An Anchor-Free Method Based on Transformers and Adaptive Features for Arbitrarily Oriented Ship Detection in SAR Images

Bingji Chen, Chunrui Yu, Shuang Zhao, and Hongjun Song

***Abstract*—Ship detection is a crucial application of synthetic aperture radar (SAR). Most recent studies have relied on convolutional neural networks (CNNs). CNNs tend to struggle in gathering adequate contextual information through local receptive fields and are also susceptible to noise. Inshore scenes in SAR images are plagued by substantial background noise, so achieving high-accuracy ship detection of arbitrary orientations within complex scenes remains an ongoing challenge when relying solely on CNNs. To address the above challenges, this article presents an anchor-free method based on transformers and adaptive features, namely, SAD-Det, which can detect rotationally invariant ship targets with high average precision in SAR images. Specifically, a transformer-based backbone network called the ship spatial pooling pyramid vision transformer is proposed to enhance the long-range dependencies and obtain sufficient contextual information for ships in SAR images. In addition, a neck network called the adaptive feature pyramid network is designed to enhance the ability of ship feature adaptation by adding fusion factors to feature layers in SAR images. Finally, a head network called the deformable head is constructed to make the network more adaptable to the characteristics of ships by adaptively detecting the spatial sampling positions of the targets in SAR images. The effectiveness of the proposed method is verified by experiments on two publicly available datasets, i.e., SAR ship detection dataset and rotated ship detection dataset in SAR images. Compared with other arbitrarily oriented object detection methods, the proposed method achieves state-of-the-art detection performance.**

***Index Terms*—Adaptive features, anchor-free, arbitrarily oriented detector, ship detection, synthetic aperture radar (SAR), transformer.**

I. Introduction

SYNTHETIC aperture radar (SAR) is a type of active microwave imaging radar. Unlike optical images, SAR is not affected by daylight or weather and can achieve all-day and all-weather observations of the Earth. It also has good information acquisition capability in complex situations [1]. Due to its excellent performance, SAR has been widely developed and utilized in various fields, including ocean monitoring [2], topographic mapping [3], agricultural monitoring [4], and disaster detection [5]. Among them, research on ship object detection based on SAR ocean images is an important application of SAR [6] and is of great significance in the fields of maritime supervision, fishery management, disaster rescue, etc.

The main characteristic of traditional ship detection methods in SAR images is manual feature extraction, which typically consists of several stages: land masking, preprocessing, prescreening, and discrimination [7]. These traditional detection methods mainly include methods based on a constant false alarm rate (CFAR) [8], methods based on a global threshold [9], methods based on visual saliency [10], methods based on wavelet transform [11], and methods based on polarization information [12]. Among them, the most widely used method is the CFAR method based on the sea clutter statistical distribution. Its basic idea is to statistically model the sea clutter around the pixels to be detected by sliding the window under the preset false alarm rate to adaptively determine the detection threshold and then compare the gray value of the pixel to be detected with the detection threshold to perform ship detection. However, these traditional methods, including CFAR, require manual design and the extraction of ship features, resulting in cumbersome algorithmic processes and low robustness. These drawbacks impede the advancement of detection performance, rendering them inadequate for contemporary ship detection tasks in SAR images.

In recent years, deep learning [13] has exhibited exceptional performance in various domains, such as image classification, object detection, semantic segmentation, and instance segmentation. Li et al. [14] released the first public SAR ship detection dataset, named SSDD, which enables end-to-end ship detection in SAR images based on deep learning. Currently, several neural network architectures are employed for image processing using deep learning, including convolutional neural networks (CNNs) [15] and transformers [16].

CNN-based object detection methods have been widely applied to ship detection in SAR images due to their excellent performance. In terms of two-stage detectors, multiple researchers have combined various improved modules and attention mechanisms based on Faster R-CNN [17], Cascade R-CNN [18],

Manuscript received 22 August 2023; revised 4 October 2023; accepted 12 October 2023. Date of publication 18 October 2023; date of current version 29 December 2023. *(Corresponding author: Hongjun Song.)*

Bingji Chen and Shuang Zhao are with the Department of Space Microwave Remote Sensing System, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: chenbingji21@mails.ucas.ac.cn; zhaoshuang18@mails.ucas.ac.cn).

Chunrui Yu is with the Beijing Institute of Tracking and Telecommunication Technology, Beijing 100094, China (e-mail: ycrzxc@163.com).

Hongjun Song is with the Department of Space Microwave Remote Sensing System, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: songhj@aircas.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3325573

and Mask R-CNN [19] to improve the performance of ship detection in SAR images. In terms of one-stage detectors, Lin et al. [20] designed a feature enhancement pyramid and shallow feature reconstruction network based on Retinanet [21] to mitigate the adverse effects of scattering noise in SAR images and improve detection accuracy for small ships. Zhang et al. [22] introduced a frequency attention module based on YOLO V5 [23], which can adaptively process the frequency domain information of SAR images to suppress noise such as sea clutter. Zhang et al. [24] proposed a network structure based on SSD [25], which takes raw SAR images and saliency maps as input and fuses their features in order to reduce the computational complexity and the number of network parameters (Params). The aforementioned detection algorithms are all based on the horizontal bounding box (HBB), which can effectively detect ships in SAR images. However, unlike natural scene images, remote sensing images, including SAR images, are usually obtained from a bird's-eye view and have distinctive characteristics [26]. For example, the ship features have arbitrary orientations, dense arrangements, scale variation, complex backgrounds, etc. Especially in ports, where a large number of ships are densely arranged and have a large aspect ratio, the HBB can introduce substantial interference from the background area and adjacent ships.

The oriented bounding box (OBB) is well suited for ship detection in SAR images because it preserves the directionality of the targets in the image, leading to more accurate positioning. For example, Zhao et al. [27] proposed a single-stage detection method that efficiently detects ships in any orientation in SAR images through multiscale feature fusion and calibration. Guo et al. [28] introduced a new encoding representation to describe the OBB and incorporated a feature adaptive module to refine each feature pyramid layer. Zhou et al. [29] presented a simple ellipse Params representation method for objects in any direction, utilizing the YOLOX [30] algorithm for directional ship detection and yielding favourable results. Zhou et al. [31] designed a new anchor-free keypoint-based detection method called KeyShip for high-precision detection of oriented ships in SAR images. These approaches are all CNN-based ship detection methods for SAR images. CNNs can obtain an image-specific inductive bias, including locality and translation equivariance, thus improving the ability to learn image features. It is worth noting that SAR images in inshore ship scenes are subject to severe background noise. However, the convolution operation in CNNs can only obtain local receptive fields, making it incapable of capturing the long-range dependencies that can bolster the representation ability, thereby limiting the utilization of contextual information [32]. In addition, CNNs are sensitive to geometric perturbations of images, such as random translations, rotations, and flips, which makes their anti-interference ability poor and their generalization ability weak [33].

In contrast to the convolution operation, the multihead self-attentions (MSAs) of the transformer can capture long-range dependencies [34]. Furthermore, related studies [35] have shown that the transformer exhibits strong adaptability to perturbations, occlusions, and domain shifts. In the field of object detection in remote sensing images, Yao et al. [36] proposed a new multimodal deep learning framework for land use and land cover classification tasks, which outperforms other backbone models based on transformers or CNNs. Li et al. [37] introduced a baseline network for hyperspectral anomaly detection, which combines a low-rank representation model with deep learning techniques to enhance the performance of hyperspectral anomaly detection. In the field of ship detection in SAR images, some researchers have also conducted relevant research based on transformers. For example, Xia et al. [38] utilized the Cascade Mask R-CNN as the basic architecture, combined with the Swin Transformer [39], to obtain a visual transformer framework based on contextual joint-representation learning. Li et al. [40] introduced a feature enhancement module based on the Swin Transformer to improve feature extraction capabilities, along with an adjacent feature fusion module to optimize feature pyramids for enhancing ship recognition and positioning capabilities in SAR images. Shi et al. [41] developed a deformable attention mechanism into a Swin Transformer and proposed a new contour-guided shape enhancement module to improve the accuracy of ship detection in SAR images. Zhou et al. [42] created an edge semantic decoupling module and integrated a transformer into the detection layer to achieve dense ship detection in inshore areas. The above algorithms are all based on the HBB, and a small number of researchers have explored detection methods based on the OBB. For example, Zhou et al. [43] modified the pyramid vision transformer (PVT) [44] model, designed a multiscale feature fusion module and adopted a new loss function to improve the detection ability for small targets and mitigate the influence of the ship's boundary scattering interference. However, the abovementioned transformer-based methods have limited context information and feature information based on ships in SAR images, so their detection performance needs to be further improved.

To address the above problems, this article presents an anchor-free method based on transformers and adaptive features for arbitrarily oriented ship detection in SAR images, called SAD-Det. The method is a hybrid structure of the transformer and CNN. First, to obtain more contextual information related to ships in SAR images, the ship spatial pooling pyramid vision transformer (SSP-PVT) is proposed. This module employs PVT to generate multilayer feature maps and then introduces the last layer of feature maps into the ship spatial pooling module (SSPM) to modulate the feature maps from different dimensions. Second, to improve the performance of ship target feature fusion, we design the adaptive feature pyramid network (AFPN). By incorporating the adaptive weight module (AWM) in different feature layers of the feature pyramid network (FPN), we leverage its self-attention characteristics to assign different weights to the feature layers of the neck network. Third, to make the network more adaptable to the characteristics of ships in SAR images, we propose the deformable head (DeHead) based on deformable convolution (DC), which enhances the adaptive ability of the detection head network. The residual connection is introduced to solve the problem of vanishing gradients.

The main contributions of this article are as follows.

1) An anchor-free method based on transformers and adaptive features for arbitrarily oriented ship detection in SAR

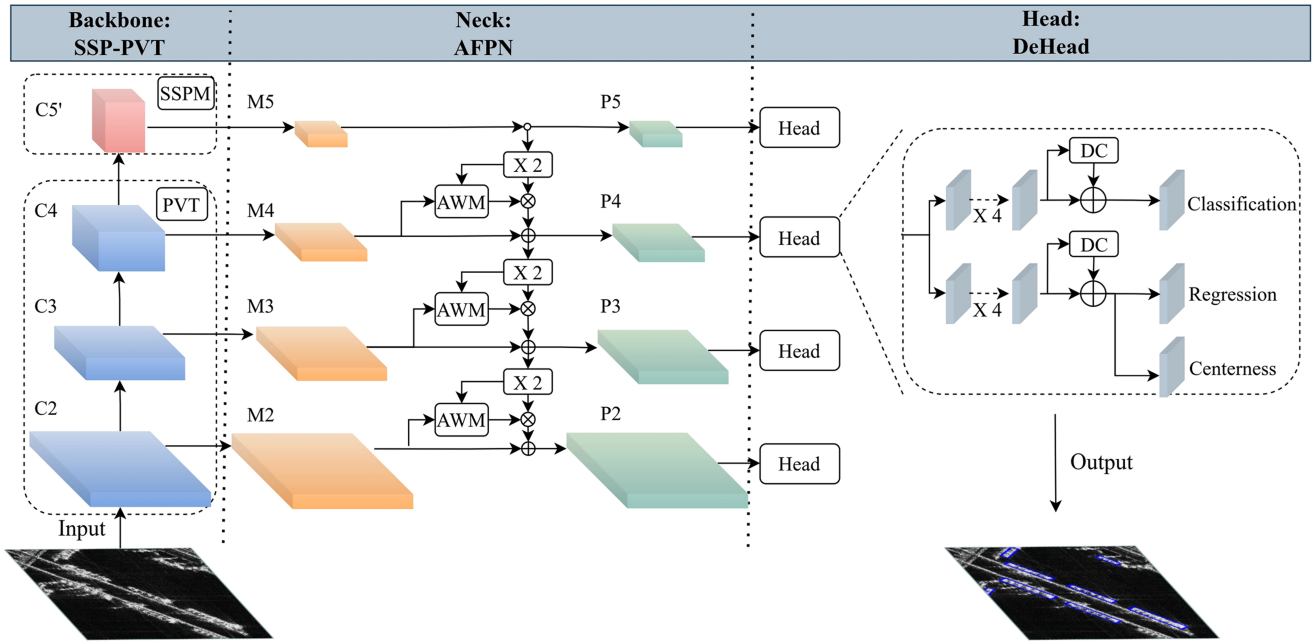


Fig. 1. Overall architecture of SAD-Det.

images is proposed, called SAD-Det. The detection performance of the arbitrarily oriented ship method for SAR images based on the anchor-free framework achieves the highest average precision, which demonstrates the potential of combining the transformer and CNN.

2) The SSP-PVT is proposed, which utilizes PVT and incorporates a module called the SSPM to enhance the long-range dependencies of ships in SAR images, thereby obtaining sufficient contextual information to improve the detection performance of ships in SAR images.
3) The AFPN is proposed, which incorporates the AWM into the neck network based on FPN. The aim of AWM is to add fusion factors to different feature layers, thereby assigning different weights to the feature layers in order to improve the performance of feature fusion for ships in SAR images. Furthermore, the DeHead is proposed, which implements DC to adaptively detect the spatial sampling positions of the targets and incorporates residual connections to optimize the network for ship characteristics in SAR images.
4) Extensive experiments on SSDD and rotated ship detection dataset in SAR images (RSDD-SAR) validate the effectiveness of the proposed module. Compared with other arbitrarily oriented object detectors, the proposed method achieves state-of-the-art detection performance.

The rest of this article is organized as follows. Section II provides a detailed description of the proposed method. Section III presents and analyses the experimental results. Finally, Section IV concludes this article.

II. Methodology

In this section, we first explain the overall architecture of the proposed method and then describe SSP-PVT of the backbone network, AFPN of the neck network, and DeHead of the head network. Finally, we illustrate the loss function used by this method.

A. Overall Architecture

The overall architecture of the proposed method, named SAD-Det, is illustrated in Fig. 1. The backbone network is SSP-PVT we propose, which is based on PVT, as depicted by the blue cube. To maintain consistency with the naming conventions of CNN-based backbone networks, such as ResNet, the feature maps outputted by SSP-PVT are named {C2,C3,C4,C5}. Then, the last layer of feature layer C5 is fed into the proposed SSPM, where it is transformed into a feature map C5' represented by the red cube. Therefore, the feature map generated by the backbone network becomes {C2,C3,C4,C5'}. Subsequently, the multi-layer feature map is input to the proposed AFPN. Initially, lateral connections are utilized to adjust them to a feature map {M2, M3, M4, M5} with an equal number of channels, as shown by the orange cube. Then, the top–down pathway and the proposed AWM are employed to merge the feature maps, resulting in the fused feature map. Finally, convolution operations are applied to obtain the output feature map {P2, P3, P4, P5}, depicted by the green cube. The feature map is then fed into the proposed DeHead, where DC and residual connections are utilized in the two branches to further process the feature map. Finally, the loss function is computed for three parts, namely, classification, regression, and centerness.

B. Ship Spatial Pooling Pyramid Vision Transformer

With the promising advancements of transformers in the field of natural language processing, researchers have extended transformers to the field of computer vision and achieved promising

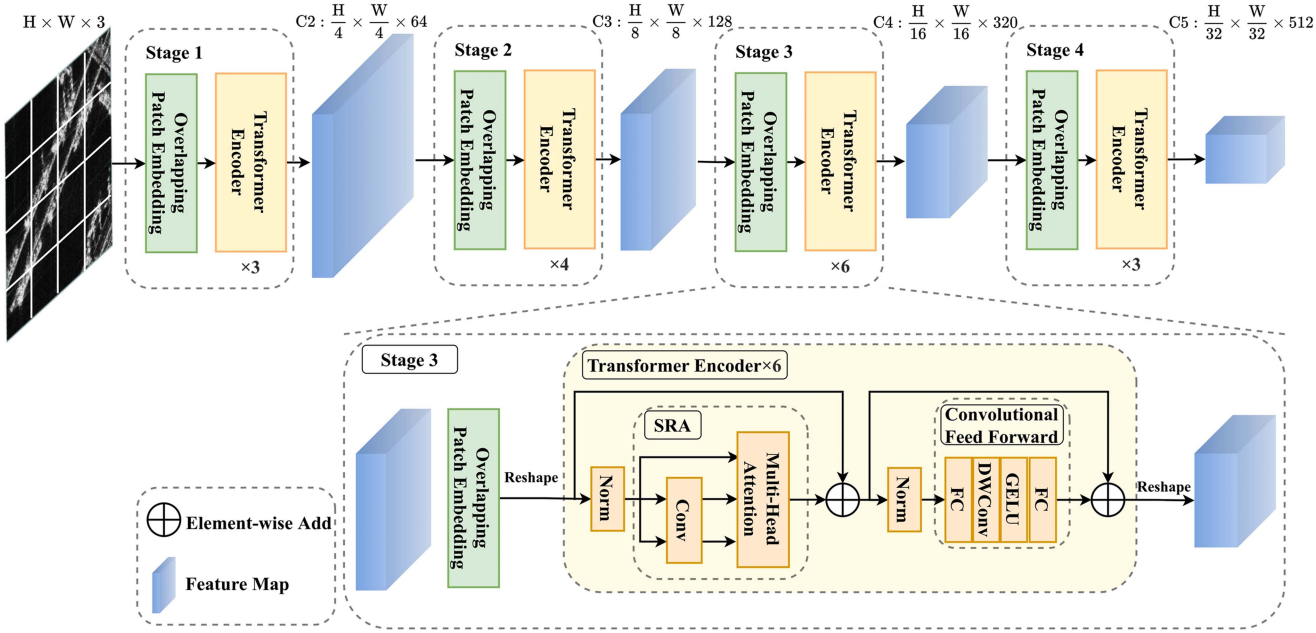


Fig. 2. Structure of PVT-V2-B2.

outcomes. For example, the Vision Transformer (ViT), proposed by Dosovitskiy et al. [45] demonstrated for the first time that a transformer could be applied to computer vision, and it achieved state-of-the-art performance in numerous experiments at that time. However, ViT is not a generic backbone, as it is only suitable for image classification, not for dense prediction tasks such as object detection and image segmentation.

PVT [44] is the first hierarchical design of the ViT, which incorporates a pyramid structure into the transformer, enabling seamless integration with downstream tasks, such as object detection. This design is similar to that of ResNet and other CNN-based backbone networks. PVT plays a crucial role among a multitude of outstanding backbone networks and has yielded superior outcomes compared with Swin-Transformer, which is another frequently employed backbone network. This will be elucidated through subsequent experiments and analyses. Unlike traditional convolutional backbones, PVT is a nonconvolutional backbone that consists of multiple independent transformer encoders stacked together. The resolution of the input image is gradually reduced through patch embedding. In addition, a related team proposed PVT-V2 [46], which reduces the computational complexity to linear and solves the problem of the high computational complexity of PVT for high-resolution images. PVT-V2 offers different models according to different parameter quantities and task requirements. To ensure a fair comparison with existing backbones, such as ResNet-50 [47] and Swin-T [39], we selected PVT-V2-B2, which possesses similar parameter quantities, as the baseline for the proposed method's backbone.

The structure of PVT-V2-B2 is illustrated in Fig. 2. Similar to the CNN-based backbone, PVT-V2-B2 contains four stages, each of which consists of a patch embedding layer and some transformer encoder layers, producing feature maps at various scales. In the first stage, the input image is set to a size of $H \times W \times 3$, and then it is divided into $\frac{HW}{4^2}$ patches, where each patch has a size of $4 \times 4 \times 3$. Then, the patches are fed into an overlapping patch embedding layer, and multiple embedded patches with a size of $\frac{HW}{4^2} \times 64$ can be obtained. Subsequently, the embedded patches are passed through three consecutive transformer encoder layers, followed by a reshaping operation to generate a feature map C2 with a size of $\frac{H}{4} \times \frac{W}{4} \times 64$. In comparison with [16], a transformer encoder layer of PVT-V2-B2 contains a spatial-reduction attention layer, which means adding convolutions before the multihead attention layer to reduce the computational complexity, and a convolutional feed forward layer, which means adding a depthwise convolution between the first FC layer and the GELU layer. By applying a similar approach to the remaining three stages, we can obtain feature maps C3, C4, and C5 with different sizes.

Compared with other targets, ship targets in SAR images have a substantial aspect ratio. PVT-V2-B2 is a general backbone network that usually detects input feature maps within a square window. However, it lacks specific optimization for ship targets, leaving room for improvement in its detection performance. Inspired by [48], we propose the SSPM to enhance the long-range dependencies of ships in SAR images. The specific structure of SSPM is illustrated in Fig. 3. We first set the feature map C5 as the input $x \in \mathbb{R}^{H' \times W' \times C'}$ and then pass x into two branches. Each branch contains a horizontal or vertical strip pooling layer. The outputs of these layers are labelled $y^h \in \mathbb{R}^{H' \times C'}$ and $y^v \in \mathbb{R}^{W' \times C'}$, respectively. The specific formulas are as follows:

$$y_{i,c}^{h} = \frac{1}{W'} \sum_{0 \le j \le W'} x_{i,j,c} \tag{1}$$

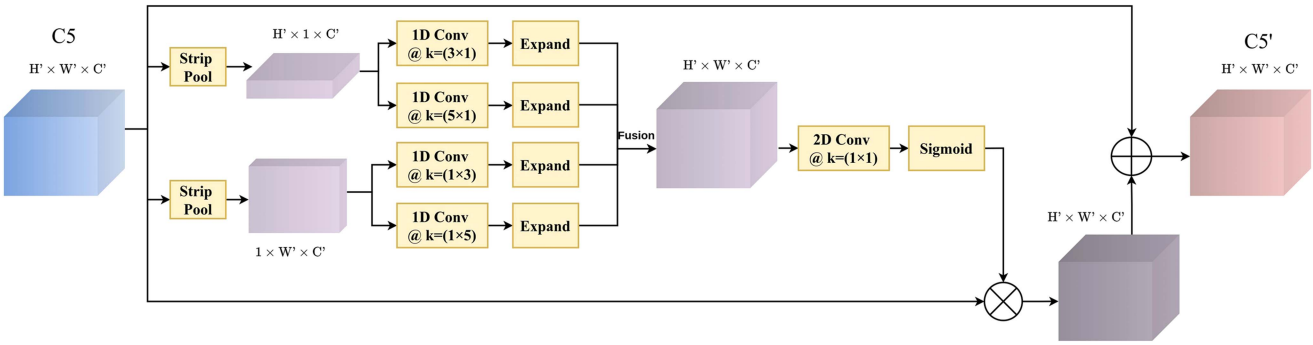


Fig. 3. Structure of SSPM.

$$y_{j,c}^{v} = \frac{1}{H'} \sum_{0 \le i \le H'} x_{i,j,c}. \tag{2}$$

Then, the feature map passes through two 1-D convolutional layers with kernel sizes of three and five, restores the feature map to its original size through the expansion operation, and performs elementwise addition on the feature map. The specific process is as follows:

$$y_{i,c}^{h} = \mathrm{EP}(\mathrm{Conv}_{3\times1}(y_{i,c}^{h})) + \mathrm{EP}(\mathrm{Conv}_{5\times1}(y_{i,c}^{h})) \tag{3}$$

$$y_{j,c}^{v} = \mathrm{EP}(\mathrm{Conv}_{1\times3}(y_{j,c}^{v})) + \mathrm{EP}(\mathrm{Conv}_{1\times5}(y_{j,c}^{v})) \tag{4}$$

$$y_{i,j,c} = y_{i,c}^{h} + y_{j,c}^{v} \tag{5}$$

where $\mathrm{Conv}_{3\times1}$, $\mathrm{Conv}_{5\times1}$, $\mathrm{Conv}_{1\times3}$, and $\mathrm{Conv}_{1\times5}$ represent convolutional layers of sizes 3×1, 5×1, 1×3 and 1×5, respectively. EP represents the expansion operation. After that, the output z is calculated as follows:

$$z = x + \mathrm{Prod}(x, \sigma(\mathrm{Conv}_{1\times1}(y))) \tag{6}$$

where $\mathrm{Conv}_{1\times1}$ represents the convolutional layer of size 1×1. σ represents the sigmoid function, and Prod represents the elementwise product. At the same time, the residual connection [47] is introduced to improve the detection results. Through the above process, the feature map C5 is transformed into C5'. In conclusion, the output feature map of SSP-PVT is composed of {C2,C3,C4,C5'}.

C. Adaptive Feature Pyramid Network

Feature fusion is a crucial step in many object detection tasks. FPN [49] is a typical feature fusion method that is widely used in various object detection networks. It can connect deep and shallow feature maps through top–down pathways and lateral connections and subsequently fuses objects of different scales to enhance the detection performance. However, the feature fusion method represented by FPN simply involves element-by-element addition or concatenation of features within the channel dimension without fully considering the characteristics of the specific detection tasks. In the field of ship detection in SAR images, ships are typically small in size, as observed in several publicly available datasets. Therefore, investigating methods of enhancing detection performance for small targets represented by ships is of great significance.

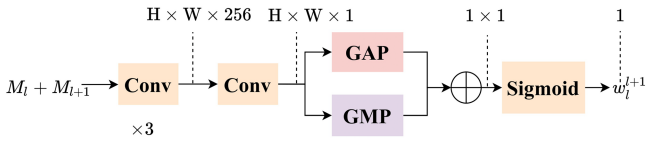


Fig. 4. Structure of AWM.

Adding fusion factors between adjacent feature layers to assign different weights to each layer is an efficient approach to enhancing the performance of FPN. Inspired by [50], we propose the AWM depicted in Fig. 4. The overall neck network is shown in Fig. 1, which we call the AFPN.

Considering that the majority of the ships in a SAR image are small in size and that the shallow feature map contains more information about small targets, we modified the input of the neck from the usual {C3,C4,C5'} to {C2,C3,C4,C5'}. The input feature map is first adjusted to the middle feature map {M2, M3, M4, M5} with a channel number of 256 through lateral connections. M5 then doubles the values of H and W through interpolation and performs scalar multiplication with the weight w_4^5 obtained by AWM. The result is then added to M4 to create a new feature map M4. This new feature map M4 is passed through a 3×3 convolution to obtain the output feature map P4. Similarly, the feature maps P2 and P3 can be obtained, and P5 can be directly obtained from M5 through a 3×3 convolution. The formulas are presented as follows:

$$M_l = \mathrm{Conv}_{1\times1}(C_l), \qquad l = 2,3,4,5 \tag{7}$$

$$M_l = M_l + w_l^{l+1} \times \mathrm{Upsample}(M_{l+1}), \qquad l = 2,3,4 \tag{8}$$

$$P_l = \mathrm{Conv}_{3\times3}(M_l), \qquad l = 2,3,4,5 \tag{9}$$

where $\mathrm{Conv}_{1\times1}$ and $\mathrm{Conv}_{3\times3}$ represent convolutional layers of sizes 1×1 and 3×3, respectively, and Upsample represents the nearest neighbour interpolation. w_l^{l+1} represents the weight factor resulting from AWM, as shown in Fig. 4. The input value is obtained by the elementwise addition of the feature maps M_l and M_{l+1}, and then the feature factor w_l^{l+1} is obtained by convolution, global average pooling, global max pooling,

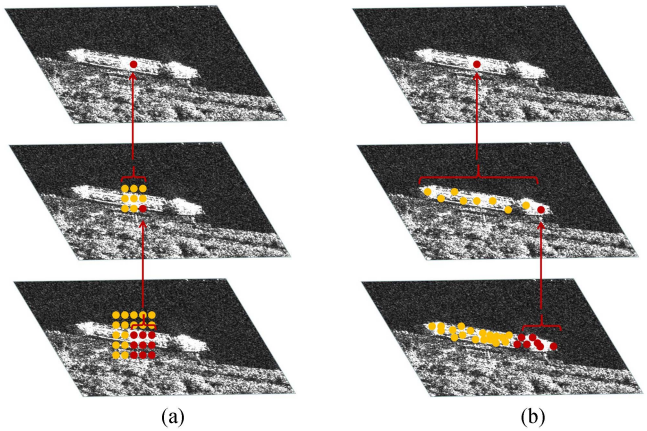


Fig. 5. Comparison of standard convolution and DC on the receptive field. (a) Standard convolution. (b) DC.

sigmoid, and other methods. The formula is as follows:

$$w_l^{l+1} = f(M_l + \text{Upsample}(M_{l+1})), \qquad l = 2, 3, 4 \tag{10}$$

where f represents AWM. After the above process, the output feature map of AFPN consists of {P2, P3, P4, P5}.

D. Deformable Head

As previously mentioned, the ship target in a SAR image has a large aspect ratio, and because standard convolution samples the fixed position of the input feature map, it cannot meet the precise requirements of ship detection in SAR images. Therefore, it is necessary to adaptively adjust the size of the receptive field to achieve more accurate positioning. DC [51], [52] can mitigate this problem. Fig. 5 shows a comparison between standard convolution and DC. The area represented by each point in Fig. 5(a) and (b) indicates the location of the receptive field. Our focus is on the activation point of the uppermost feature map, which is set at the center of the ship. This point needs to be derived from the nine points marked by the feature map of the middle layer through 3×3 convolution. These nine points need to be derived from multiple points of the feature map in the lowest layer through 3×3 convolution. After comparison, it is evident that the standard convolution in Fig. 5(a) has a fixed receptive field, while the DC in Fig. 5(b) has an adaptive receptive field. The sampling positions of DC are more in line with the shape and size of the object itself.

The head design of the proposed method is illustrated in Fig. 1. Similar to the head network of FCOS [53], the head of the proposed method is shared among different feature layers. However, a notable difference is the inclusion of an additional angle branch to facilitate the merging process in the regression task. The head of each layer consists of two branches: one branch predicts classification tasks, while the other predicts regression and centerness tasks. The four-layer feature map output by the neck network is used as the input of the head network. Initially, the feature maps pass through four 3×3 convolutional layers with 256 channels. Subsequently, they are passed through a deformable convolutional layer, which takes into account the residual connection [47] that we propose to improve the detection results. Finally, a 3×3 convolutional layer, with respective channel numbers of 1, 5, and 1, is employed for classification tasks, regression tasks, and centerness tasks.

E. Loss Function

The loss function of the proposed method consists of three parts, which can be calculated as

$$\begin{aligned} \text{Loss} = & \frac{1}{N_{\text{pos}}} \sum_{x,y} L_{\text{cls}}(p_{x,y}, c^*_{x,y}) \\ & + \frac{\lambda_1}{N_{\text{pos}}} \sum_{x,y} 1_{\{c^*_{x,y}>0\}} L_{\text{reg}}(t_{x,y}, t^*_{x,y}) \\ & + \frac{\lambda_2}{N_{\text{pos}}} \sum_{x,y} 1_{\{c^*_{x,y}>0\}} L_{\text{ctrness}}(s_{x,y}, s^*_{x,y}) \end{aligned} \tag{11}$$

where L_{cls} represents the classification loss using the focal loss function [21]. L_{reg} represents the regression loss using the rotated intersection over union (IoU) loss function [54]. L_{ctrness} represents the centerness loss using the cross entropy loss function [55]. N_{pos} represents the number of positive samples. λ_1 and λ_2 represent the balance weights for L_{reg} and L_{ctrness}, respectively, and both default to 1. $1_{\{c^*_{x,y}>0\}}$ is 1 when the (x,y) point in the feature map is identified as a positive sample; otherwise, it is 0. $p_{x,y}$ indicates the score predicted as a ship at that point. $c^*_{x,y}$ represents the real label corresponding to that point, with 1 for a ship and 0 for the background. $t_{x,y}$ represents the target bounding box information predicted at the point, and $t^*_{x,y}$ represents the real target bounding box information corresponding to the point. $s_{x,y}$ represents the predicted centerness at that point, and $s^*_{x,y}$ represents the true centerness corresponding to that point.

III. Experimental Results

A. Datasets and Experimental Settings

We use two datasets to evaluate the performance of the proposed method, namely, SSDD [14], [56] and RSDD-SAR [57].

SSDD is the first public ship dataset for SAR images. It has two versions, and we specifically chose the 2021 version due to its unified OBB annotations and detailed usage standards compared with the 2017 version. This dataset contains a total of 1160 images depicting 2456 ships. It encompasses various polarization methods, resolutions, and image sizes. As one of the most widely used ship datasets in SAR images, SSDD serves as a baseline for evaluating the performance of the proposed method. Specific details can be found in Table I.

RSDD-SAR is one of the few publicly available ship datasets with OBB annotations for SAR images. It comprises 7000 images, encompassing multiple imaging modes, polarization modes, and resolutions. The images in this dataset contain a total of 10 263 ships. The model trained by this dataset demonstrates a strong generalization ability, thus making RSDD-SAR an ideal choice for evaluating the performance of ship detection based on the OBB. Details can be found in Table I.

TABLE I
DETAILS OF SSDD AND RSDD-SAR

Dataset	SSDD	RSDD-SAR
Date	2017 or 2021	2022
Sensors	RadarSat-2, TerraSAR-X, Sentinel-1	Gaofen-3, TerraSAR-X
Polarization	HH, VV, VH, HV	HH, VV, VH, HV, DH, DV
Resolution (m)	1–15	2–20
Scenes	Inshore, Offshore	Inshore, Offshore
Image Size (pixels)	160–668	512×512
Image Number	1160	7000
Ship Number	2456	10263
Annotation	Vertical, Oriented, Polygon	Oriented

The SSDD dataset comprises images of various sizes, which we resized to 500 × 350 pixels for our experiments. Following the recommendation in [56], we divided the images into a training set and a testing set using an 8:2 ratio. Regarding the RSDD-SAR dataset, the default image input size is specified as 512 × 512 pixels in [57]. This dataset is randomly split into a training set and a test set using a 5:2 ratio.

The experiment involving the proposed method employed the AdamW optimizer with an initial learning rate of 0.0001 and a weight decay of 0.0005. The batch size was set to 8. For training on the SSDD dataset, the model underwent 120 epochs, with the learning rate decreasing by 0.1 times at 60, 90, and 110 epochs. On the other hand, the training of the RSDD-SAR dataset involved 36 epochs, with the learning rate decreasing by 0.1 times at 24 and 33 epochs. The other methods utilized a stochastic descent optimizer with an initial learning rate of 0.001, a weight decay of 0.0001, and a momentum of 0.9. All experiments were implemented on the following software and hardware platforms. The software platform employed a toolbox called MMRotate [58], which is based on the deep learning framework PyTorch, and was run on the Ubuntu 22.04 operating system. The hardware platform consisted of a computer equipped with an Intel Core i9-12900KF processor and two Nvidia GeForce RTX 3090 GPUs.

B. Evaluation Metrics

To quantitatively evaluate the performance of the proposed method, we utilized three evaluation metrics from the MS COCO dataset [59]: AP, AP50 and AP75. In addition, we used floating point operations (FLOPs) and Params to measure the time complexity and space complexity of the model. AP is the average precision calculated from IoU values ranging from 0.5 to 0.95 with a step size of 0.05. AP50 represents the average precision at an IoU threshold of 0.5, aligning with the evaluation metric used in Pascal VOC [60]. AP75 represents the average precision at an IoU threshold of 0.75.

Generally, the definition of AP is the area under the precision-recall curve (PRC). A higher AP value indicates better detection performance of the method. The formula for AP is given by

$$\mathrm{AP} = \int_0^1 P(R)dR \tag{12}$$

where P represents the precision, which is the proportion of correctly detected ship samples to all predicted positive ship samples. R represents the recall, which is the proportion of correctly detected ship samples to all labelled positive ship samples. The formulas for both are as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{13}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{14}$$

where TP represents true positives, referring to the correctly detected positive samples. FP represents false positives, indicating incorrectly detected positive samples. FN represents false negatives, denoting the incorrectly detected negative samples. Given the IoU threshold, and setting recall as the abscissa and precision as the ordinate, we can obtain the PRC mentioned earlier.

C. Analysis of the Proposed Method

1) Ablation Experiments: To verify the effectiveness of the proposed method, we gradually incorporate each module into the baseline and analyze the impact of each module on SSDD. The corresponding results are presented in Table II. The baseline network we utilize is FCOS (OBB), which incorporates an angle branch solely to facilitate the detection of rotationally invariant objects, in contrast to FCOS. To ensure a fair comparison, we modify the neck feature layer of the baseline network from {P3, P4, P5, P6, P7} to {P2, P3, P4, P5} as employed by the proposed method, and we adopt the same training strategy as the proposed method. Unless otherwise specified, the following experiments are conducted on SSDD by default.

According to the results presented in Table II, some conclusions can be drawn from the analysis. The three proposed modules demonstrate a significant improvement in detection performance compared with the baseline network, thus validating their effectiveness. Furthermore, the proposed method exhibits an increase of 3.8%, 2.4%, and 7.6% in the value of AP, AP50, and AP75, respectively, compared with the baseline network. In addition, it is evident from the values of FLOPs and Params in the table that SSP-PVT has similar space complexity and time complexity compared with ResNet-50 from the baseline network, thereby demonstrating the effectiveness of the transformer-based architecture in ships in SAR images. Moreover, the other two modules incorporate adaptive methods that aid in detecting ship features. However, this leads to an increase in the parameter quantity, consequently affecting the high-speed performance of detection. Fig. 6 illustrates the corresponding PRCs for different module combinations at an IoU of 0.75 and provides a more intuitive representation of the benefits brought by each module. Subsequently, a detailed analysis is conducted to examine the impacts of the three proposed modules.

2) Effect of SSP-PVT: Next, ablation experiments and extensive discussions are conducted for SSP-PVT. To comprehensively examine the impact of SSP-PVT, it is compared against other widely employed backbone networks, including ResNet-50, Swin-T, and PVT-V2-B2. The experimental results

TABLE II
ABLATION EXPERIMENTS WITH THE PROPOSED METHOD ON SSDD

	SSP-PVT	AFPN	DeHead	AP	AP50	AP75	FLOPs (G)	Params (M)
Baseline	–	–	–	55.5%	94.4%	60.3%	95.94	31.37
SAD-Det	√	–	–	58.2%	95.9%	65.6%	92.01	37.59
	√	√	–	58.4%	96.7%	66.3%	118.25	44.55
	√	√	√	**59.3%**	**96.8%**	**67.9%**	125.61	45.39

The bold items indicate the best value for each column.

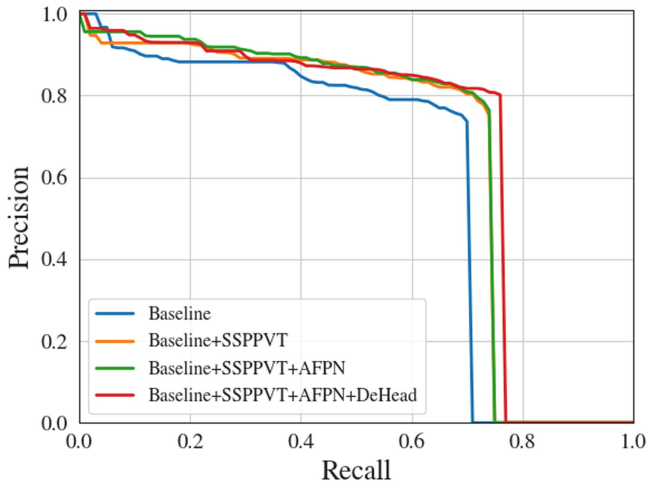


Fig. 6. PRCs with IoU = 0.75 when adding different modules.

TABLE III
COMPARISON OF PERFORMANCE BETWEEN SSP-PVT AND OTHER BACKBONE NETWORKS

Type	AP	AP50	AP75	FLOPs (G)	Params (M)
ResNet-50	55.5%	94.4%	60.3%	95.94	31.37
Swin-T	54.8%	94.7%	59.8%	97.71	34.99
PVT-V2-B2	57.8%	95.7%	64.4%	91.86	32.21
SSP-PVT	**58.2%**	**95.9%**	**65.6%**	92.01	37.59

The bold items indicate the best value for each column.

are presented in Table III. Each experiment utilizes FPN as the neck network and FCOS (OBB) head as the head network. The similarity in the values of AP, FLOPs, and Params between Swin-T and ResNet-50 validates the viability of the transformer architecture in ship detection for SAR images. Furthermore, PVT-V2-B2 exhibits superior performance compared to Swin-T, resulting in increments of 3.0%, 1.0%, and 4.6% for the value of AP, AP50, and AP75, respectively. This phenomenon is likely due to Swin-T's utilization of windows -head self-attentionMSA for self-attention in local areas at the expense of direct global relationship modeling, an essential feature of ViTs. In addition, SSP-PVT outperforms PVT-V2-B2, yielding increments of 0.4%, 0.2%, and 1.2% for the value of AP, AP50, and AP75, respectively. This is likely because the inclusion of SSPM strengthens the long-range dependencies of ships in SAR images. These experiments demonstrate that SSP-PVT can efficiently extract the features of ships in SAR images.

We also represent these methods visually. Fig. 7 depicts some sparse small targets in an offshore scene of the SSDD dataset. The detection results using ResNet-50 and Swin-T as the backbone network reveal the presence of some false positives. However, when utilizing PVT and SSP-PVT as the backbone network, all objects are detected accurately, thereby intuitively verifying the effectiveness of this module.

TABLE IV
COMPARISON OF PERFORMANCE BETWEEN AFPN AND OTHER NECK NETWORKS

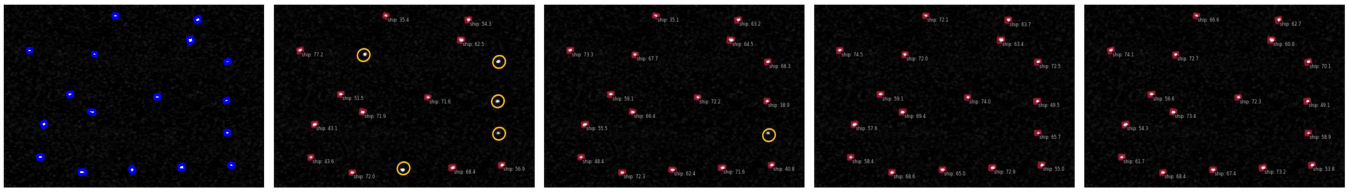
Type	AP	AP50	AP75	FLOPs (G)	Params (M)
FPN	57.8%	95.7%	64.4%	91.86	32.21
PAFPN	57.2%	94.8%	65.1%	96.22	35.75
BiFPN	57.0%	95.6%	62.5%	95.78	34.21
AFPN	**57.9%**	**96.4%**	**65.5%**	118.1	39.30

The bold items indicate the best value for each column.

3) Effect of AFPN: Subsequently, ablation experiments and further discussion of AFPN are conducted. AFPN is compared with other commonly used neck networks, and the results are presented in Table IV. For these experiments, considering the influence of the transformer on the experiment, the backbone network of each experiment utilizes PVT-V2-B2, and the head network adopts FCOS (OBB) head. The results demonstrate that FPN-based variants, such as PAFPN [61] and BiFPN [62], are not as effective as the baseline structure FPN, leading to decreased values of AP, AP50, and AP75. This is likely because the complex neck networks, involving more upsampling and downsampling, result in the loss of feature information, thereby negatively impacting detection result. However, when compared with FPN, AFPN exhibits improved performance with increases of 0.1%, 0.7%, and 1.1% in the value of AP, AP50, and AP75, respectively. This improvement can be attributed to AWM incorporated in AFPN, which adaptively changes the weight of each feature map based on the specific characteristics of each group of images, thus achieving more accurate detection. It is worth noting that due to the inclusion of AWM, the FLOPs value and Params value in AFPN are slightly higher than those in FPN, mildly affecting the network's high-speed performance.

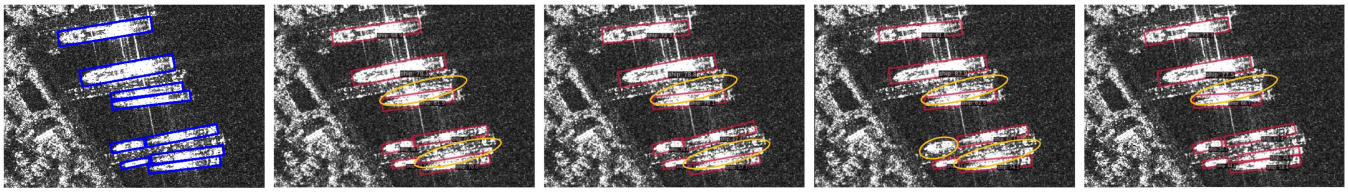
The visualization of these methods is given in Fig. 8, depicting an inshore scene with densely packed targets from the SSDD dataset. It is evident that all methods exhibit various levels of missed detections. Notably, AFPN demonstrates superior performance compared with other methods, with the fewest false positives in the detection results of the neck network. This illustrates the effectiveness of AFPN, but there is still potential for improvement in dense scenes.

4) Effect of DeHead: In the following analysis, we examine DeHead. The head network of FCOS (OBB) is used as a baseline and compared with DeHead. The results are presented in Table V. In each experiment, the backbone network utilizes PVT-V2-B2, while the neck network adopts FPN. It can be seen



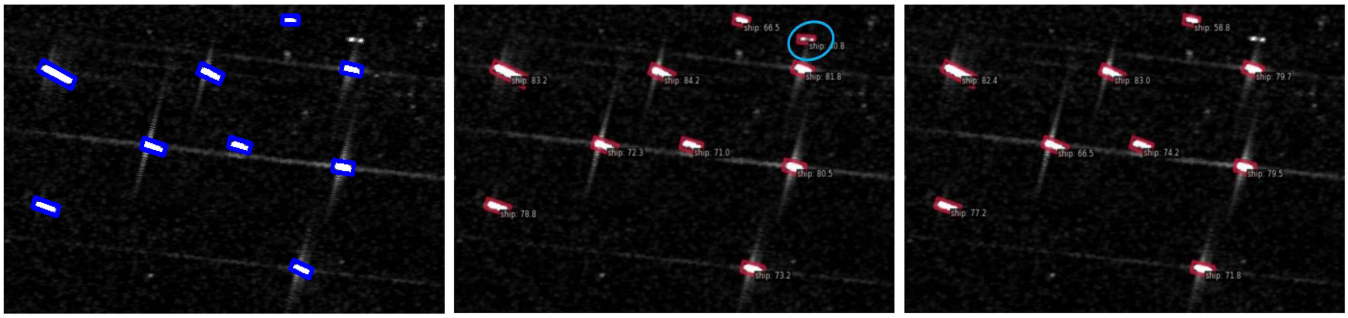
(a) (b) (c) (d) (e)

Fig. 7. Visualization of the detection results for different backbone networks. (a) Ground truth. (b) ResNet-50. (c) Swin-T. (d) PVT-V2-B2. (e) SSP-PVT. The blue boxes indicate the ground truth, the red boxes indicate true positives, and the orange circles indicate false positives.



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

Fig. 8. Visualization of the detection results for different neck networks. (a) Ground truth. (b) FPN. (c) PAFPN. (d) BiFPN. (e) AFPN. The blue boxes indicate the ground truth, the red boxes indicate true positives, and the orange circles indicate false positives.



(a) (b) (c)

Fig. 9. Visualization of the detection results for different head networks. (a) Ground truth. (b) Baseline. (c) DeHead. The blue boxes indicate the ground truth, the red boxes indicate true positives, and the light blue circles indicate false negatives.

TABLE V
COMPARISON OF PERFORMANCE BETWEEN DEHEAD AND THE BASELINE

Type	AP	AP50	AP75	FLOPs(G)	Params(M)
Baseline	57.8%	95.7%	64.4%	91.86	32.21
DeHead	**58.4%**	**95.9%**	**65.8%**	99.31	33.05

The bold items indicate the best value for each column.

that DeHead outperforms the baseline, resulting in the value of AP, AP50, and AP75 increasing by 0.4%, 0.2%, and 1.2%, respectively. This improvement stems from the introduction of DC, which enhance the network's adaptability to ship characteristics through adaptive detection of the objects' spatial sampling positions. In addition, the introduction of residual connections at deeper positions alleviates the issue of vanishing gradients. Furthermore, the inclusion of modules results in a slight increase in time complexity and space complexity.

Different head networks are also shown, revealing a sparse small target in an offshore scene of the SSDD dataset, as depicted in Fig. 9. It is evident that the baseline's head network exhibits false negatives, whereas DeHead network displays none, thereby substantiating the validity of the proposed module from a qualitative perspective.

D. Comparison With General Arbitrarily Oriented Object Detection Methods

To further assess the effectiveness of the proposed method, this section compares the proposed SAD-Det with several commonly used arbitrarily oriented object detection methods on both SSDD and RSDD-SAR. These methods include two-stage and anchor-based algorithms, such as Oriented R-CNN [63], Faster R-CNN (OBB) [64], and ReDet [65]. In addition, one-stage and anchor-based algorithms, such as R3Det [66] and S2A-Net [67],

TABLE VI
DETECTION RESULTS OF DIFFERENT ARBITRARILY ORIENTED OBJECT DETECTION METHODS ON SSDD

Methods	AP	AP50	AP75	FLOPs (G)	Params (M)
Two-stage, anchor-based					
Oriented R-CNN	51.9%	90.4%	55.0%	47.85	41.13
Faster R-CNN(OBB)	43.1%	86.2%	40.7%	47.83	41.12
ReDet	54.8%	91.6%	63.6%	36.80	31.54
One-stage, anchor-based					
R3Det	46.4%	89.2%	43.6%	56.57	41.58
S2A-Net	52.3%	92.7%	56.2%	33.76	36.18
One-stage, anchor-free					
FCOS(OBB)	52.2%	91.4%	57.3%	35.47	31.89
Based on a transformer					
SAD-Det	**59.3%**	**96.8%**	**67.9%**	125.61	45.39

The bold items indicate the best value for each column.

as well as one-stage and anchor-free algorithms, such as FCOS (OBB) [53].

The results of different arbitrarily oriented object detection methods on SSDD are presented in Table VI. The results demonstrate that SAD-Det outperforms the other methods in terms of AP, AP50, and AP75, achieving respective values of 59.3%, 96.8%, and 67.9%. Taking the AP50 indicator as an example, SAD-Det's performance witnessed a noticeable improvement of 5.2%, 4.1%, and 5.4% compared with that of ReDet, S2A-Net, and FCOS (OBB), respectively. These experimental findings verify that SAD-Det, which combines transformer and CNN architectures, yields superior detection outcomes relative to other CNN-based arbitrarily oriented detection methods and can even rival the object detection performance of HBB-based methods. It is evident that the proposed method has higher time and space complexity compared with the other methods, as indicated by its FLOPs and Params. The primary reason for this discrepancy is that the backbone network of the proposed method produces feature maps {C2, C3, C4, C5}, while other methods that neglect the largest feature map C2 only output feature maps {C3, C4, C5}. Furthermore, it should be noted that the calculation of FLOPs is partly influenced by the area of the feature maps. Consequently, although the value of FLOPs of the proposed method is significantly higher than that of other methods, it is still within an acceptable range. Moreover, Fig. 10 illustrates the PRCs corresponding to these methods at IoU = 0.75, enabling an enhanced understanding of their performance and further substantiating the advantages of the proposed method from an alternative perspective.

We select several methods for visualization. The pictures consist of two offshore scenes and two inshore scenes from the SSDD dataset, as depicted in Fig. 12. Fig. 12(a) portrays sparse small objects against an ocean background, and all methods achieve good detection results for this particular scenario. In contrast to a few methods that produce a small number of FPs or FNs, other methods, including SAD-Det, accurately detect all objects. In Fig. 12(b), in addition to sparse small targets, the scene contains disturbances, such as islands. At this juncture, most methods generate a considerable number of FPs and FNs. Nonetheless, SAD-Det only has one FP, highlighting its superior

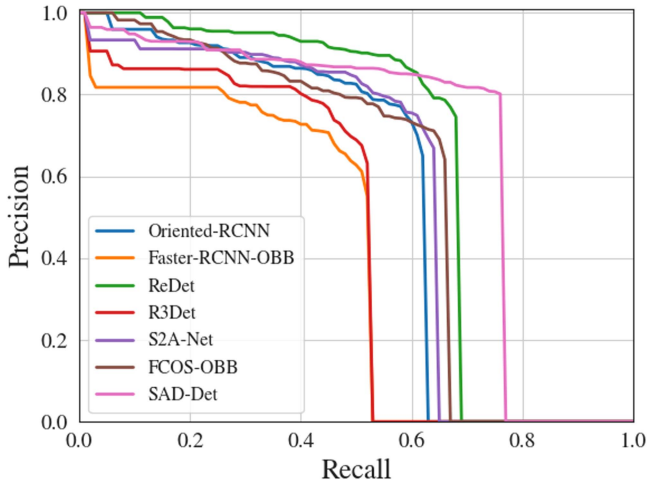


Fig. 10. PRCs of different methods on SSDD when IoU = 0.75.

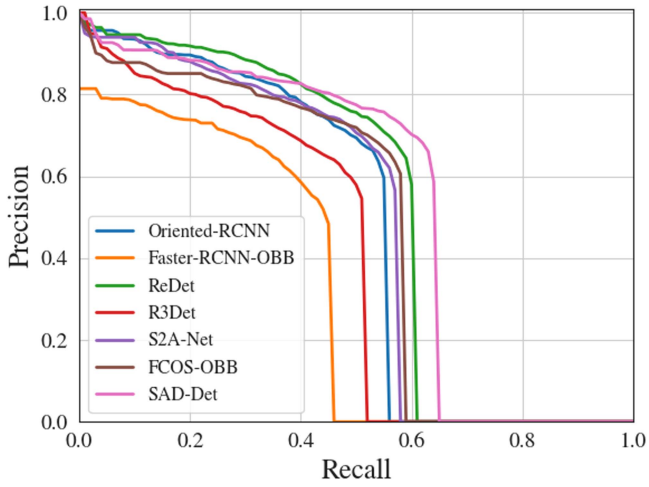


Fig. 11. PRCs of different methods on RSDD-SAR when IoU = 0.75.

resistance to interference compared with that of other methods. Fig. 12(c) portrays dense large-scale ship targets in an inshore scene. In contrast to other methods, the proposed method detected all targets accurately, showcasing its effectiveness in this type of environment. In Fig. 12(d), dense ship targets are once again present in an inshore scene, with ships positioned close to the shore, thereby increasing the difficulty of detection. All methods exhibit a high number of FPs, yet SAD-Det demonstrates superior detection accuracy. This further emphasizes the need to enhance the proposed method's resilience to interference in inshore scenes with dense targets.

To evaluate the generalization ability of the proposed method, experiments are conducted on RSDD-SAR. The results of different arbitrarily oriented object detection methods can be seen in Table VII. Compared with the other methods, SAD-Det achieves high scores in the AP, AP50, and AP75 indicators, reaching 52.0%, 93.8%, and 53.9%, respectively. For instance, in terms of the AP50 indicator, SAD-Det outperforms ReDet, S2A-Net, and FCOS (OBB) with improvements of 3.1%, 2.9%, and

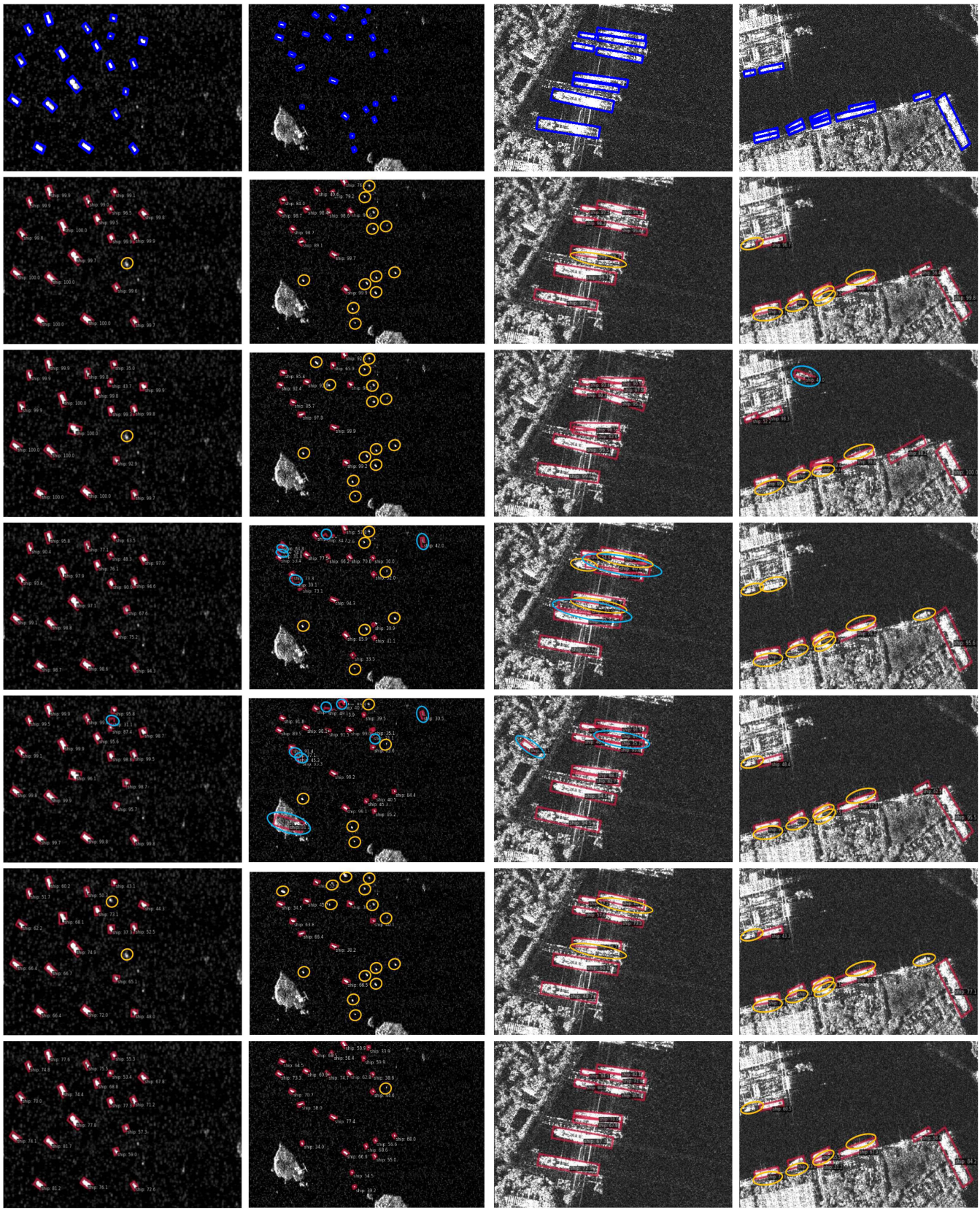


Fig. 12. Visualization of the detection results of various methods on SSDD. (a) Offshore Scene 1. (b) Offshore Scene 2. (c) Inshore Scene 1. (d) Inshore Scene 2. From top to bottom: ground truth, oriented R-CNN, faster R-CNN(OBB), R3Det, S2A-Net, FCOS (OBB), SAD-Det. The blue boxes indicate the ground truth, the red boxes indicate true positives, the orange circles indicate false positives, and the light blue circles indicate false negatives.

TABLE VII
DETECTION RESULTS OF DIFFERENT ARBITRARILY ORIENTED OBJECT DETECTION METHODS ON RSDD-SAR

Methods	AP	AP50	AP75	FLOPs (G)	Params (M)
Two-stage, anchor-based					
Oriented R-CNN	48.0%	89.6%	46.5%	63.28	41.13
Faster R-CNN(OBB)	40.6%	85.7%	32.2%	63.25	41.12
ReDet	50.4%	90.7%	51.7%	40.88	31.54
One-stage, anchor-based					
R3Det	45.4%	89.5%	39.9%	82.17	41.58
S2A-Net	48.6%	90.9%	47.5%	49.05	36.18
One-stage, anchor-free					
FCOS(OBB)	47.9%	89.3%	47.1%	51.55	31.89
Based on a transformer					
SAD-Det	**52.0%**	**93.8%**	**53.9%**	182.82	45.39

The bold item indicates the best value for each column.

TABLE VIII
DETECTION RESULTS OF DIFFERENT ARBITRARILY ORIENTED SHIP DETECTION METHODS ON RSDD-SAR

Methods	All–AP50	Inshore–AP50	Offshore–AP50
KeyShip	89.8%	–	–
SaDet*	91.7%	74.3%	95.5%
AEDet	90.1%	77.8%	90.5%
MT-FANet	90.8%	66.9%	95.7%
SAD-Det	**93.8%**	**78.7%**	**96.5%**

[1] * indicates that the structure is reproduced.
[2] The bold items indicate the best value for each column.

4.5%, respectively, highlighting the advantages of the proposed method in detecting OBBs. The corresponding PRCs are shown in Fig. 11, confirming the superior performance of the proposed method compared with other approaches.

We show the results of several methods on RSDD-SAR in Fig. 13. Fig. 13(a) displays some ship targets under high-sea states. While the other methods exhibit FPs and FNs, SAD-Det accurately detects all objects. The other three images in Fig. 13(b)–(d) depict ship targets in inshore scenes. Due to the interference of artificial facilities on the shore, small islands in the sea and high-sea states, most methods generate missed and false detections, and some FPs and FNs appear. In contrast, SAD-Det accurately detects all objects in these three images. This indicates that the proposed method possesses superior robustness and anti-interference ability compared with the other methods.

E. Comparison With Arbitrarily Oriented Ship Detection Methods

To further validate the effectiveness of the proposed method for ship detection, this section compares the proposed SAD-Det with several arbitrarily oriented ship detection methods on RSDD-SAR from both inshore and offshore perspectives. These methods include KeyShip [31], SaDet [27], AEDet [29], and MT-FANet [68]. Since the aforementioned methods are not open source, we reproduce the structure of SaDet and conduct experiments, citing the best experimental results from the other studies as our comparison values. The specific results are presented in Table VIII, which demonstrates that the AP50 of SAD-Det achieves the best value in all scenes, inshore scenes, and offshore scenes. This clearly illustrates the effectiveness of the proposed method in the field of ship detection.

TABLE IX
SPECIFIC CONFIGURATION OF A LARGE-SCALE ALOS-2 SAR IMAGE

Parameter	Value
Position	Tokyo Bay
Polarization	HH
Waveband	L
Resolution (m)	3
Image Size (pixels)	10389×6487
Time	2014-04-10

F. Verification on a Large-Scene SAR Image

To assess the performance of the proposed method on a large-scene SAR image, we conduct experiments using a large-scale ALOS-2 SAR image that includes both inshore and offshore scenes. This image is not part of the SSDD and RSDD-SAR datasets, and its specific details can be found in Table IX. The model weights trained on SSDD are applied to this image, and the results can be seen in Fig. 14. The performance is satisfactory for the seaside region, with only one missed detection observed in the case of multiple small targets. In contrast, the inshore scenes exhibit a higher number of false and missed detections. This discrepancy can be attributed to the presence of interference, such as artificial facilities in inshore scenes, as well as the absence of large land scenes on SSDD. This suggests the necessity of further improving the anti-interference capability of the proposed method in inshore scenes and its robustness when applied to other datasets.

IV. DISCUSSION

For deep learning or machine learning models, we not only require them to fit well with the training dataset, which means having a small training error, but also hope that they can fit well with unknown datasets, such as the testing dataset, which means having strong generalization ability. The effectiveness of generalization ability can be evaluated by examining the occurrence of model overfitting and underfitting, as illustrated in Fig. 15. As the number of epochs increases, the training loss gradually decreases, and the validation loss initially decreases and then increases. Overfitting and underfitting are two conditions used to describe the model's behavior during the training process. Both conditions adversely impact the neural network's generalization ability, with overfitting being the primary concern in current research.

In the field of ship detection from SAR images, overfitting is a concern due to the limited number of samples in the ship dataset. To address this issue, several methods are employed in this study to prevent overfitting. Specifically, data augmentation is applied to expand the size of the dataset. This involves preprocessing the images by applying techniques, such as image rotation, scaling, and random cropping to generate additional images. In addition, L2 regularization and dropout regularization are employed to reduce model complexity.

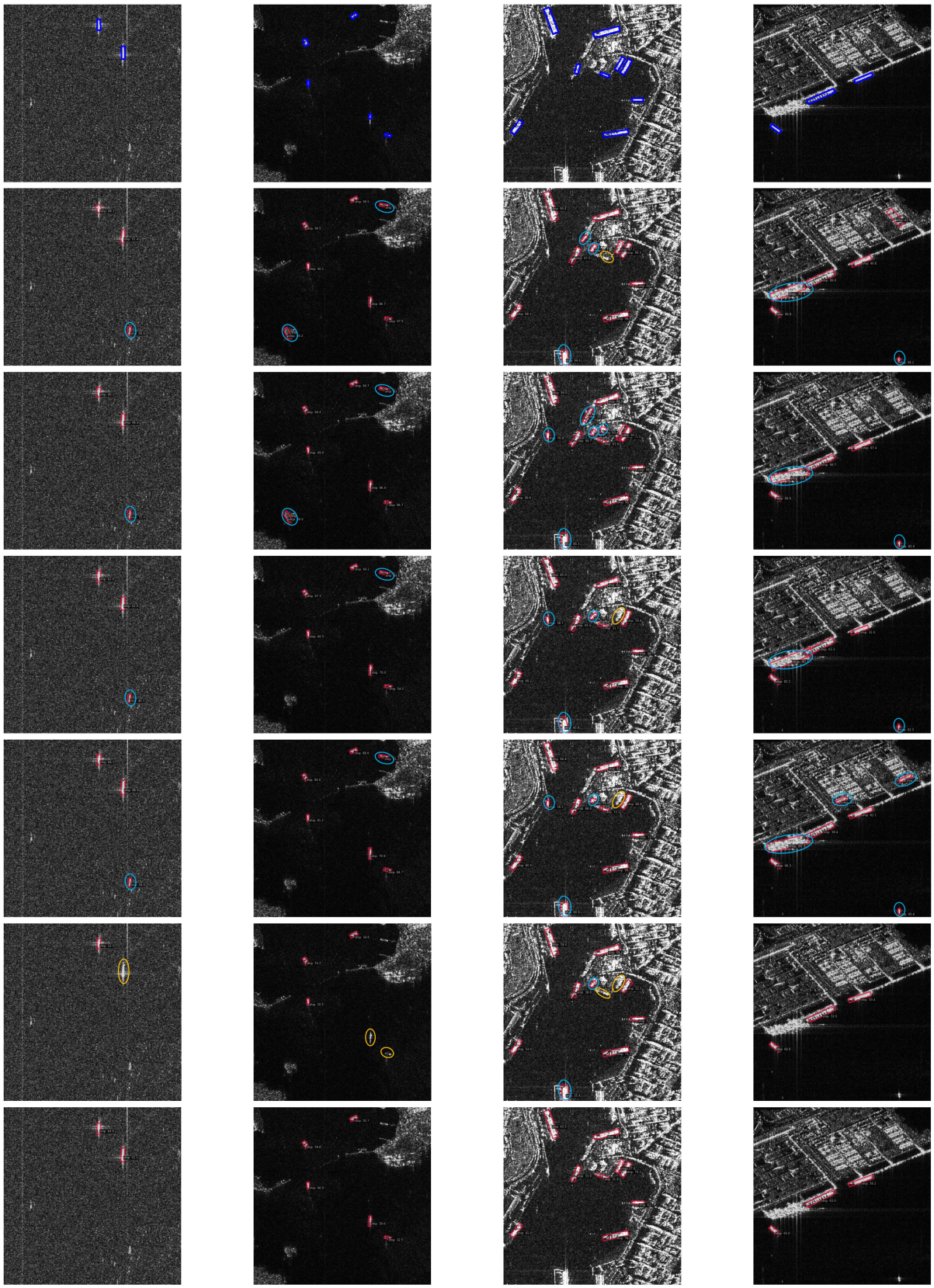


Fig. 13. Visualization of the detection results of various methods on RSDD-SAR. (a) Offshore Scene 1. (b) Inshore Scene 1. (c) Inshore Scene 2. (d) Inshore Scene 3. From top to bottom: ground truth, oriented R-CNN, faster R-CNN (OBB), R3Det, S2A-Net, FCOS (OBB), SAD-Det. The blue boxes indicate the ground truth, the red boxes indicate true positives, the orange circles indicate false positives, and the light blue circles indicate false negatives.

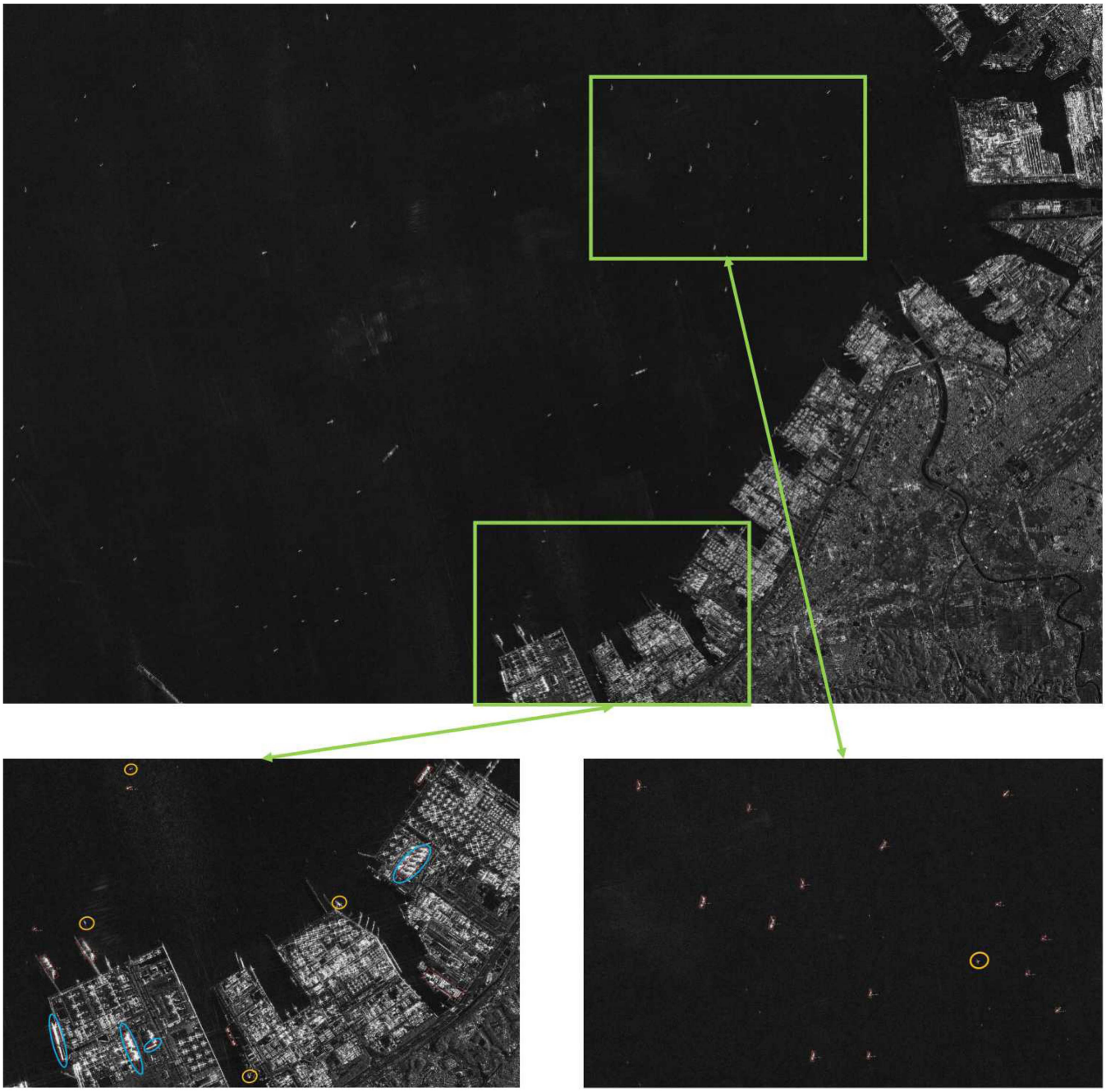


Fig. 14. Visualization of the detection results for a large-scene ALOS-2 SAR image. The red boxes indicate true positives, the orange circles indicate false positives, and the light blue circles indicate false negatives.

The performance of the proposed model is evaluated as follows. First, increasing the number of epochs leads to a rise in validation loss and a subsequent decrease in the AP value if the model is overfitting. Fig. 16 illustrates the relationship between AP value and epochs for the proposed method in this study. Fig. 16(a) presents the test results on SSDD, while Fig. 16(b) shows the test results on RSDD-SAR. It can be observed that as the number of epochs increases, the values of AP, AP50, and AP75 initially increase and then stabilize without decreasing. This indicates that the proposed method is not affected by overfitting. Furthermore, a hallmark of overfitting is limited generalization ability. Since the proposed model yields impressive outcomes across two datasets and a large-scene SAR image, it can be inferred to a certain extent that the model avoids overfitting.

Judging from the experimental results, it can be observed that the trained model did not exhibit signs of overfitting on the two datasets, thereby confirming the effectiveness of the proposed method. Furthermore, it is important to explore lightweight and high-speed techniques for ship detection in SAR images, as this will be the primary focus of my future research.

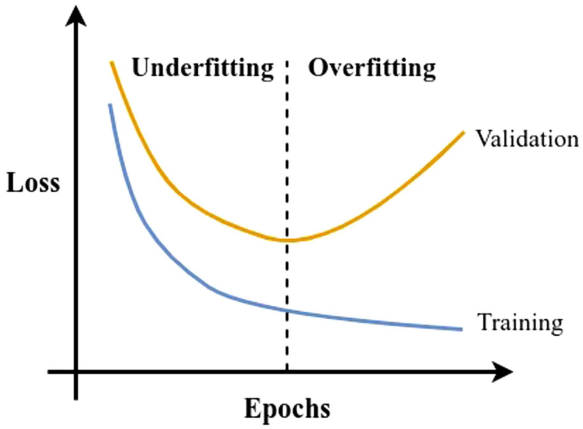


Fig. 15. Schematic diagram of overfitting and underfitting.

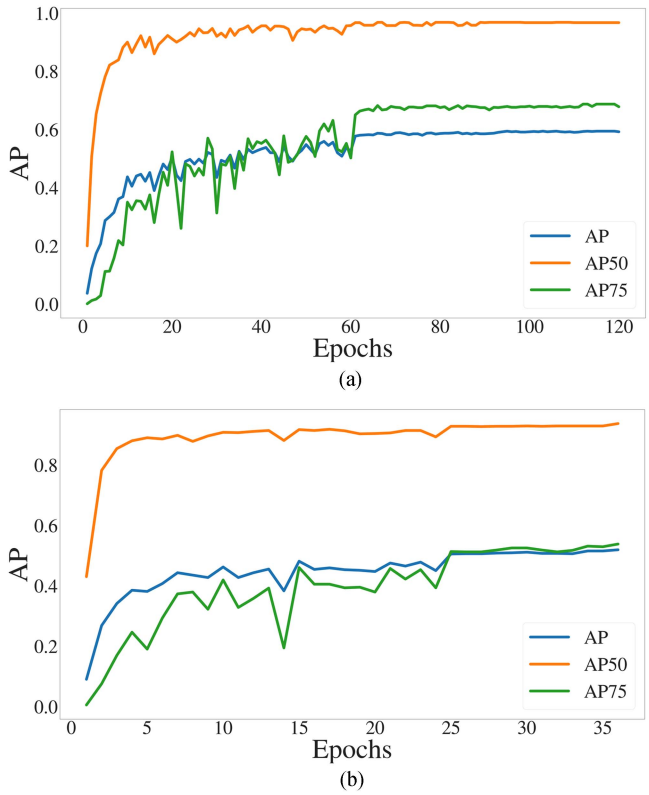


Fig. 16. Relationship between AP and epochs. (a) Results on SSDD. (b) Results on RSDD-SAR.

V. Conclusion

In this article, we propose an anchor-free method based on transformers and adaptive features for arbitrarily oriented ship detection in SAR images, namely, SAD-Det. SAD-Det is a hybrid structure of a transformer and CNN that can detect rotationally invariant ship targets with high average precision in SAR images. SAD-Det comprises three main components: SSP-PVT, AFPN, and DeHead. SSP-PVT is a backbone network that is based on PVT and combined with SSPM. SSP-PVT can enhance the long-range dependencies of ships in SAR images and obtain sufficient context information to improve the detection performance of ships in SAR images. AFPN is a neck network that is based on FPN and augmented with AWM. The addition of fusion factors allows different weights to be assigned to feature layers, thereby enhancing the performance of feature fusion in ships in SAR images. DeHead is a head network that utilizes DC to adaptively detect the spatial sampling positions of the targets and combines residual connections to make the network more adaptable to the characteristics of ships in SAR images. The effectiveness of each module in the proposed method is verified by experiments. Compared with other arbitrarily oriented object detection methods, this method achieves state-of-the-art detection performance. In addition, the inclusion of specific modules and targeted network designs significantly increases the computational complexity of the proposed method. Future research will explore high-speed ship detection methods for SAR images.

References

[1] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou, "A tutorial on synthetic aperture radar," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 1, pp. 6–43, Mar. 2013, doi: 10.1109/MGRS.2013.2248301.

[2] R. Gens, "Oceanographic applications of SAR remote sensing," *GIScience Remote Sens.*, vol. 45, no. 3, pp. 275–305, 2008, doi: 10.2747/1548-1603.45.3.275.

[3] R. Bürgmann, P. A. Rosen, and E. J. Fielding, "Synthetic aperture radar interferometry to measure earth's surface topography and its deformation," *Annu. Rev. Earth Planet. Sci.*, vol. 28, no. 1, pp. 169–209, 2000, doi: 10.1146/annurev.earth.28.1.169.

[4] C. Liu, Z. Chen, Y. Shao, J. Chen, T. Hasi, and H. Pan, "Research advances of SAR remote sensing for agriculture applications: A review," *J. Integrative Agriculture*, vol. 18, no. 3, pp. 506–525, 2019, doi: 10.1016/S2095-3119(18)62016-7.

[5] A. C. Mondini, F. Guzzetti, K.-T. Chang, O. Monserrat, T. R. Martha, and A. Manconi, "Landslide failures detection and mapping using synthetic aperture radar: Past, present and future," *Earth- Sci. Rev.*, vol. 216, 2021, Art. no. 103574, doi: 10.1016/j.earscirev.2021.103574.

[6] J. Li, C. Xu, H. Su, L. Gao, and T. Wang, "Deep learning for SAR ship detection: Past, present and future," *Remote Sens.*, vol. 14, no. 11, 2022, Art. no. 2712, doi: 10.3390/rs14112712.

[7] D. J. Crisp, "The state-of-the-art in ship detection in synthetic aperture radar imagery," Defence Science and Technology Organisation Salisbury (Australia) Info Sciences Lab, Tech. Rep. ADA426096, 2004.

[8] G. Gao, "Statistical modeling of SAR images: A survey," *Sensors*, vol. 10, no. 1, pp. 775–795, 2010, doi: 10.3390/s100100775.

[9] A. Renga, M. D. Graziano, and A. Moccia, "Segmentation of marine SAR images by sublook analysis and application to sea traffic monitoring," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1463–1477, Mar. 2019, doi: 10.1109/TGRS.2018.2866934.

[10] L. Li, L. Du, and Z. Wang, "Target detection based on dual-domain sparse reconstruction saliency in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4230–4243, Nov. 2018, doi: 10.1109/JSTARS.2018.2874128.

[11] 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, doi: 10.1109/LGRS.2016.2631638.

[12] T. Liu, Z. Yang, Y. Jiang, and G. Gao, "Review of ship detection in polarimetric synthetic aperture imagery," *J. Radars*, vol. 10, no. 1, pp. 1–19, 2021, doi: 10.12000/JR20155.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[14] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. SAR Big Data Era: Models, Methods Appl.*, 2017, pp. 1–6, doi: 10.1109/BIGSARDATA.2017.8124934.

[15] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," Nov. 2015, *arXiv:1511.08458*.

[16] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[17] 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, doi: 10.1109/LGRS.2018.2882551.

[18] M. Li, S. Lin, and X. Huang, "SAR ship detection based on enhanced attention mechanism," in *Proc. 2nd Int. Conf. Artif. Intell. Comput. Eng.*, 2021, pp. 759–762, doi: 10.1109/ICAICE54393.2021.00148.

[19] X. Nie, M. Duan, H. Ding, B. Hu, and E. K. Wong, "Attention mask R-CNN for ship detection and segmentation from remote sensing images," *IEEE Access*, vol. 8, pp. 9325–9334, 2020, doi: 10.1109/ACCESS.2020.2964540.

[20] L. Bai, C. Yao, Z. Ye, D. Xue, X. Lin, and M. Hui, "Feature enhancement pyramid and shallow feature reconstruction network for SAR ship detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1042–1056, Jan. 2023, doi: 10.1109/JSTARS.2022.3230859.

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

[22] L. Zhang, Y. Liu, W. Zhao, X. Wang, G. Li, and Y. He, "Frequency-adaptive learning for SAR ship detection in clutter scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art no. 5215514, doi: 10.1109/TGRS.2023.3249349.

[23] G. Jocher, "YOLOv5 by ultralytics," May 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[24] G. Zhang, Z. Li, X. Li, C. Yin, and Z. Shi, "A novel salient feature fusion method for ship detection in synthetic aperture radar images," *IEEE Access*, vol. 8, pp. 215904–215914, 2020, doi: 10.1109/ACCESS.2020.3041372.

[25] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[26] K. Wang, Z. Li, A. Su, and Z. Wang, "Oriented object detection in optical remote sensing images using deep learning: A survey," Feb. 2023, *arXiv:2302.10473*.

[27] S. Zhao, Q. Liu, W. Yu, and J. Lv, "A single-stage arbitrary-oriented detector based on multiscale feature fusion and calibration for SAR ship detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8179–8198, Sep. 2022, doi: 10.1109/JSTARS.2022.3206822.

[28] P. Guo, T. Celik, N. Liu, and H. Li, "Break through the border restriction of horizontal bounding box for arbitrary-oriented ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Apr. 2023, Art no. 4005505, doi: 10.1109/LGRS.2023.3270897.

[29] K. Zhou et al., "Arbitrary-oriented ellipse detector for ship detection in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 7151–7162, Apr. 2023, doi: 10.1109/JSTARS.2023.3267240.

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

[31] J. Ge, Y. Tang, K. Guo, Y. Zheng, H. Hu, and J. Liang, "KeyShip: Towards high-precision oriented SAR ship detection using key points," *Remote Sens.*, vol. 15, no. 8, 2023, Art. no. 2035, doi: 10.3390/rs15082035.

[32] A. A. Aleissaee et al., "Transformers in remote sensing: A survey," *Remote Sens.*, vol. 15, no. 7, 2023, Art. no. 1860, doi: 10.3390/rs15071860.

[33] E. Rodner, M. Simon, R. B. Fisher, and J. Denzler, "Fine-grained recognition in the noisy wild: Sensitivity analysis of convolutional neural networks approaches," Oct. 2016, *arXiv:1610.06756*.

[34] N. Park and S. Kim, "How do vision transformers work?," Feb. 2022, *arXiv:2202.06709*.

[35] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M. H. Yang, "Intriguing properties of vision transformers," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 23296–23308.

[36] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jun. 2023, Art. no. 5514415, doi: 10.1109/TGRS.2023.3284671.

[37] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art no. 5513412, doi: 10.1109/TGRS.2023.3279834.

[38] R. Xia et al., "CRTransSar: A visual transformer based on contextual joint representation learning for SAR ship detection," *Remote Sens.*, vol. 14, no. 6, 2022, Art. no. 1488, doi: 10.3390/rs14061488.

[39] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[40] K. Li, M. Zhang, M. Xu, R. Tang, L. Wang, and H. Wang, "Ship detection in SAR images based on feature enhancement swin transformer and adjacent feature fusion," *Remote Sens.*, vol. 14, no. 13, 2022, Art. no. 3186, doi: 10.3390/rs14133186.

[41] H. Shi, B. Chai, Y. Wang, and L. Chen, "A local-sparse-information-aggregation transformer with explicit contour guidance for SAR ship detection," *Remote Sens.*, vol. 14, no. 20, 2022, Art. no. 5247, doi: 10.3390/rs14205247.

[42] Y. Zhou, F. Zhang, Q. Yin, F. Ma, and F. Zhang, "Inshore dense ship detection in SAR images based on edge semantic decoupling and transformer," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4882–4890, May 2023, doi: 10.1109/JSTARS.2023.3277013.

[43] Y. Zhou, X. Jiang, G. Xu, X. Yang, X. Liu, and Z. Li, "PVT-SAR: An arbitrarily oriented SAR ship detector with pyramid vision transformer," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 291–305, Nov. 2023, doi: 10.1109/JSTARS.2022.3221784.

[44] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.

[45] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," Oct. 2020, *arXiv:2010.11929*.

[46] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022, doi: 10.1007/s41095-022-0274-8.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[48] Q. Hou, L. Zhang, M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4003–4012.

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

[50] Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao, and Z. Han, "Effective fusion factor in FPN for tiny object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1160–1168.

[51] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.

[52] 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.*, 2019, pp. 9308–9316.

[53] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2022, doi: 10.1109/TPAMI.2020.3032166.

[54] D. Zhou et al., "IoU loss for 2D/3D object detection," in *Proc. Int. Conf. 3D Vis.*, 2019, pp. 85–94, doi: 10.1109/3DV.2019.00019.

[55] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," Jun. 2014, *arXiv:1406.2080*.

[56] T. Zhang et al., "SAR ship detection dataset (SSDD): Official release and comprehensive data analysis," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3690, doi: 10.3390/rs13183690.

[57] C. Xu et al., "RSDD-SAR: Rotated ship detection dataset in SAR images," *J. Radars*, vol. 11, no. 4, pp. 581–599, 2022, doi: 10.12000/JR22007.

[58] Y. Zhou et al., "MMRotate: A rotated object detection benchmark using pytorch," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 7331–7334, doi: 10.1145/3503161.3548541.

[59] T. Y. Lin et al., "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[60] M. Everingham and J. Winn, "The Pascal visual object classes challenge 2012 (voc2012) development kit," *Pattern Anal. Statist. Model. Comput. Learn.*, vol. 2007, no. 1–45, pp. 12–14 2012.

[61] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[62] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10781–10790.

[63] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3520–3529.

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

[65] J. Han, J. Ding, N. Xue, and G. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2786–2795.

[66] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, pp. 3163–3171, 2021, doi: 10.1609/aaai.v35i4.16426.
[67] J. Han, J. Ding, J. Li, and G. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art no. 5602511, doi: 10.1109/TGRS.2021.3062048.
[68] Q. Liu, D. Li, R. Jiang, S. Liu, H. Liu, and S. Li, "MT-FANet: A morphology and topology-based feature alignment network for SAR ship rotation detection," *Remote Sens.*, vol. 15, no. 12, 2023, Art. no. 3001, doi: 10.3390/rs15123001.

Bingji Chen received the B.E. degree in optoelectronic information science and engineering from Jilin University, Changchun, China, in 2017, and the M.E. degree in electronics and communication engineering from Peking University, Beijing, China, in 2020. He is currently working toward the Ph.D. degree in communication and information system with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, and the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing.

His research interests include computer vision and synthetic aperture radar images object detection and recognition.

Chunrui Yu received the Ph.D. degree in signal processing from the National University of Defense Technology, Changsha, China, in 2012.

He is currently a Research Associate with the Beijing Institute of Tracking and Telecommunication Technology, Beijing, China. His research interests include synthetic aperture radar (SAR) system design, SAR jamming and antijamming, and space-time adaptive processing.

Shuang Zhao received the B.S. degree in communication engineering from the Harbin Engineering University, Harbin, China, in 2018, and the Ph.D. degree in communication and information system with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China and the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, in 2023.

Her research interests include computer vision and synthetic aperture radar (SAR) image processing, especially on SAR object detection and recognition.

Hongjun Song received the B.E. degree in electronics engineering from the University of Science and Technology of China, Hefei, China, in 1991 and the M.E. and the Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 1994 and 1998, respectively.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include the signal processing and system design of novel synthetic aperture radar modes.