the improved YOLO (You Only Look Once) algorithm. In addition, considering the real-time requirements and computational constraints in mobile applications, the YOLOv4 network is modified to make it more lightweight. Moreover, decoupled head and coordinate attention are introduced to preserve YOLOv4's superb detection performance as much as possible after lightweighting it. First, as the detection head of the YOLOv4 degrades the performance, this study decouples the classification and regression tasks. Second, since the channel attention mechanism ignores the spatial position information, coordinate attention is used to obtain long-range dependencies and accurate position information in the spatial domain. Moreover, the effects of the coordinate attention mechanism in different hierarchical YOLOv4 structures are analyzed. Furthermore, on the basis of the YOLOv4 backbone, another lightweight backbone is added to the model structure to improve model detection performance. Experimental results on the SAR ship detection dataset (SSDD) and the high-resolution SAR images dataset (HRSID) demonstrate that the proposed method can achieve high detection accuracy in complex scenes. The proposed lightweight model has fewer parameters compared to the original YOLOv4 structure. Furthermore, two massive SAR images are used to confirm the proposed model's migration application performance. The experimental results demonstrate that the proposed model has a strong migration ability and can be used in maritime monitoring.

INDEX TERMS Ship detection, YOLO, coordinate attention, decoupled head, SAR.

I. INTRODUCTION

Compared with optical, infrared, and hyperspectral sensors, synthetic aperture radars are active microwave imaging sensors, which have an all-day and all-weather operational capability. In recent years, the number and quality of SAR images have greatly improved due to the advancements in SAR imaging technology. Therefore, the detection of ship targets has become a research hotspot. The traditional ship detection methods usually perform image preprocessing, sea-land segmentation, and candidate region extraction. Constant false alarm rate (CFAR) [1], multi-resolution [2], [3], polarization information [4], and conversion [5] have been commonly used. However, these methods have high computational complexity, weak mobility, and are considerably laborious.

It is noteworthy that CNNs can effectively address the problems of traditional methods by automatically learning from SAR images in a robust manner.

Girshick introduced a region-based convolution neural network (RCNN) [6] to the field of target identification and recognition. Subsequently, the Fast RCNN [7] can realize end-to-end detection by adopting various optimizations, such

The associate editor coordinating the review of this manuscript and approving it for publication was Zhongyi Guo.

as shared convolution, ROI pooling, and multitask loss. The Faster RCNN [8] is based on the RPN network, and it uses the anchor mechanism to connect region generation with a CNN to realize real-time detection. The YOLO [9] used the idea of regression to complete classification and localization directly by using a one-stage network. The SSD [10] uses a fixed frame for region generation and multi-layer feature information to improve the detection speed and accuracy to a certain extent. By modifying the loss function, the RetinaNet [11] has addressed the class imbalance problem in one-stage methods.

For the Faster RCNN, Zhang et al. [12] employed binary normed gradient and cascaded CNN to improve the accuracy. Yang et al. [13] used RetinaNet and improved the loss function to reduce the false alarm rate. The SSD was enhanced by Wang et al. [14] to improve the detection speed. Furthermore, the authors enhanced the detection performance for small targets. Zhu et al. [15] used the YOLO to design an integrated multi-scale mechanism for the detection of small ships.

The attention mechanism modeled on the human visual system has been a research focus in the computer vision field. The research on spatial attention has been mostly based on recurrent neural networks (RNNs) [16]. The RAM [17] and subsequent studies, such as DRAW [18], GlimpseNet [19], STN [20], DCN [21], and DCNv2 [22] use subnetworks, were used to predict the target regions explicitly. The GeNet [23] uses the attention mechanism to predict a soft mask implicitly. It should be noted that the research on the channel attention mechanism has been mainly based on the SENet [24]. The improvement versions of the SENet include the GospNet [25], FcaNet [26] (the squeeze module is improved), ECANet [27] (the excitation module is improved), SRM [28], and GCT [29]. The channel and spatial attention mechanism includes the research of split-channel attention and spatial attention, and some of the proposed models are CBAM [30] the SCSE [31]. In contrast, the three-dimensional attention maps have been estimated directly by a residual attention mechanism. The follow-up work on split-channel attention and spatial attention has proposed triple attention [32] for cross-dimensional interaction, coordinated attention [33] for long-term dependence, and DANet [34] and RGA [35] for relationship perceived attention. There have been fewer studies on the attention mechanism in the field of ship target detection. Lin et al. [36] improved the detection performance of the Fast RCNN by introducing an attention mechanism. To enable multi-scale ship detection, the CBAM was introduced to the detection method by Cui et al. [37]. Further, Zhao et al. [38] implemented an attention mechanism into the FPN to achieve multi-scale detection. Fu et al. [39] designed a hierarchy-based attention network and a space-based attention network to improve detector performance.

To address the shortcomings of the existing research, this work introduces a coordination attention module to ship target detection. This solves the problem that the SENet considers only the internal channel information while ignoring the localization information. At the same time, the problems that the CBAM captures only local relationships from multiple channels of each location and is unable to obtain the long-range dependency relationship are solved. In addition, this work combines recent achievements in the research on decoupled heads and solves the contradictions between classification and regression. The decoupled head is used to enhance the performance of a deep learning network.

The main contributions of this work can be summarized as follows:

(1) The decoupled head is used to decouple the classification and regression tasks in the traditional coupled head;
(2) Compared with the conventional SE and CBAM attention mechanisms, the coordinate attention mechanism used in this work can obtain long-range dependencies and accurate position information in the spatial domain;
(3) A two-way trunk is used to improve the detection model's performance for small targets;
(4) Lightweight networks for mobile applications are presented.

In this paper, part II introduces YOLOv4's structure, part III introduces the optimization methods in this paper, including coordinate attention mechanism, decoupled head, loss function optimization and two-way trunk, part IV mainly introduces the experimental results of this paper, including comparative experiments and ablation experiments, part V mainly introduces the lightweight part, part VI summarizes our paper.

II. YOLO-v4 ARCHITECTURE

The YOLOv4 architecture is presented in Figure 1. The YOLOv4 uses the original YOLO structure but adopts well-known optimization strategies developed in recent years, including the CIO loss function, improved NMS [40], SPPNet [41], and PANet [42].

Compared with the YOLOv3, the backbone of the YOLOv4 is changed from the original Darknet-19 to CSPDarknet-53, which retains the original residual connection module but avoids network performance degradation. The CSPdarknet-53 contains five cross-stage partial (CSP) block backbones composed of a five-layer residual network named the resblock_body. The resblock_body incorporates a special convolution operation to reduce the input resolution. As is shown in Figure 2, a cross-stage partial network (CSPNet) [43] solves the problem of gradient information repetition in network optimization. Moreover, it also reduces the number of computations and ensures higher accuracy.

The SPP module is used in the CSPDarknet-53's last feature layer, which uses 5 × 5, 9 × 9, and 13 × 13 max-pooling layers to conduct multi-scale fusion. The obtained feature maps are used to expand the receptive field and introduce contextual features. The YOLOv4 model uses the output feature map of the SPP structure as input of the feature pyramid. By using the fusion method PANet, the final feature map is

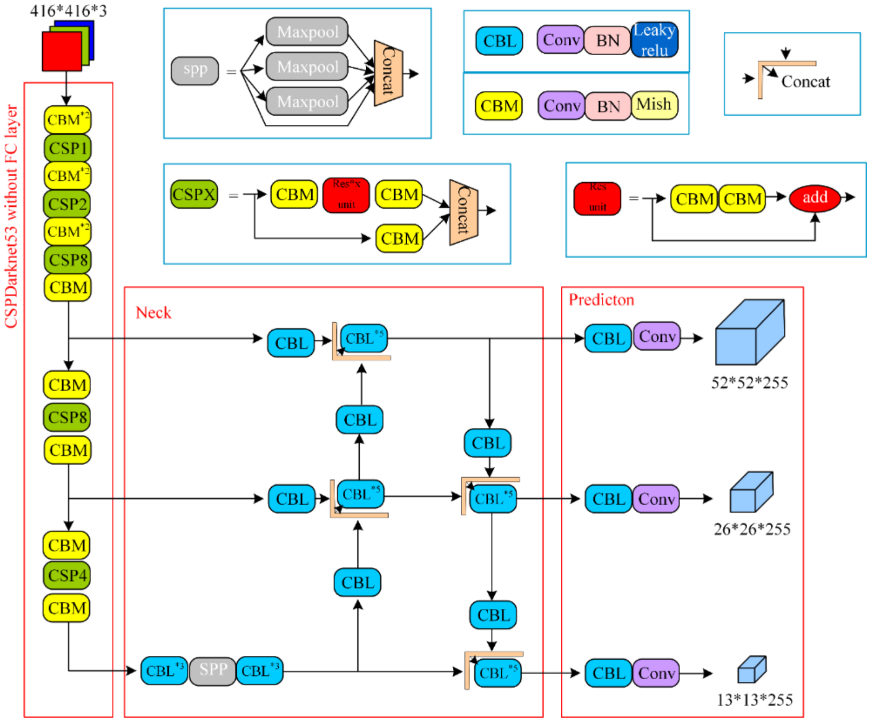


FIGURE 1. The YOLO-v4 architecture.

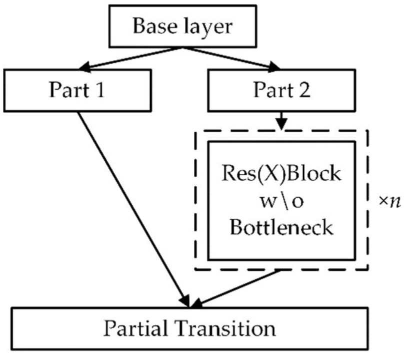


FIGURE 2. The schematic diagram of CSPNet.

provided to the YOLO detection head for final classification and localization.

The CIOU loss function, which considers not only the overlapping area of the predicted box and the ground truth in the GIOU loss function but also the distance between the center point of the predicted box and the ground truth in the DIOU loss function, is used in the YOLO-v4 model. In this loss function, both the predicted box's length-width ratio and the ground truth's length-width ratio are considered.

$$\mathrm{L_{CIOU}} = 1 - \mathrm{IOU} + \frac{\rho^2\left(b, \mathrm{b^{gt}}\right)}{c^2} + \alpha v \tag{1}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{\mathrm{w^{gt}}}{\mathrm{h^{gt}}} - \arctan\frac{w}{h})^2, \tag{2}$$

$$\alpha = \frac{v}{(1 - \mathrm{IOU}) + v}, \tag{3}$$

where c is the diagonal length of the smallest box that can simultaneously cover both the ground-truth and predicted boxes; $\rho^2(\mathrm{b}, \mathrm{b^{gt}})$ denotes the Euclidean distance between the ground truth and the predicted frame's center point; $\frac{\mathrm{w^{gt}}}{\mathrm{h^{gt}}}$ is the ground truth's aspect ratio; $\frac{w}{h}$ is the prediction frame's aspect ratio.

It should be noted that the YOLO detection head in the YOLOv4 model couples the classification and regression tasks, thus degrading network performance. At the same time, there is no explicit application attention mechanism in the YOLOv4 model, which affects the detection effect for dense and small target objects.

III. DECOUPLED HEAD AND COORDINATED ATTENTION DETECTION METHOD

To address the limitations of the YOLO network, an improved high-resolution ship detection method is proposed. The flowchart of the proposed method is presented in Figure 3, where it can be seen that the proposed architecture includes three main sections: a backbone, a neck, and a head. The head adopts the decoupled detection head, which is discussed in detail in Section III-*A*. In addition, coordinate attention is added to the residual blocks of stages 3–5 of the cross-stage partials to enhance the detection effect of a small target, which is explained in detail in Section III-*B*. In the neck, the input image has a size of 416 × 416. After passing through the backbone, processes 1–3 generate 13 × 13, 26 × 26, and

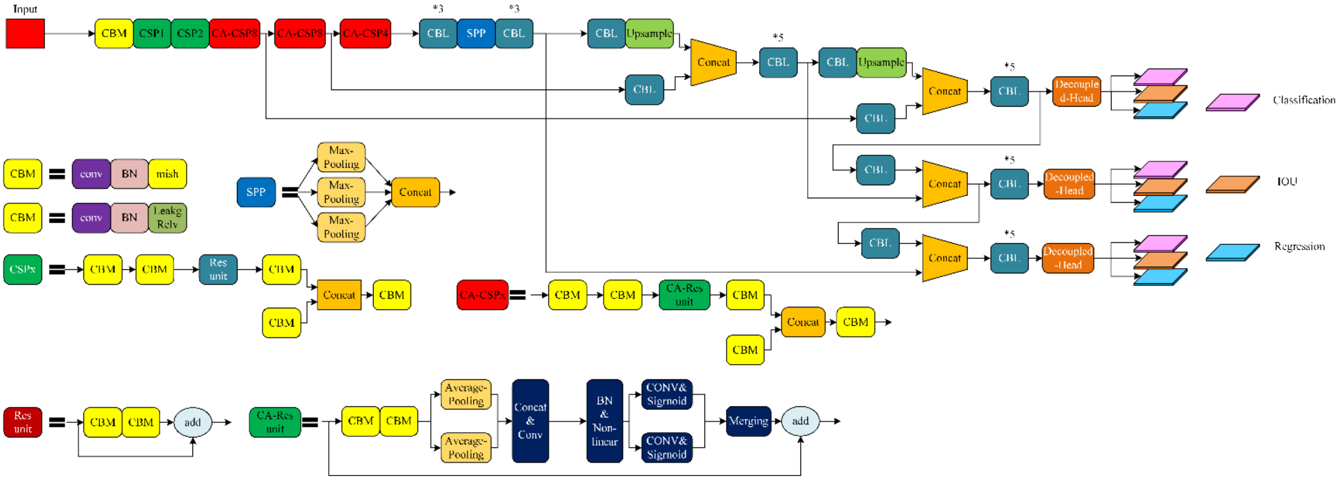

Input
CBM
CSP1
CSP2
CA-CSP8
CA-CSP8
CA-CSP4
*3
CBL
SPP
*3
CBL
CBL
Upsample
CBL
Concat
*5
CBL
CBL
Upsample
CBL
Concat
*5
CBL
Decoupled-Head
CBL
Concat
*5
CBL
Decoupled-Head
CBL
Concat
*5
CBL
Decoupled-Head
Classification
IOU
Regression
CBM = conv BN mish
CBM = conv BN Leakg Relv
SPP = Max-Pooling Max-Pooling Max-Pooling Concat
CSPx = CBM CBM Res unit CBM CBM Concat CBM
CA-CSPx = CBM CBM CA-Res unit CBM CBM Concat CBM
Res unit = CBM CBM add
CA-Res unit = CBM CBM Average-Pooling Average-Pooling Concat & Conv BN & Non-linear CONV& Sigmoid CONV& Sigmoid Merging add

FIGURE 3. The architecture of the proposed model.

53 × 53 feature maps, respectively. Process 2 upsamples input 13 × 13 feature maps to 26 × 26 feature maps and fuses them with the backbone's 26 × 26 feature maps. Process 3 upsamples the output feature maps of process 2 to the size of 52 × 52 and fuses them with the backbone's 52 × 52 feature maps, and then passes the resulting data to the detection head. The bottom-up integration process of the PAN is implemented based on processes 4 and 5. Process 4 downsamples the output feature maps of process 3 to the size of 26 × 26, fuses the obtained maps with the output feature map of process 2 and then passes the resulting data to the detection head. Process 5 downsamples the output feature maps of process 4 to 13 × 13 feature maps and then fuses them with the output feature maps of process 1 and passes them to the detection head.

A. DECOUPLED HEAD

The conflict between classification and regression tasks in object detection has been a widely analyzed problem [44]. The YOLOX [45] shows that the coupled head degrades network performance to a certain extent. It is worth noting that the decoupled head can enhance a network's convergence speed. Therefore, the decoupled head is crucial in end-to-end models [46], [47]. In addition, it should be noted that the coupled detection head has been used in YOLOv3–YOLOv5 models, and it consists of a 1 × 1 convolution layer. For instance, when the coupled detection head of the YOLOv4 model is used for performing detections on the COCO dataset, three boxes are preset. Each box needs to predict the confidence of a target, four regression bounding box parameters, and 80 category probabilities. This leads to a reduction in network performance and the inability to determine the location of a target accurately. Therefore, this study decouples the detector head and derives a branch responsible for finding the target object and regression bounding box, as well as a branch to handle the target category. Finally, the two branches are integrated into the prediction to avoid performance degradation in the traditional detector head. The architecture of the proposed decoupled detection head with anchors is presented in Figure 4. The decoupled head reduces the characteristic channels using a 1 × 1 convolution layer and adds two parallel branches. For the classification and regression tasks, each of the branches has two 3 × 3 convolution layers.

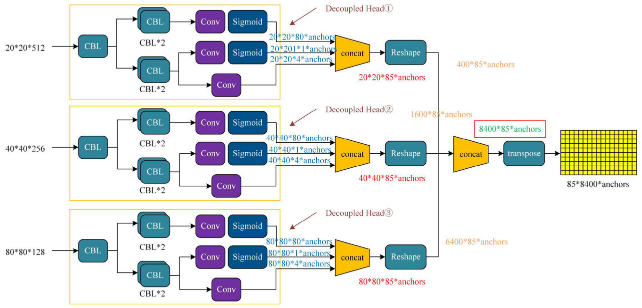


FIGURE 4. The structure of the proposed decoupled head with anchors.

B. COORDINATE ATTENTION

The SENet structure can be briefly described as follow. First, the global average pooling operation defined in (4) is employed to compress global spatial information in statistical data of the channel dimension, which is based on a squeeze module; namely, the input of a size of $C \times H \times W$ is converted to the size of $C \times 1 \times 1$. The numerical distribution of C characteristic graphs is stored in the output. Second, the channel correlation is obtained by the excitation module. Because the size of the feature map obtained by the squeeze module is $C \times 1 \times 1$, the result obtained by it is fed to the fully-connected layer. Therefore, the result obtained after passing it through the fully-connected layer is $C / r \times 1 \times 1$, where r represents the scaling factor. Then, after passing through a nonlinear layer and another fully connected layer, the output dimension becomes $C \times 1 \times 1$. Finally, according to the output weights of the excitation module, the reweighting process allocates the weight of the characteristic channel, which is performed to finish the recalibration of the original feature in the channel dimension.

The CBAM structure is as follows. The CBAM combines channel and spatial information. Its operation begins by compressing the feature maps in the spatial dimension by using both average and maximum pooling to obtain a one-dimensional vector. Next, a multi-layer perceptron is used to process this vector. Then, both average and maximum pooling are applied to the feature map in the channel dimension. Afterward, the two results are concatenated according to the channel dimensions. Further, the dimension is reduced to a single channel after the convolution process. The final spatial attention feature map is obtained by multiplying this feature map with the input feature map. It should be noted that the SENet considers only channel information but ignores the position information. Conversely, the CBAM considers location information by introducing the weighting coefficients. The weighting process captures only the local relationships and is unable to obtain long-range dependencies. However, coordinated attention can effectively address the aforementioned problems.

Moreover, the coordinate attention module decomposes (4) into horizontal and vertical parts, as shown in (5) and (6), respectively. Particularly, given an input x, each channel is encoded along the horizontal and vertical directions based on two spatial ranges, $(h, 1)$ and $(1, w)$, of the pool core. To build a pair of direction sensing feature maps, the features are aggregated in the two spatial directions. This enables the attention block to capture the long-term dependence and attain accurate location information. In addition, this significantly helps the network to locate an object of interest precisely.

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum\nolimits_{i=1}^{H} \sum\nolimits_{j=1}^{W} \mathrm{u_c}(i, j), \quad (4)$$

$$\mathrm{z_c^h}(h) = \frac{1}{\mathrm{W}} \sum\nolimits_{0 \leq i \leq W} \mathrm{x_c}(h, i), \quad (5)$$

$$\mathrm{z_c^w}(w) = \frac{1}{H} \sum\nolimits_{0 \leq j \leq H} \mathrm{x_c}(j, w). \quad (6)$$

As shown in Figure 5, it concatenates the results obtained by (5) and (6). These results are finally sent to a 1 × 1 convolution layer, which is given by:

$$f = \delta\left(\mathrm{F_1}\left(\left[\mathrm{z^h}, \mathrm{z^w}\right]\right)\right), \quad (7)$$

where δ denotes the sigmoid function, F1 is the convolution function, $[\cdot]$ represents the concatenation operation, and f is a $C/r \times (H + W)$, and it is divided into two parts, $f^h$and f^w. The output results are expressed by using the following mathematical expressions:

$$\mathrm{g^h} = \delta\left(\mathrm{F_h}\left(\mathrm{f^h}\right)\right), \quad (8)$$

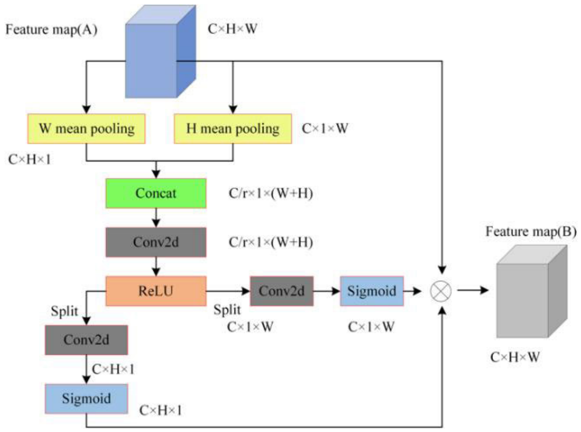


FIGURE 5. The architecture of the coordinate attention.

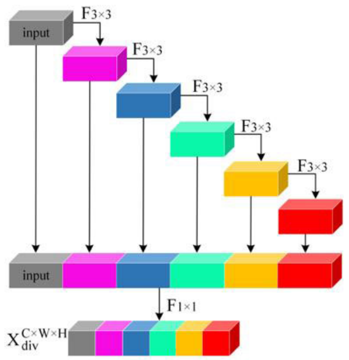


FIGURE 6. The OSA architecture.

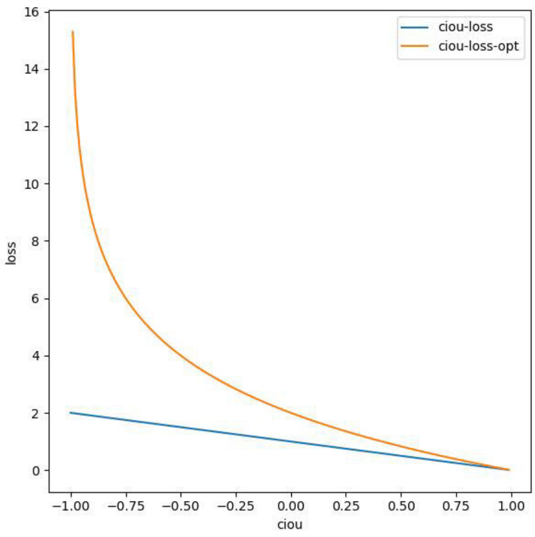


FIGURE 7. The CEIOU and ICEIOU curves.

$$\mathrm{g^w} = \delta\left(\mathrm{F_w}\left(\mathrm{f^w}\right)\right). \quad (9)$$

Finally, the output of the coordinate attention module is given by:

$$\mathrm{y_c}(i, j) = \mathrm{x_c}(i, j) \times \mathrm{g_c^h}(i) \times \mathrm{g_c^w}(j). \quad (10)$$

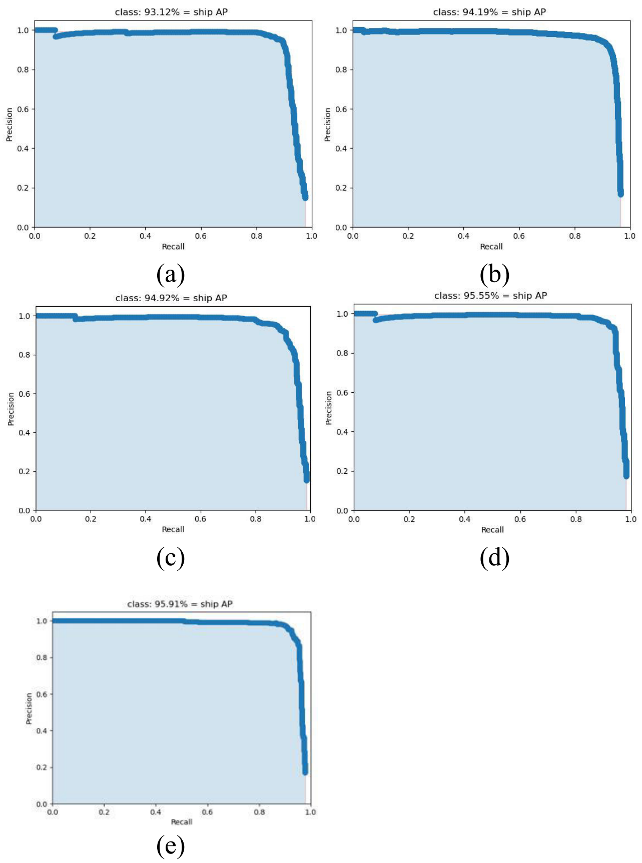


FIGURE 8. The P-R curve of the HRSID ablation experiment. (a) Baseline; (b) baseline with decoupled head; (c) baseline with coordinate attention; (d) baseline with decoupled head and coordinate attention; (e) baseline with decoupled head, coordinate attention, and double trunk.

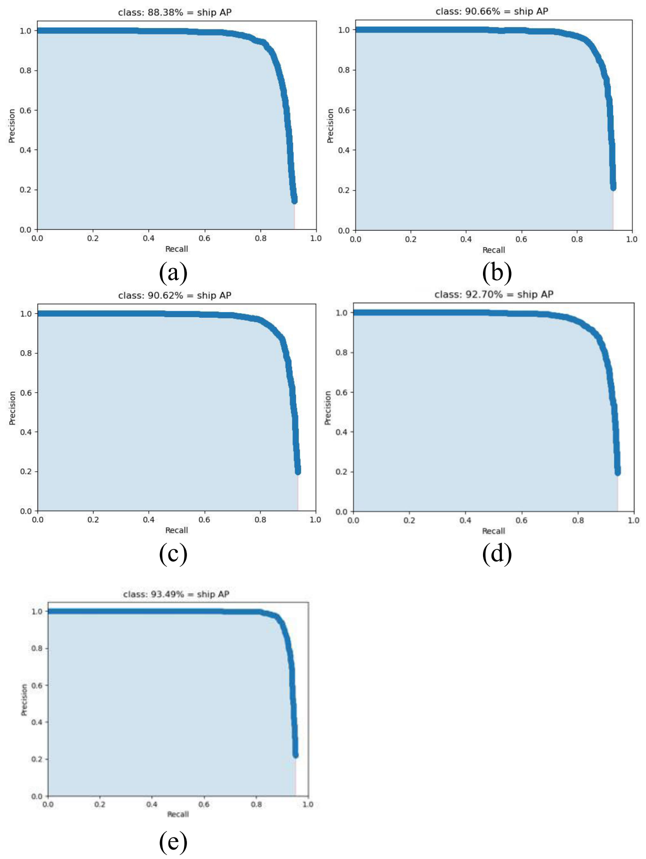


FIGURE 9. The P-R curve of the HRSID ablation experiment. (a) Baseline; (b) baseline with decoupled head; (c) baseline with coordinate attention; (d) baseline with decoupled head and coordinate attention; (e) baseline with decoupled head, coordinate attention, and double trunk.

TABLE 1. The map of the ablation experiment obtained on the SSDD dataset.

	Double Trunk	Decoupled Head	Coordinate Attention	mAP (%)
(a)	×	×	×	93.12
(b)	×	√	×	94.19 (+1.07)
(c)	×	×	√	94.92 (+1.80)
(d)	×	√	√	95.55 (+2.43)
(e)	√	√	√	95.91 (+2.79)

TABLE 2. The map of the ablation experiment obtained on the HRSID dataset.

	Double Trunk	Decoupled Head	Coordinate Attention	mAP (%)
(a)	×	×	×	88.38
(b)	×	√	×	90.66(+2.28)
(c)	×	×	√	90.62(+2.24)
(d)	×	√	√	92.70(+4.32)
(e)	√	√	√	93.49(+5.11)

In addition, (c) can be easily integrated into mobile networks, such as MobileNetv2 [48] and EfficientNet [49]. This is also conducive to the lightweight work of subsequent models.

C. DOUBLE TRUNK

To enhance the extraction ability of the network model for small targets further, a two-way backbone is introduced into the VoVNet [50] on the basis of the CSPDarknet53. The VoVnet can effectively extract various feature information by using the one-time aggregation (OSA) module, which connects the subsequent layers, as shown in Figure 6. Because the OSA module can capture multi-scale receptive fields, the diversified feature maps enhance the multi-scale target detection ability of the target detection model, especially for small target detection.

In this paper, the feature information extracted from the input feature map by the backbone of the VoVNet is fused with the feature map extracted by another Yolo backbone, the CSP8. More feature information is extracted through feature fusion, which is conducive to enhancing the model's detection ability for small ship targets.

TABLE 3. Comparison results on the SSDD dataset.

Method	Parameter (M)	Precision (%)	Recall (%)	F1-score (%)	$mAP_{0.5:0.95}$ (%)	$mAP_{0.5}$ (%)	$mAP_{0.75}$ (%)	FPS
Faster RCNN	137	46.27	90.08	61.83	42.65	87.52	34.19	8.66
RetinaNet	37	93.31	83.67	89.12	45.79	93.90	45.17	21.96
SSD	**26**	93.05	60.94	73.34	42.52	89.82	38.26	**47.20**
ImYOLOv4	65	93.54	90.95	92.91	50.64	94.16	58.19	42
Ours	115	**95.33**	**93.09**	**93.31**	**58.75**	**95.55**	**68.17**	26.48

TABLE 4. Comparison results on the HRSID dataset.

Method	Parameter (M)	Precision (%)	Recall (%)	F1-score (%)	$mAP_{0.5:0.95}$ (%)	$mAP_{0.5}$ (%)	$mAP_{0.75}$ (%)	FPS
Faster RCNN	137	81.63	81.45	81.23	37.24	81.11	34.71	9.05
RetinaNet	37	86.28	81.43	84.85	44.63	84.01	48.72	22.38
SSD	**26**	85.21	60.08	70.41	39.36	82.97	32.16	**47.27**
ImYOLOv4	65	92.02	88.95	90.65	45.62	92.34	60.42	42
Ours	115	**93.04**	**90.07**	**92.42**	**48.82**	**93.49**	**65.21**	26.28

D. LOSS FUNCTION

In the loss function part, the CEIOU loss function [51] reflects the distance between the predicted frame and the ground truth. When the CEIOU value is large, the closer the predicted frame is to the ground truth, the smaller the loss function value is. However, in the training process of a network model, the gradient of the CEIOU loss function in the training process will not change, thus affecting the training effect. In order to solve this problem, the ICEIOU loss function, which is presented in Figure 7, is introduced to improve the model training effect.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the simulation setup, datasets, evaluation measures, and implementation of the proposed method are presented. In addition, ablation and comparative experiments are introduced.

A. DATASETS

The SAR ship detection dataset (SSDD) [12] containing 1,160 images and 2,456 ship targets was used in the experimental verification. This dataset was collected by using Sentinel-1, TerraSAR-X, and RadarSat-2 sensors, including target ships with HH, HV, VV, and VH polarization modes. The resolution for the dataset was 1 m–15 m. Please note that the images of target ships were acquired on seas, as well as in nearshore areas. The high-resolution SAR images dataset (HRSID) [52] contained 5,064 SAR images and 16,951 ship targets. It was collected using TanDEM-X, TerraSAR-X, and Sentinel-1B sensors installed on target ships with HH, HV, and VV polarization modes and different backgrounds. The dataset included data with resolutions of 0.5 m, 1 m, and 3 m.

B. SIMULATION SETUP

For implementation, PyTorch 1.8.0, CUDA 11.1, CUDNN 8805, Intel(R) Xeon(R) Gold 6130, and Tesla P100 were used. Model training included 200 epochs; a learning rate was 0.0003, and an AdaBelief optimizer was adopted. It should be noted that optimizer selection has a significant effect on the convergence of a trained deep learning model [53], [54]. The AdaBelief optimizer was selected because it has both the fast convergence characteristics of the Adam optimizer and the good generalization capability of the SGD [55]. Further, the learning rate used a cosine annealing strategy. The detection threshold for IOU in all experiments was 0.5.

C. EVALUATION METRICS

In this study, precision, recall, F1_score, FPS parameters, and GFLOPs were used to evaluate the proposed method's detection performance. The precision and recall were respectively calculated by:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{11}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{12}$$

The F1_score is a mathematical expression of the harmonic average of accuracy and recall, and it is defined by:

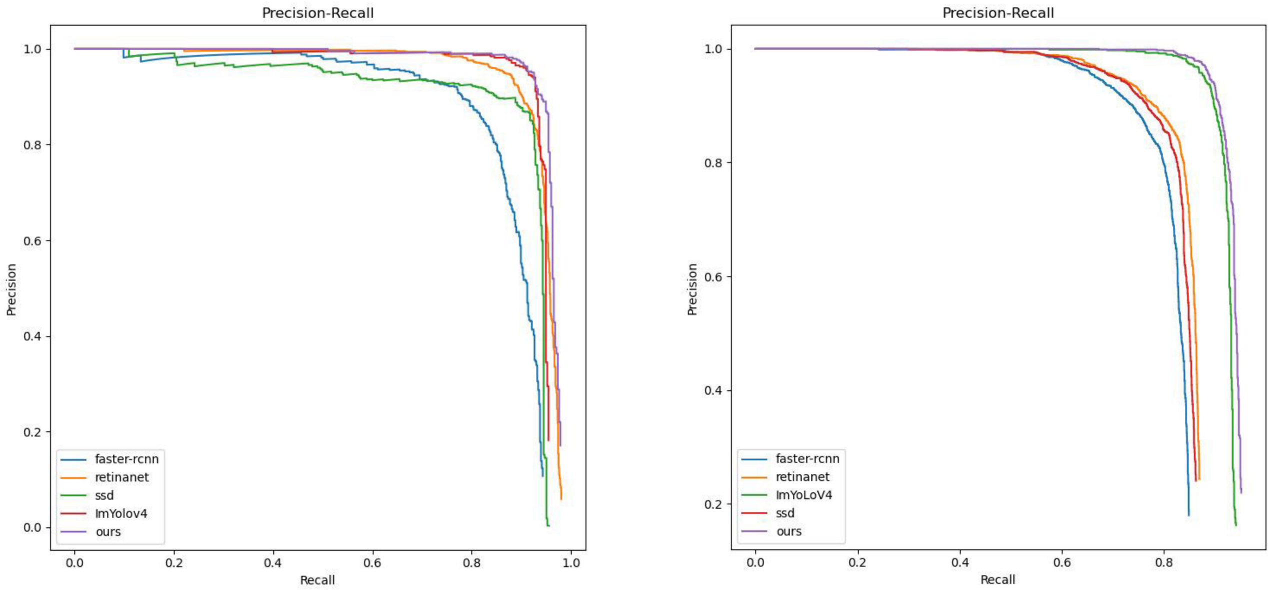
$$\text{F1\_score} = \left(\frac{1}{n}\sum \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\right)^2. \tag{13}$$

Further, the AP is mathematically expressed as follows:

$$\text{AP} = \int_0^1 P(R)\text{d}R. \tag{14}$$

Generally, mAP (0.5:0.95) represents IOU from 0.5 to 0.95. In this study, mAP was calculated at intervals of 0.05 and then averaged; mAP0.5 and mAP0.75 denoted the map values for IOU of 0.5 and 0.75, respectively.

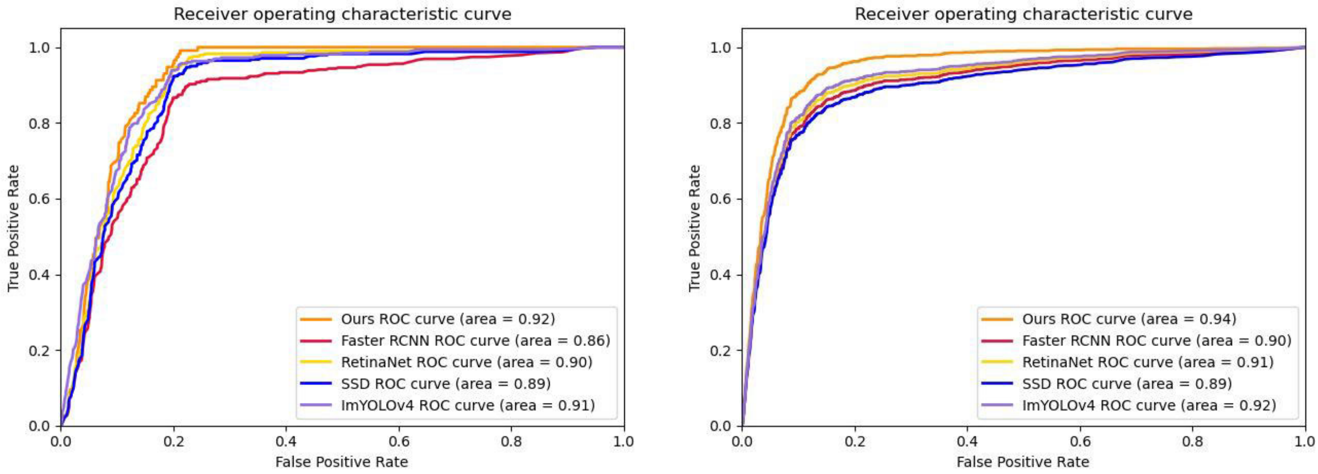
The receiver operation characteristic (ROC) curve describes the relationship between the true positive rate (TPR) and the false positive ratio (FPR). The AUC is defined



(a)

(b)

FIGURE 10. **The P-R curves: (a) SSDD dataset; (b) HRSID dataset.**



(a)

(b)

FIGURE 11. **The ROC Curves. (a) SSDD. (b) HRSID.**

as the area between the ROC curve and the coordinate axis. The TPR and FPR are respectively defined as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{15}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}. \tag{16}$$

where *FPS* represents the detection speed, and it is given by:

$$\mathrm{FPS} = \frac{N}{T}. \tag{17}$$

where N denotes the number of samples in the test set, and T is the amount of time required for testing the model on the test set.

The GFLOPs are used to measure the computation amount. Namely, network complexity is proportional to the number of performed calculations. The GFLOPs represent the number of parameters in the network. In a neural network, parameters generally refer to the weights and biases that are learned during the training process.

D. ABLATION EXPERIMENT

To confirm the efficiency of the decoupled head and coordinated attention module in detecting target ships in SAR images, ablation experiments were performed on the SSDD and HRSID datasets. Figure 8 presents the P-R curves of

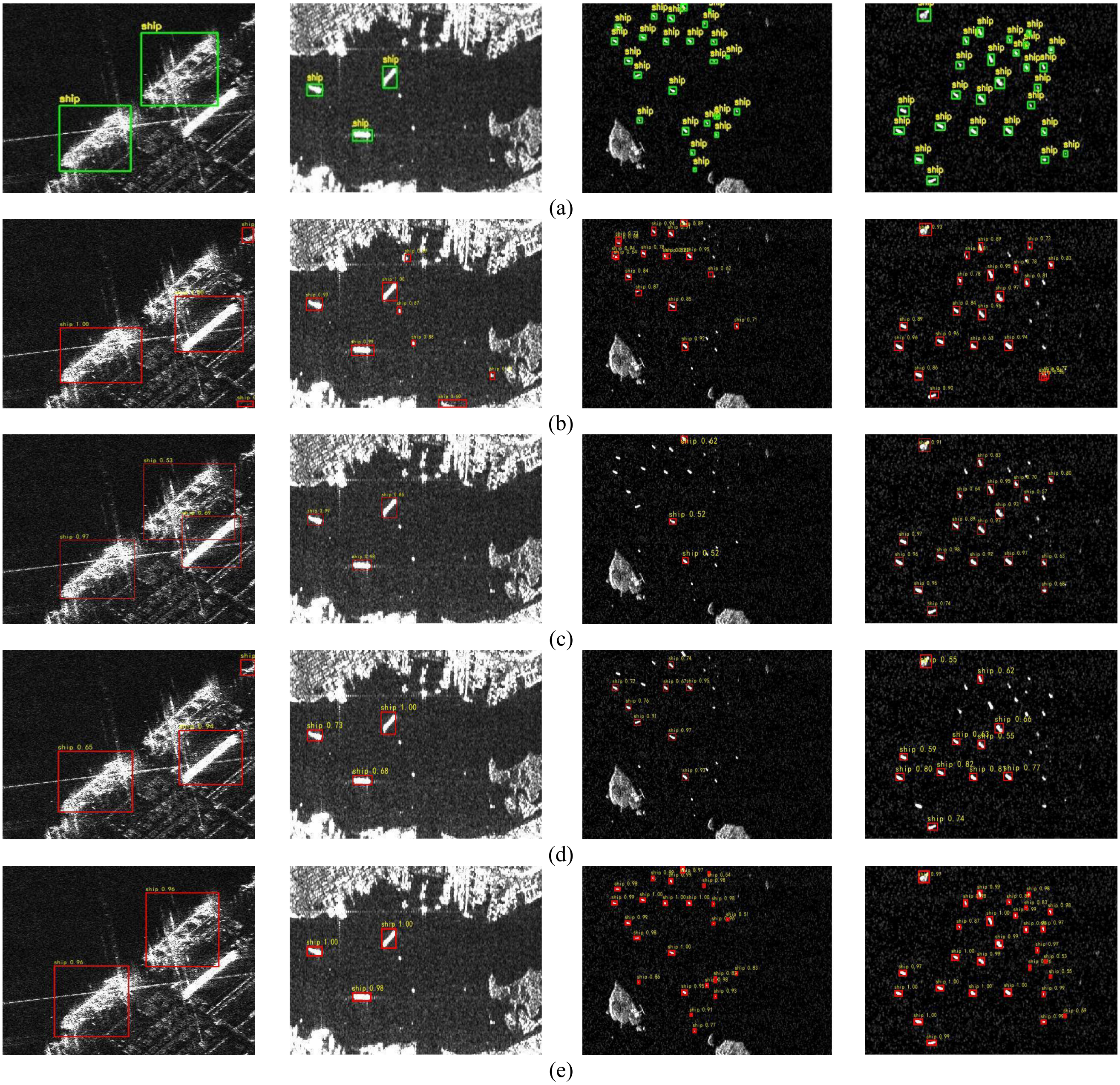


FIGURE 12. The detection result of different algorithms on the SSDD dataset. (a) Ground truth; (b) Faster RCNN; (c) RetinaNet; (d) SSD; (e) the proposed method.

TABLE 5. Performance comparison of different attention mechanisms on the SSDD dataset.

Method	$mAP_{0.5}$ (%)	Precision (%)	Recall (%)	F1-score (%)
YOLOv4_Decouple	94.19	91.03	92.21	91.62
+SE	94.76 (+0.57)	91.05	92.05	91.55
+CBAM	94.85 (+0.66)	**97.11**	89.36	93.07
+CA	**95.55 (+1.36)**	95.33	**93.09**	**93.31**

the ablation experiment on the SSDD dataset. The mAP results of the ablation experiment are given in Table 1. The results showed that the decoupled detection head used to replace the coupled detection head for decoupling could

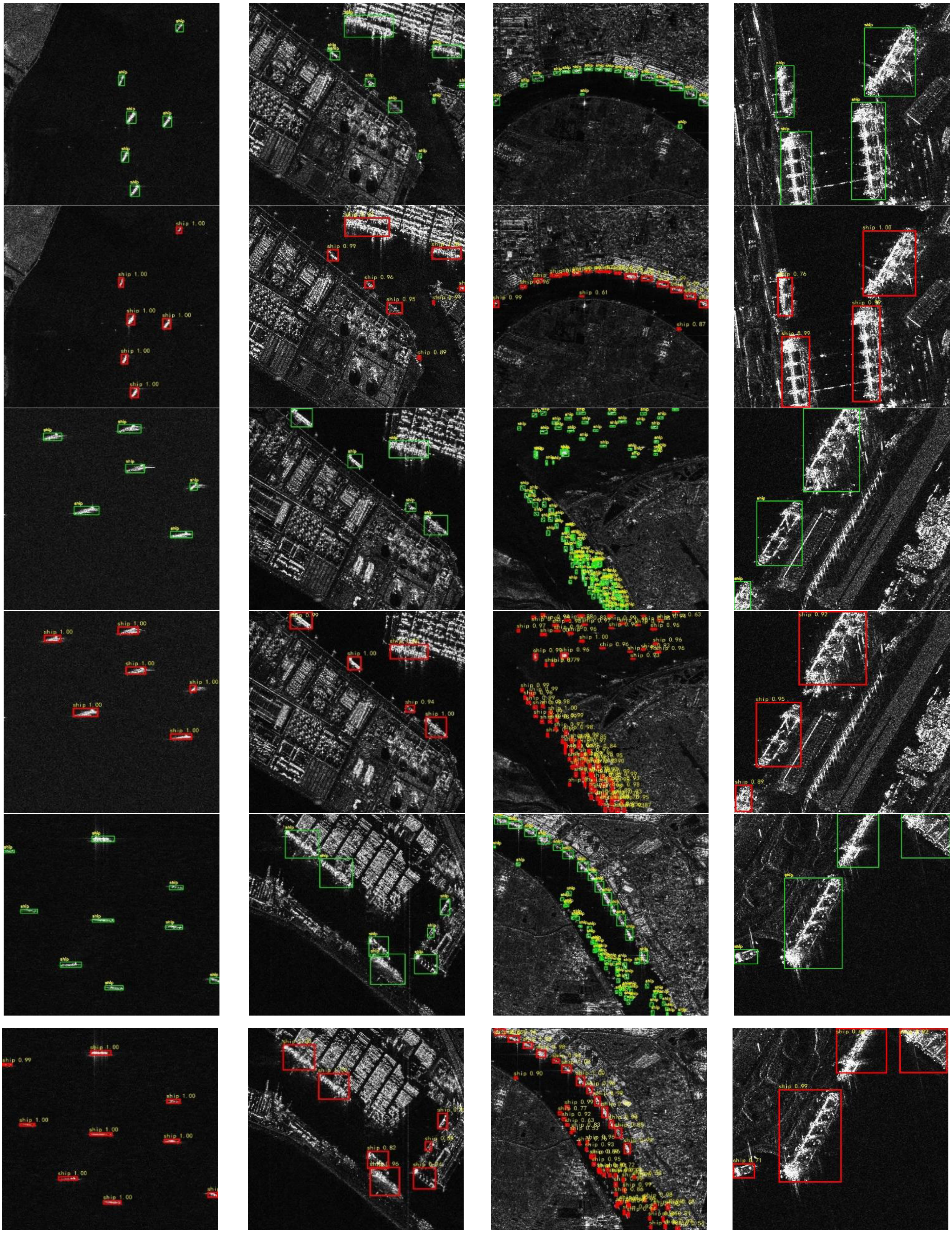


FIGURE 13. The detection results on the HRSID dataset.

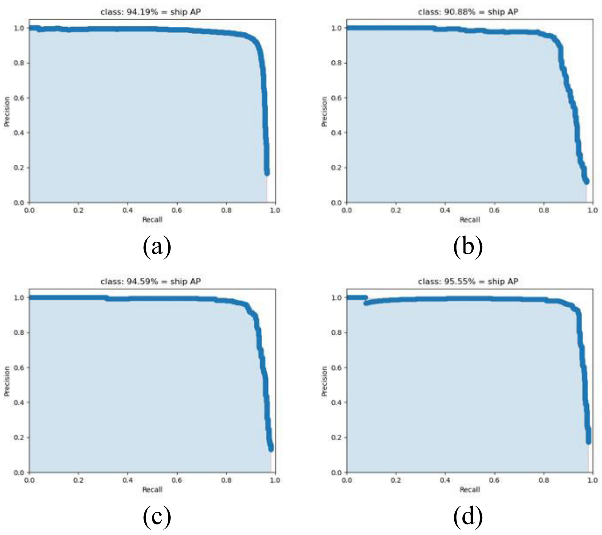
handle classification and regression problems. This improved the mAP by 1.07% and significantly enhanced network performance. In addition, inspired by the ConvNeXt [56], the coordinate attention module was added to the third stage for achieving robust feature learning. The coordinate attention module was also incorporated in the fourth and fifth phases with deeper semantics. This improved the mAP by 1.80% compared to the baseline network. The network obtained by combining decoupled detection head and coordinate attention showed a mAP improvement of 2.43% compared with the baseline network. Figure 9 shows the P-R curve of the ablation experiment on the HRSID dataset, and the corresponding mAP results are presented in Table 2. Compared with the baseline, the decoupled detection head increased the mAP by 2.28%, the coordinate attention enhanced the mAP by 2.24%, and the combination of decoupled detection head and coordinate attention improved the mAP by 4.32%. It is noteworthy that the HRSID dataset included a large number of small targets and rich data, thus enabling the decoupled detection head and coordinate attention to affect the network's performance significantly.

E. COMPARATIVE EXPERIMENT

Next, the proposed algorithm was compared with the Faster RCNN, SSD, RetinaNet, and ImYOLOv4. The comparison results are presented in Tables 3 and 4. The P-R curves of different algorithms obtained on the SSDD and HRSID datasets are presented in Figure 10. The ROC curves of different algorithms obtained on the SSDD and HRSID datasets are presented in Figure 11. The experimental results showed that although the number of parameters in the proposed algorithm was slightly larger than in the SSD and RetinaNet, it outperformed the other methods by more than 4% on the F1_score. In terms of precision and recall, the proposed method surpassed the previous methods by more than 2% and 3%, respectively. Further, in terms of mAP0.5, the performance of the proposed method was more than 1.4% higher than those of the other methods; also, the proposed method achieved a significant improvement in mAP0.5:0.95. As shown in Table 4, the proposed method outperformed other methods on the F1_score by more than 1.2% and by more than 1% and 2% on precision and recall, respectively. The proposed method improved the mAP0.5 by more than 1.1%, as well as mAP0.75. Moreover, the proposed method had the lowest computational complexity of 32.27 GFLOPs, while those of the Faster R-CNN, RetinaNet, and SSD were 109.7, 87.7, and 107.5 GFLOPs, respectively.

To emphasize the benefits of the proposed strategy even further, various types of targets, such as small targets, nearshore targets, and dense targets, were selected to compare the detection results of different methods on the SSDD dataset, as shown in Figure 12, where the ground truth is denoted by the green box, and the predictions of the algorithms are represented by the red boxes.

Figures 12(a) show the detection results of the ground truth, Fast RCNN, RetinaNet, SSD, and the proposed algorithm, respectively. As presented in Figures 12(b)-12(d), for nearshore and small targets, the degrees of false alarm and missed detection were obvious. The proposed method reduced the probability of missed detections and false alarms and had a good performance. For small and dense targets, Faster RCNN had a smaller number of missed detections than the other algorithms, as shown in Figures 12(c) and 12(d); however, the missed detection problem was still prominent. In contrast, the proposed algorithm could effectively recognize dense and small objects, having a low percentage of missed detection. The detection results obtained on the HRSID are shown in Figure 13, where the green box represents the ground truth, and the red box represents the prediction box. The results suggested that the proposed model could detect targets that were close to shore and dense and small.



(a) (b) (c) (d)

FIGURE 14. **The P-R curves of the attention mechanism experiment. (a) Decoupled head model; (b) Experiment 1; (c) Experiment 2; (d) Experiment 3.**

TABLE 6. **The description of two large SAR images.**

No.	Place	Time	Polarization	Mode	Resolution	Size
Image1	Singapore Strait	6 June 2020	VV	IW	5 m × 20 m	25,650 × 16,768
Image2	Malacca Strait	12 April 2020	VV	IW	5 m × 20 m	25,427 × 16,769

F. ATTENTION MECHANISM EXPERIMENT

Further, an experimental investigation of various effects of coordinate attention in the YOLOv4 hierarchical structures was conducted. Namely, the effects of coordinated attention, SE, and CBAM in the ship target detection task from SAR images were investigated using the YOLOv4 as a baseline network.

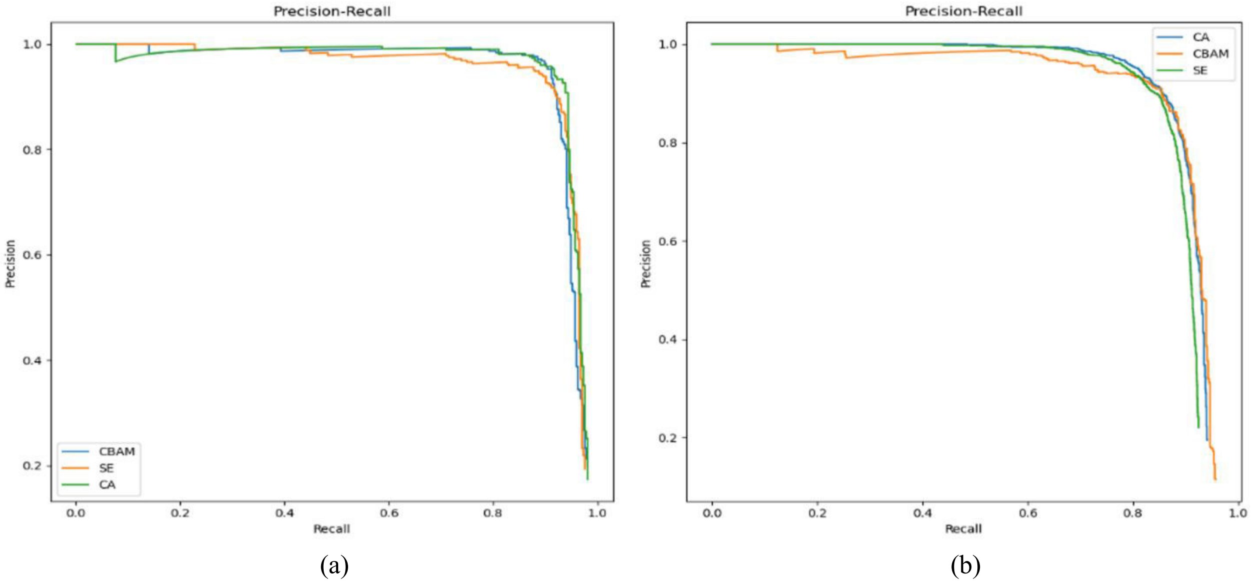


FIGURE 15. The P-R curve of the different attention mechanisms. (a) SSDD. (b) HRSID.

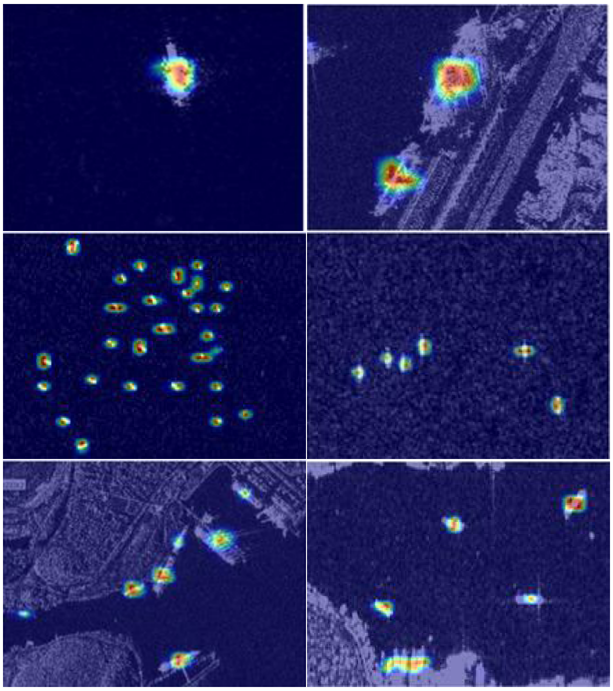


FIGURE 16. The intermediate feature visualization results.

First, a comparative experiment was performed on the SSDD dataset using different addition positions. In Experiment 1, coordinate attention was added before all residual layers, and in Experiment 2, coordinate attention was added to all residual layers. Inspired by the ConvNeXt, the features learned by the model in the third stage were the most robust. At the same time, the MobileNet showed that the features learned by the model in the fourth and fifth stages had high semantics. Thus, in Experiment 3, the residual layers of P3, P4, and P5 received coordinate attention. Figure 14 shows the P-R curve obtained by adding the attention mechanism to the decoupled optimization model in different positions.

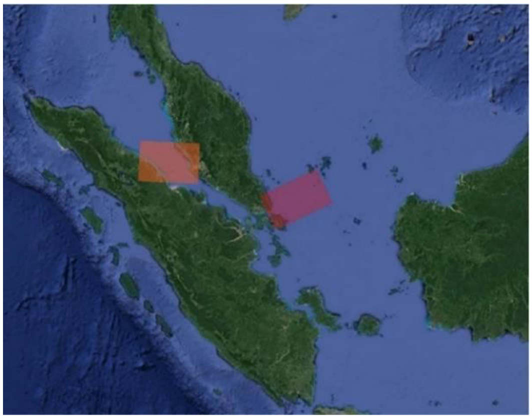


FIGURE 17. The annotated Google map of the two aforementioned locations.

The results of Experiment 1 showed that when the CA module was added to all residual layers, the mAP was reduced by 3.31% compared with the decoupled optimization model. The results of Experiment 2 demonstrated that when the CA module was added to all residual layers, the mAP increased slightly, namely by only 0.4%, compared with the decoupled optimization model. Thus, the effect of the attention mechanism was not obvious. The results of Experiment 3 showed that the addition of the attention mechanism to the residual layer of P3, P4, and P5 increased the mAP by nearly 1.4% compared with the decoupled optimization model. The three

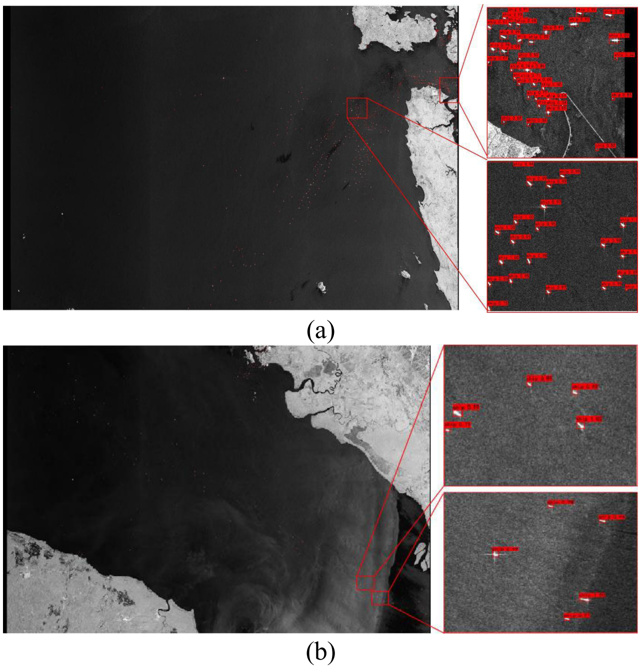


FIGURE 18. The detection results of the adaptability experiment. (a) Image 1; (b) Image 2.

experiments showed that adding coordinate attention to the residual layers of P3, P4, and P5 could be effective.

Figure 15 presents the P-R curves of different attention methods obtained on the SSDD and HRSID datasets. As shown in Figure 15, the coordinate attention mechanism outperformed the SE and CBAM in the ship detection task from SAR images. Table 5 compares the performances of different attention mechanisms on the SSDD dataset. As shown in Table 5, all three attention models could improve the model's map, but the CBAM and SE reduced the decoupled model's recall rate. The CA improved the model's map the most, while also improving the model's recall rate. Therefore, among the three attention mechanisms, the CA provided the greatest enhancement to the model's detection.

In addition, as shown in Figure 16, the intermediate feature visualization results demonstrated the benefits of the coordinate attention module proposed in this paper.

After introducing the coordinate attention mechanism, the model could effectively deal with the multi-scale problem in the task of ship target detection in SAR images.

G. ADAPTABILITY EXPERIMENT

To ensure that the selected model could be migrated easily, two large SAR images were selected, and their real geographical locations were marked, as presented in Figure 17. In Figure 17, the Strait of Malacca is denoted by the orange color, and the Strait of Singapore is represented in red. These locations represent famous shipping routes in the world. The descriptions of the two large scene images are given in Table 6. As shown in Table 6, the VV polarized target ships with high backscattering value and interference broadband (IW) mode of sentry 1 were selected. Due to the limited GPU memory, it was impossible to use large-scale images in model training directly. Therefore, the training and test were performed by segmenting the subgraphs in the document [57]. The adaptability of the proposed algorithm was also tested, and the obtained results are shown in Figure 18.

The detection results in Figure 18 show that the suggested model could successfully detect the majority of the target ships.

V. DISCUSSION

Currently, the three popular factions of target detection architecture on mobile terminals include the ShuffleNet [58, 59], FBnet [60], and MobileNet [61]. The mobile target detection architecture adopted in this work is MobileNetv3, and it is presented in Figure 19. The core idea in the MobileNet series refers to deep separable convolutions. The deep separable convolution divides an ordinary convolution into deep and point-by-point convolutions. During the process of deep convolution, the convolution kernel is divided based on the channel dimension and convoluted with the input feature map. However, it is noteworthy that the reduction in dimensions of the characteristic map leads to the loss of useful information. To address this issue, MobileNetV3 introduces point-by-point convolution after deep convolution to ensure the number of channels of the output feature map. Based on these operations, the depth separable convolution reduces the number of computations and parameters to about one-ninth to one-eighth of the standard convolution at the cost of

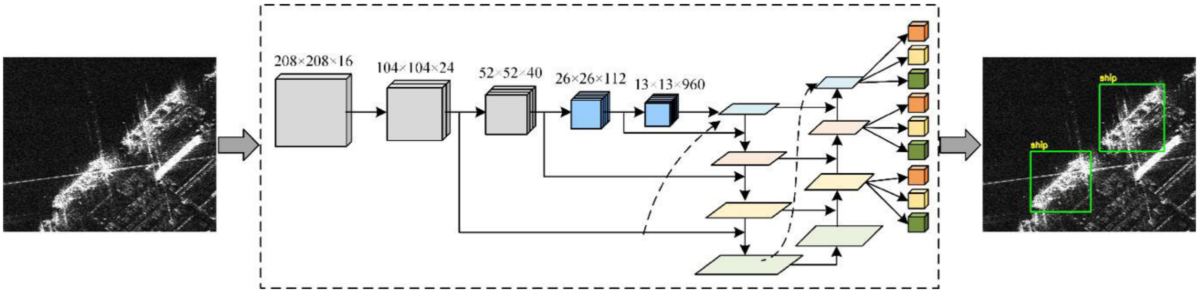


FIGURE 19. The architecture of the MobileNetv3 is used in this work.

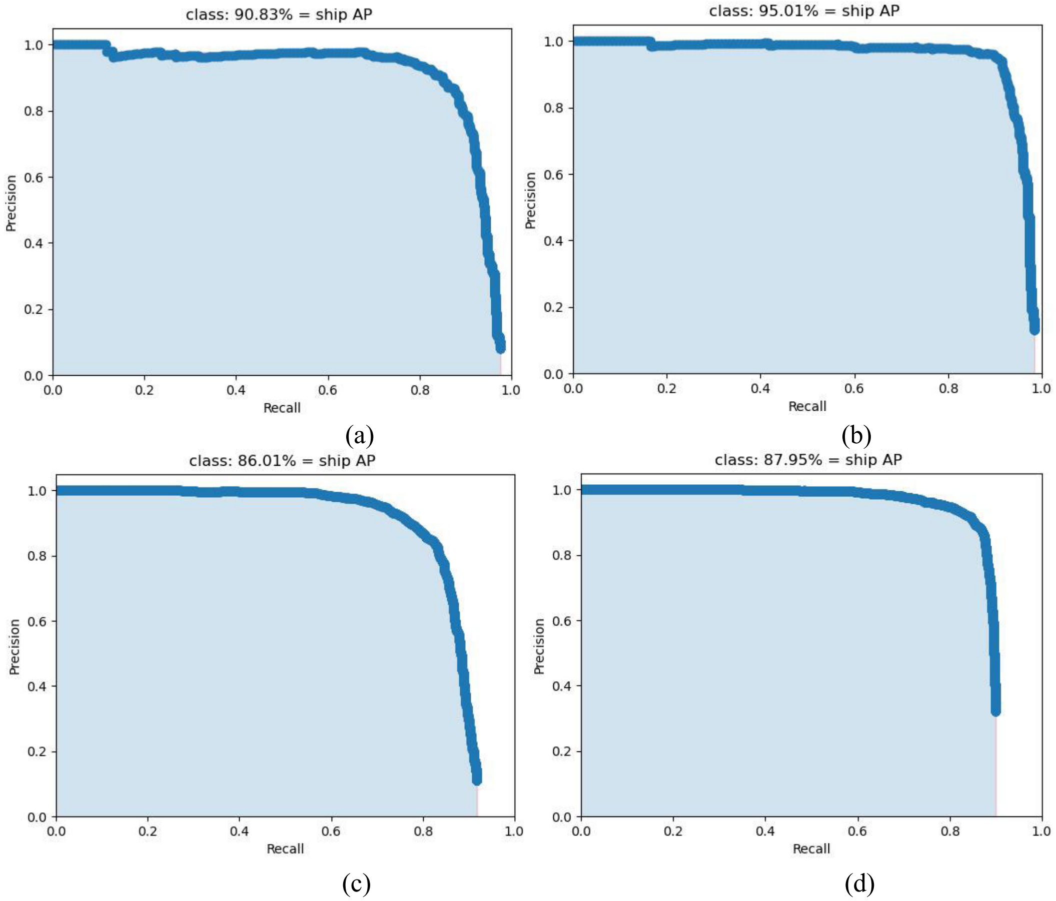


FIGURE 20. The P-R curve of the lightweight model. (a) lightweight baseline on the SSDD dataset; (b) lightweight baseline with decoupled head and coordinate attention on the SSDD dataset; (c) lightweight baseline on the HRSID dataset; (d) lightweight baseline with decoupled head and coordinate attention on the HRSID dataset.

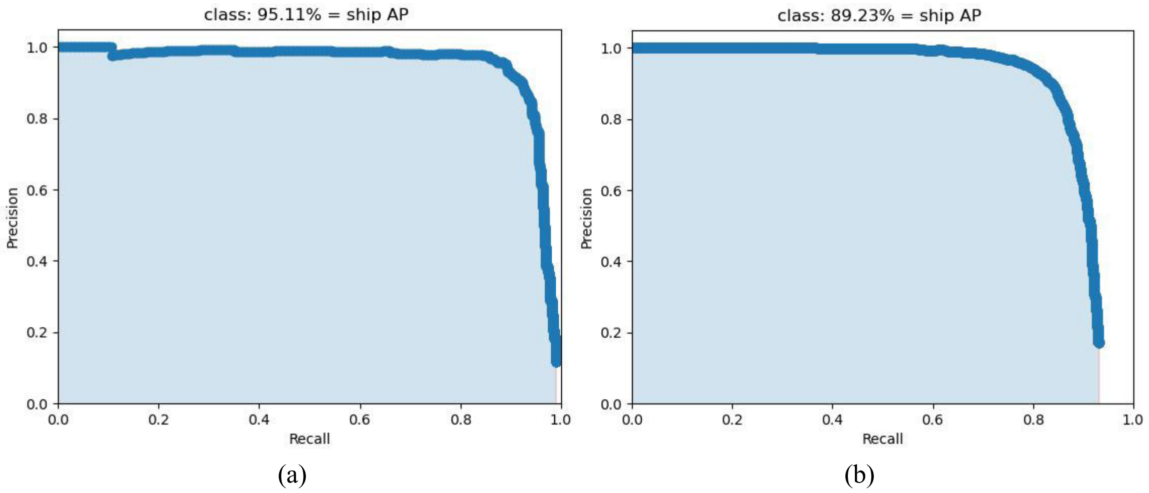


FIGURE 21. The P-R curve of the lightweight model after the introduction of the focus; (a) SSDD dataset; (b) HRSID dataset.

a 1% reduction in accuracy. Second, the inverse residual structure with a linear bottleneck is implemented to reduce information loss during the training process caused by a low-dimensional ReLU. The pointwise (PW) convolution is used to upgrade the dimensions before performing deep convolution, and then convolution is performed in a high

TABLE 7. The lightweight comparison on the SSDD dataset.

Method	Parameter (M)	Precision (%)	Recall (%)	F1-score (%)	$mAP_{0.5:0.95}$ (%)	$mAP_{0.5}$ (%)	$mAP_{0.75}$ (%)
MBv3	**5.5**	86.77	87.23	87.05	48.36	90.83	44.34
+CA+DecoupledHead	7.1	90.34	91.76	91.04	53.17	95.01	50.14
+Focus	7.5	**91.51**	**92.02**	**91.76**	**55.45**	**95.11**	**54.23**

TABLE 8. The lightweight comparison on the HRSID dataset.

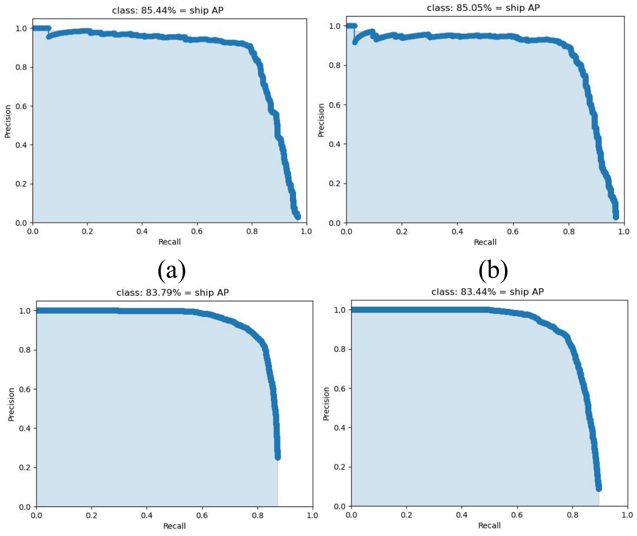
Method	Parameter (M)	Precision (%)	Recall (%)	F1-score (%)	$mAP_{0.5:0.95}$ (%)	$mAP_{0.5}$ (%)	$mAP_{0.75}$ (%)
MBv3	**5.5**	87.39	79.36	83.18	42.72	86.01	40.23
+CA+DecoupledHead	7.1	89.51	80.69	84.87	46.41	87.95	47.21
+Focus	7.5	**91.45**	**82.62**	**86.81**	**48.12**	**89.23**	**50.12**

dimensional space to extract the features. The residual connection structure introduced after the last layer's activation function is replaced by a linear function.

During the lightweight experiment, the backbone of the baseline network was replaced with the MobileNetV3. In addition, the coupling and attention mechanisms were used to improve the performance of lightweight networks. Figure 20 shows the P-R curve of the lightweight model. The results in Figures 20(a) and 20(b) obtained on the SSDD dataset show that after introducing the decoupled head and coordinate attention module, the mAP increased by more than 4% compared to the lightweight baseline. The results presented in Figures 20(c) and 20(d) obtained on the HRSID dataset showed that by introducing decoupled head and coordinate attention module, the mAP increased by approximately 2% compared to the lightweight baseline. This further indicated that the decoupled head and coordinate attention module could be successfully applied to a lightweight network. Moreover, the number of parameters of the lightweight model was 7.1M, accounting for only 9.1% of the parameters of the model presented in Figure 3. The accuracy achieved on the SSDD dataset was lowered by 0.5%, whereas the accuracy on the HRSID dataset was reduced by approximately 5%. In addition, the computational complexity of our lightweight model was 3.52 GFLOPs, and its detection frame rate was 49.32 FPS.

Next, the focus layer was introduced to the lightweight model for further analysis, and the corresponding results are shown in Figure 21. Besides, Tables 7 and 8 show the comparison of various performance metrics of the MobileNetv3 lightweight baseline network, MobileNetv3 lightweight network with coordinate attention and decoupled head, and focus optimized network. Tables 7 and 8 show the results on the SSDD and HRSID datasets, respectively.

To minimize the number of parameters in the lightweight model, the original lightweight large and small models were analyzed. The small model had 1.8 M parameters, which was one 1/4 of the original lightweight model. The effectiveness was verified on the SSDD and HRSID datasets. Figure 22 depicts the P-R curves of the large and small models.



(a) (b)

(c) (d)

FIGURE 22. The P-R curve of the large and small, lightweight models. (a) The P-R curve of the large lightweight model on the SSDD; (b) the P-R curve of the small, lightweight model on the SSDD; (c) The P-R curve of the large lightweight model on the HRSID. (d) The P-R curve of the small, lightweight model on the HRSID.

VI. CONCLUSION

By introducing the well-known optimization strategy, this study improves the original YOLOv4's detection accuracy. However, the complex YOLOv4 structure is not conducive to mobile deployment. To address this problem, this paper proposes a decoupled head and coordinate attention method. On the basis of lightweighting the YOLOv4 network, the proposed method ensures good detection performance. In addition, a decoupled head is proposed to optimize model performance. Moreover, to address the problems of the channel attention system's inability to gather precise position information and CBAM's inability to capture long-range dependencies in the spatial domain, the coordinate attention module is added to the third stage with the most robust learning features, the fourth and fifth stages with higher semantics. Further, by introducing a two-way trunk, the detection performance of the model for small targets is further improved. According to the experimental results on two public datasets, compared to the other five SAR ship detectors based on CNN, the proposed decoupled head and coordinate attention method is feasible and has higher detection performance. Moreover, by using the proposed method, satisfactory detection results can be obtained in two large-scene images, indicating the excellent migration ability of the proposed model in marine monitoring.

The results presented in this study can be useful for further research on SAR ship detection.

REFERENCES

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

[2] 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, p. 21, 2019.

[3] F. Zhang and B. Wu, "A scheme for ship detection in inhomogeneous regions based on segmentation of SAR images," *Int. J. Remote Sens.*, vol. 29, no. 19, pp. 5733–5747, Oct. 2008.

[4] C. Wang, Z. Wang, H. Zhang, B. Zhang, and F. Wu, "A PolSAR ship detector based on a multi-polarimetric-feature combination using visual attention," *Int. J. Remote Sens.*, vol. 35, no. 22, pp. 7763–7774, 2014.

[5] C. P. Schwegmann, W. Kleynhans, and B. P. Salmon, "Synthetic aperture radar ship detection using Haar-like features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 154–158, Feb. 2017.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014.

[7] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[8] 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.

[9] 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)*, Seattle, WA, USA, Jun. 2016.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, Netherlands: Springer, Aug. 2016.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[12] T. W. Zhang, "SAR ship detection dataset (SSDD): Official release and comprehensive data analysis," *Remote Sens.*, vol. 13, no. 18, p. 41, 2021.

[13] R. Yang, G. Wang, Z. Pan, H. Lu, H. Zhang, and X. Jia, "A novel false alarm suppression method for CNN-based SAR ship detector," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1401–1405, Aug. 2021.

[14] Y. Y. Wang, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sens.*, vol. 11, no. 7, p. 14, 2019.

[15] M. Zhu, G. Hu, H. Zhou, C. Lu, Y. Zhang, S. Yue, and Y. Li, "Rapid ship detection in SAR images based on YOLOv3," in *Proc. 5th Int. Conf. Commun., Image Signal Process. (CCISP)*, Nov. 2020.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] V. Mnih, "Recurrent models of visual attention," in *Proc. 28th Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2014.

[18] K. Gregor, "DRAW: A recurrent neural network for image generation," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, 2015.

[19] K. Xu, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, 2015.

[20] M. Jaderberg, "Spatial transformer networks," in *Proc. 29th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2015.

[21] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017.

[22] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019.

[23] J. Hu, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Proc. 32nd Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2018.

[24] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2017.

[25] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019.

[26] B. Su, J. Liu, X. Su, B. Luo, and Q. Wang, "CFCANet: A complete frequency channel attention network for SAR image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11750–11763, 2021.

[27] H. Xue, M. Sun, and Y. Liang, "ECANet: Explicit cyclic attention-based network for video saliency prediction," *Neurocomputing*, vol. 468, pp. 233–244, Jan. 2022.

[28] H. Lee, H.-E. Kim, and H. Nam, "SRM: A style-based recalibration module for convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019.

[29] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Gated channel transformation for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020.

[30] S. H. Woo, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany: Springer, 2018.

[31] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel 'squeeze and excitation' blocks," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 540–549, Feb. 2019.

[32] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021.

[33] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021.

[34] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019.

[35] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020.

[36] 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.

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

[38] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, 2020.

[39] J. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *Remote Sens.*, vol. 99, pp. 1–14, Jun. 2020.

[40] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017.

[41] K. M. He, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland: Springer, 2014.

[42] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018.

[43] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020.

[44] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," 2020, *arXiv:2003.07540*.

[45] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[46] Z. Sun, M. Dai, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "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.

[47] Z. Sun, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 21, p. 4209, Oct. 2021.

[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

[49] M. X. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019.

[50] Y. Lee, J.-W. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and GPU-computation efficient backbone network for real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019.

[51] H. Peng and X. Tan, "Improved YOLOX's anchor-free SAR image ship target detection," *IEEE Access*, vol. 10, pp. 70001–70015, 2022.

[52] S. Wei, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.

[53] M. A. Günen, U. H. Atasever, and E. Besdok, "Analyzing the contribution of training algorithms on deep neural networks for hyperspectral image classification," *Photogrammetric Eng. Remote Sens.*, vol. 86, no. 9, pp. 581–588, Sep. 2020.

[54] M. A. Günen, "Performance comparison of deep learning and machine learning methods in determining wetland water areas using EuroSAT dataset," *Environ. Sci. Pollut. Res.*, vol. 29, no. 14, pp. 21092–21106, Mar. 2022.

[55] J. Zhuang, "AdaBelief optimizer: Adapting stepsizes by the belief in observed gradients," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2020.

[56] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," 2022, *arXiv:2201.03545*.

[57] T. W. Zhang, "LS-SSDD-v1.0: A deep learning dataset dedicated to small ship detection from large-scale Sentinel-1 SAR images," *Remote Sens.*, vol. 12, no. 18, p. 37, 2020.

[58] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018.

[59] N. N. Ma, "ShuffleNet v2: Practical guidelines for efficient CNN architecture design," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*. Munich, Germany: Springer, 2018.

[60] B. Wu, K. Keutzer, X. Dai, P. Zhang, Y. Wang, F. Sun, Y. Wu, Y. Tian, P. Vajda, and Y. Jia, "FBNet: Hardware-aware efficient ConvNet design via differentiable neural architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019.

[61] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019.

QINZUO LI is currently pursuing the M.S. degree with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include SAR ship detection and deep learning.

DENGJUN XIAO received the B.S. degree in electronic engineering from the University of Electronic Science and Technology of China, in 1999. He is currently a Senior Engineer at the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include radar system engineering, radio frequency, microwave systems, and broadband digital signal source.

FANGYING SHI is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences. Her research interest includes image processing.