PSFNet: Efficient Detection of SAR Image Based on Petty-Specialized Feature Aggregation

Peng Zhou, Peng Wang, *Senior Member, IEEE*, Jie Cao, Daiyin Zhu, *Member, IEEE*, Qiyuan Yin, Jiming Lv, Ping Chen, Yongshi Jie, and Cheng Jiang

***Abstract*—With the rapid development of deep learning, convolutional neural networks have achieved milestones in synthetic aperture radar (SAR) image object detection. However, object detection in SAR images is still a great challenge due to the difficulty in distinguishing targets from complex backgrounds. At the same time, most of the targets in SAR images are small and unevenly distributed, which makes it challenging to extract sufficient feature information. To solve these issues mentioned above, an efficient object detection network for SAR images based on Swin transformer and YOLOv7 is proposed in this article. First, we design a novel feature aggregation module Petty-specialized feature aggregation (PS-FPN) to enrich small targets' semantic and spatial features while keeping the model lightweight. PS-FPN module uses the fusion of deep and shallow features by using cross-layer feature aggregation and single-branch feature aggregation to enhance the detection of small targets. Second, a novel attention mechanism strategy mix-attention is proposed to find more attention regions. Finally, we add one more prediction head to extract shallow features that effectively preserve small targets' feature information. To verify the effectiveness of the proposed algorithm, extensive experiments are carried out on several challenging SAR image datasets. The results show that, compared with other state-of-the-art detectors, the proposed method can achieve significant performance based on lightweight detection.**

***Index Terms*—Feature aggregation, mix-attention, object detection, synthetic aperture radar (SAR), swin transformer, YOLOv7.**

Manuscript received 22 August 2023; revised 28 September 2023; accepted 20 October 2023. Date of publication 25 October 2023; date of current version 23 November 2023. This work was supported in part by Fundamental Research Funds for the Central Universities in Nanjing University of Aeronautics and Astronautics under Grant NS2023020 and Grant NJ2023029, in part by the Postgraduate Research and Practice Innovation Program of NUAA under Grant xcxjh20220414, in part by Open Project Funds for the Key Laboratory of Space Photoelectric Detection and Perception (Nanjing University of Aeronautics and Astronautics), Ministry of Industry and Information Technology, under Grant NJ2023029-1, in part by Key Laboratory of Radar Imaging and Microwave Photonics, Nanjing University of Aeronautics and Astronautics, Ministry of Education, under Grant NJ20230005, in part by Beijing Key Laboratory of Urban Spatial Information Engineering under Grant 20230114, in part by Beijing Key Laboratory of Advanced Optical Remote Sensing Technology under Grant AORS202311, in part by Shanxi Key Laboratory of Signal Capturing and Processing under Grant 2023-002, in part by the State Key Laboratory of Geoinformation Engineering and Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, CASM, under Grant 2023-03-09, in part by the Key Laboratory of System Control and Information Processing, Ministry of Education, under Grant Scip20230104, in part by Science Foundation of Donghai Laboratory under Grant DH-2022KF01011, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20221478, in part by the Hong Kong Scholars Program under Grant XJ2022043, in part by the Youth Promotion Talent Project of Jiangsu Association for Science and Technology under Grant TJ-2023-010, and in part by the National Natural Science Foundation of China under Grant 61801211. *(Corresponding author: Peng Wang; Jie Cao.)*

Peng Zhou is with the Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China, also with the Beijing Key Laboratory of Urban Spatial Information Engineering, Beijing Institute of Surveying and Mapping, Beijing 100038, China, and also with the Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, Chinese Academy of Surveying and Mapping, Beijing 100039, China (e-mail: zhou_peng@nuaa.edu.cn).

Peng Wang is with the Key Laboratory of Space Photoelectric Detection and Perception (Nanjing University of Aeronautics and Astronautics), Ministry of Industry and Information Technology, Nanjing 210016, China, also with the Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai 200030, China, also with the Donghai Laboratory, Zhoushan 316021, China, and also with the Beijing Key Laboratory of Advanced Optical Remote Sensing Technology, Beijing Institute of Space Mechanics and Electricity, Beijing 100094, China (e-mail: pengwang-b614080003@hotmail.com).

Jie Cao is with the Key Laboratory of Small- and Medium-Sized UAV Advanced Technology, Ministry of Industry and Information Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210000, China (e-mail: caoj@nuaa.edu.cn).

Daiyin Zhu and Jiming Lv are with the Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: zhudy@nuaa.edu.cn; jmlv_nj@nuaa.edu.cn).

Qiyuan Yin is with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: yscyin@nuaa.edu.cn).

Ping Chen is with the Shanxi Key Laboratory of Signal Capturing and Processing, North University of China, Taiyuan 030051, China (e-mail: chenping@nuc.edu.cn).

Yongshi Jie and Cheng Jiang are with the Beijing Key Laboratory of Advanced Optical Remote Sensing Technology, Beijing Institute of Space Mechanics and Electricity, Beijing 100094, China (e-mail: jie_yongshi@163.com; cheng3515523@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3327344

I. Introduction

SYNTHETIC aperture radar (SAR) is an indispensable and vital monitoring tool in the field of remote sensing. The unique advantage of SAR is that it is not limited by time or any adverse weather conditions. SAR has the capability of performing multiview imaging and exhibiting a specific perspective ability, which can be widely applied in various remote sensing fields. Object detection is an essential application of SAR imagery, which plays a significant role in military and civil fields, such as maritime surveillance, forestry construction, and agricultural monitoring. This article focuses on improving the accuracy of SAR image object detection while maintaining lightweight performance, which provides help for the above practical applications.

At present, there is an increasing number of SAR signal processing and image processing algorithms, which have provided access to high-quality and large-scale SAR imagery data [1], [2]. Li et al. [3] proposed an optimal time selection algorithm based on multiscatterer time–frequency analysis, which improves the azimuth focusing performance of SAR images and better characterizes the spatial variation of SAR targets' Doppler frequencies. Li et al. [4] also introduced a novel time-domain notched filtering algorithm that demonstrates capabilities in pulse interference detection and suppression of radio frequency interference. Chen et al. [5] employed a similar pixel number

indicator that effectively reflects the differences between pixels and filters out speckle noise. These advanced SAR signal processing algorithms have generated higher quality images, enabling improved detection performance in SAR image object detection.

In addition, both high-quality SAR images and excellent detection algorithms are essential. Suitable detection algorithms can significantly enhance SAR object detection performance. Traditional SAR object detection algorithms are mainly divided into three procedures: region selection, feature extraction, and classification [6]. The principal purpose of region selection is to regress the targets. The feature extraction based on region selection accurately expresses the target region, which is the core task of object detection. Traditional feature extraction methods include scale-invariant feature transform [7] and histogram of oriented gradient [8]. However, the features extracted by traditional detection methods are ineffective for object detection in SAR images. Speckle noise and motion blur in SAR images may cause unnecessary differences between targets, which makes it difficult for traditional SAR object detection [9]. In the theoretical research of object detection in SAR images, the constant false alarm rate (CFAR) detection algorithm is the most widely used and effective detection method. However, the CFAR algorithm is greatly affected by the background statistical area. Generally speaking, increasing the statistical area of the background increases the accuracy of the target description. However, this approach may lead to increased background clutter and a higher false alarm rate [10]. Therefore, it is necessary to address these issues when using traditional methods for SAR image object detection.

In recent years, thanks to the development of hardware and computing power, convolutional neural network (CNN) has made significant progress in SAR image object detection. Compared with the traditional methods, the method based on deep learning promotes the development of object detection due to its excellent feature expression and learning ability. At present, detection methods based on deep learning are mainly divided into two types, namely, two-stage and single-stage detectors.

The two-stage detector divides the detection process into two parts. First, multiple candidate regions that may contain objects are generated. Then, the detector performs classification and regression tasks on each candidate region to obtain the detection results. Typical two-stage object detection methods include R-CNN series [11], [12], [13]. Although the two-stage detector performs well in accuracy, processing a large number of candidate regions will incur additional computational overhead. Unlike the two-stage algorithm, the single-stage detector directly generates the category and location information of the target, sacrificing detection accuracy for computational efficiency. Typical single-stage detectors include YOLO series [14], [15], [16], SSD series [17], etc. However, due to the intrinsic locality of convolution operation, it is still difficult for these methods to extract the global information in space-time [18]. Therefore, some methods have been proposed to solve this problem. Chen et al. [19] used ResNet [20] as the backbone network, Zhang et al. [21] used the dilated convolution to replace the traditional convolution, and Fang et al. [22] and Peng et al. [23] used attention mechanism to improve the global feature extraction ability of the model. Although these methods increase the receptive field of the network, the ability to extract global information still fails to achieve satisfactory results in object detection.

Vaswani et al. [24] proposed the transformer model in 2017, which is a significant breakthrough in artificial intelligence. In recent years, many scholars have introduced transformer into the field of computer vision (CV) [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36]. Due to the amazing semantic representation capability of the transformer, it performs as well or even better than CNN in various CV tasks, such as image classification [25], [26], object detection [29], [30], [31], semantic segmentation [27], [29], [32], image generation [33], [34], and super-resolution [35], [36], [37]. Unlike CNN networks, which require attention mechanism modules to focus on attention features, the vision transformer (ViT) [25] network can directly extract global attention. In fact, ViT is the first model to apply the global attention mechanism to image applications, making it a promising approach for a wide range of CV tasks. ViT network takes 2-D image patches with location information and class tokens as input, achieving comparable performance to CNN-based methods on large-scale image classification tasks. Although ViT has achieved dazzling breakthroughs in image feature extraction, its computational overhead makes it challenging to deploy in practical applications. Therefore, Liu et al. [28] proposed the Swin transformer network, which uses the shifted window scheme to calculate the self-attention in the local window, effectively reducing the computational complexity of self-attention and acquiring the best results in several CV tasks. Furthermore, DeiT [26] leverages the transformer network for efficient image classification and object detection while enhancing the model's robustness and generalization capabilities through the utilization of knowledge distillation and image embedding techniques. Given the remarkable performance of the transformer and its variants in CV tasks, particularly in image classification, the development potential of these models in object detection looms large.

However, due to the significant differences in distance between SAR and targets, targets in SAR images tend to be small in size and densely packed. Traditional object detection methods struggle to meet the requirements of practical applications when detecting densely packed small targets in SAR images, in terms of detection accuracy and lightweight performance. In addition, the presence of speckle noise makes the SAR image background more complex. Therefore, achieving high-precision and lightweight object detection in SAR images has emerged as a formidable challenge. To address the above problems, this article proposes an efficient detection method PSFNet for SAR images based on Petty-specialized feature aggregation (PS-FPN). First, the PS-FPN is designed to focus on cross-layer information in multiscale features and the enhancement of shallow feature information. Second, a novel attention mechanism mix-attention is proposed to suppress the interference of SAR image background in extracting target features, which is more conducive to discovering more target regions. Moreover, a new prediction head with a smaller receptive field is added at 4×4 times downsample of

the network, which will retain more detailed feature information for small object detection. Our main contributions are as follows.

1) In order to enhance the feature extraction of small targets in complex SAR scenes while maintaining the model's lightweight performance, we have designed a novel feature fusion module PS-FPN. The PS-FPN performs cross-layer feature extraction through multiple parallel feature aggregation paths, extracting deep abstract semantic information to the shallow layer for fusion. It also uses single-branch feature aggregation to strengthen the extraction and fusion of shallow feature information. In addition, PS-FPN introduces depthwise separable convolution (DSC) to replace a portion of the convolution in PSFNet. This effectively reduces the model parameters and computational complexity, resulting in improved lightweight performance for PSFNet.
2) To further suppress the interference of complex backgrounds in SAR images, this article proposes a novel attention mechanism strategy mix-attention. Specifically designed for the differences in detected targets between high- and low-level detection layers, mix-attention allocates attention extraction schemes reasonably, effectively extracting small targets from complex backgrounds. At the same time, considering the small receptive field of the shallow detection layer, we add a new detection layer to the backbone network for more effective detection of small targets.
3) Furthermore, to validate the effectiveness of our proposed method, this article conducts a comprehensive experimental analysis of the algorithm. It compares the impact of different modules of PSFNet on detection and explores the advantages of the algorithm. In addition, this article compares the proposed algorithm with other classical SAR image detection networks, demonstrating its superiority and applicability.

The rest of this article is organized as follows. Section II introduces the related work of object detection. Section III introduces the proposed model in detail. Section IV describes the contrast and ablation experimental procedure in detail and displays the experimental results, respectively. Finally, Section V concludes the article.

II. Related Work

Object detection is a classic task of CV. Despite ongoing efforts by researchers to develop more robust object detection algorithms, the task of detecting objects in both general and specific contexts remains a formidable challenge. In the following, we first introduce the relevant methods of general small object detection, and then review the methods especially used for SAR images.

A. General Small Object Detection

As GPU computing power has improved, CNN in deep learning has increasingly replaced traditional machine learning for object detection. As a result, many CNN-based object detection algorithms with excellent detection performance have emerged in recent years, including Faster R-CNN, YOLOv3, SSD, etc.

However, these algorithms mentioned above cannot achieve high accuracy in the task of small target detection. Unlike regular objects, small objects usually have less feature map information, which makes general object detection algorithms unable to extract enough feature information. Many researchers have attempted to enhance the performance of algorithms due to the negative impact of small object characteristics. Image pyramid and feature pyramid methods are the commonly used techniques to achieve this goal, as they are particularly effective in improving the detection of small objects. Image pyramids have the capability to take images of different sizes, which can improve model's performance in detecting small objects. However, the use of multiscale training strategies in image pyramids can lead to extreme differences in scale. To overcome this issue, the SNIP method was proposed [38]. During gradient backpropagation, SNIP only provides monitoring signals for objects within a specific scale range, ignoring those too large or too small. This approach effectively prevents extreme scale changes.

Although image pyramid can improve the detection of small objects, its computational load is deemed unacceptable. Consequently, object detection algorithms usually rely on feature pyramids, which employ the high-resolution feature map to predict small objects and the low-resolution feature map to predict large objects. This approach facilitates the encoding of multiscale features akin to image pyramids, circumventing the issue of excessive computational burden. However, the limited representation ability of shallow feature maps will restrict the detection performance of small objects. To address this problem, the feature pyramid network (FPN) [39] was developed. FPN allows information transfer between feature maps of different scales, thereby significantly improving the detection performance of small objects. Based on this foundation, several variants of FPN have been created, including path aggregation network [40], bidirectional FPN [41], and neural architecture search FPN [42].

B. SAR Image Object Detection

At present, with the advancement of SAR imaging technology, researchers have been able to generate high-quality SAR datasets, such as MSAR-1.0 [43], SSDD [44], and SAR-Ship-Dataset [45]. This has made SAR object detection a current research hotspot. Most traditional SAR object detection methods focus on the classical CFAR algorithm, which is a method based on gray features. Among them, the two-parameter CFAR method is a classical local adaptive object detection method [46]. This method achieves object detection by traversing the SAR image through a preset sliding window and comparing the pixel gray level in the window with an adaptive threshold to distinguish targets from clutter. Ai et al. [47] proposed a CFAR detection method based on bilateral fine-tuning statistics, which uses a strategy based on the bilateral threshold to eliminate outliers and improve detection performance in ocean scenes automatically. Notwithstanding their efficacy, these methods are constrained by their reliance on high contrast between targets and clutter in

SAR images, which is necessary to conform to the statistical distribution of clutter. Moreover, their adaptability to complex background scenes in object detection is limited. Therefore, Huo et al. [48] used the maximum-stable extreme region algorithm to select candidate regions for detection and obtain the threshold to detect the target in SAR images. On the other hand, Shi et al. [49] utilized the directed gradient histogram for feature extraction to achieve ship-background separation. However, these methods are commonly constrained by their inability to adapt to background transformations in large datasets and their limited capacity for autonomous feature learning.

In recent years, the increasing improvement in hardware computing power has led to the wider use of deep learning techniques. Compared with traditional methods, CNN methods have several advantages. They can achieve high accuracy in object detection tasks without designing complex feature extraction methods. Chen et al. [50] proposed an effective detection network algorithm that can detect multiscale SAR ships in complex scenes with the use of an attention mechanism. MSRIHL-CNN proposed by Ai et al. [51] combines low-level texture edge features and high-level depth features. Cui et al. [6] introduced the spatial shuffle-group enhance module into CenterNet [52] to extract stronger semantic features while suppressing some noise. Sun et al. [53] proposed an anchor-free method for ship target detection in high-resolution SAR images. This method combines the fully convolutional one-stage object detection network with a category-position module to detect ship targets. Kang et al. [54] combined a high-resolution region proposal network with an object detection network featuring contextual features to enhance the detection performance of ships. By integrating deep semantic and shallow high-resolution features, they achieved improved ship detection performance. At the same time, some researchers have investigated the impact of ground truth proximity to the target on the performance of SAR object detection. Sun et al. [55] proposed a novel SAR rotated ship detector, named BiFA-YOLO, based on the YOLO framework. This detector incorporates bidirectional feature fusion and angular classification techniques to enhance the detection performance of arbitrary-oriented ships. Furthermore, research works on algorithm lightweighting are also necessary. Zhou et al. [56] designed a lightweight and scattering feature extraction backbone that is more suitable for SAR image data and a new multiscale feature fusion neck for the multiscale feature discrepancy. Ma et al. [57] introduced Light-YOLOv4, a new lightweight object detection network that achieves model lightweight design by applying L1 regularization to channel scaling factors and performing network channel pruning. Additionally, knowledge distillation is employed to enhance detection accuracy through model retraining.

CNN-based methods are not good at effectively extracting long-term global features, which limits the detection performance of object detection algorithms. Inspired by the success of transformer in vision tasks, more scholars have devoted themselves to the research of SAR object detection based on transformer. Zhou et al. [58] introduced a new aggregation module AggregationS to achieve high-precision detection of SAR images. This module was achieved by encoding support and query features into the same feature subspace. Zhou et al. [59] introduced overlapping block embeddings and hybrid transformer encoder modules, which overcome the impact of target density and insufficient data on detection accuracy. Then, the normalized Gauss–Wasserstein distance loss is used to suppress the influence of the scattered interference from the ship boundary. However, both CNN-based and transformer-based networks face the challenge of achieving the best performance balance between detection accuracy and model lightweight when detecting small targets in SAR images. Therefore, this article combines the Swin transformer and YOLOv7 to explore the algorithm that can better handle the detailed features of SAR images.

III. Method

A. Framework Overview

Swin transformer is one of the most popular CV models due to its global attention view, which both convolutions and attention mechanisms lack. Therefore, we choose it as the backbone of our network. In addition, considering that YOLOv7 [16] has better feature extraction ability and speed advantages in model inference, we decide to incorporate some modules of YOLOv7 for prediction. Specifically, we have employed the spatial pyramid pooling and convolutional spatial pyramid pooling and extended efficient layer aggregation network. Consequently, we propose an efficient object detection method for SAR images based on PS-FPN.

The overall framework of the proposed method is illustrated in Fig. 1, where the backbone consists of Swin transformer blocks. The neck is responsible for fusing multiscale features and enhancing the details of small targets via the PS-FPN and mix-attention. In this section, we present the overall process and highlight the key technical points of our proposed method. Subsequently, each branch of the proposed method is described in detail.

B. Swin Transformer

Due to the complex background interference in SAR images, extracting representative target features using common convolutional structures becomes challenging. Therefore, this article utilizes Swin transformer to extract global information and allocate attention, significantly enhancing the algorithm's ability to extract features from SAR images. Furthermore, when compared to other transformer-based models, window based multihead self-attention (W-MSA) and shifted window based multihead self-attention (SW-MSA) modules within the Swin transformer demonstrate a remarkable ability to reduce computational complexity while maintaining high-performance feature extraction. This characteristic is crucial for ensuring that our model remains lightweight. The following is a detailed description of W-MSA and SW-MSA modules.

1) W-MSA Module: The ViT [25] and DeiT [26] use the standard MSA module, which calculates global self-attention on the entire feature map. However, their computational complexity is proportional to the input size squared, resulting in a notable

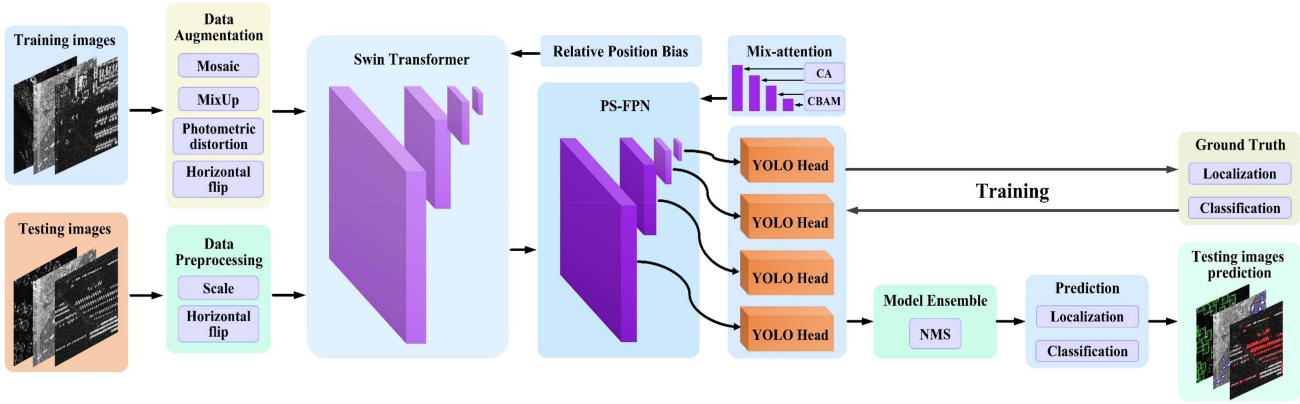


Fig. 1. Overall network framework of PSFNet. The backbone is composed of Swin transformer blocks. The neck realizes multiscale fusion and small target details enhancement through the PS-FPN module and mix-attention. The head predicts the object category and boundary. Here, the purple squares represent the feature layers, and the arrows indicate the direction of information flow.

computational burden. To address this issue, the Swin transformer [28] introduced the W-MSA module, which performs local self-attention calculation in each window. The W-MSA module partitions the input feature map into multiple nonoverlapping windows of size $M \times M$. As a result, the computational complexity of the W-MSA module is linearly proportional to the input size, reducing the computational overhead. The formulas for the computational complexity of the MSA and W-MSA modules are presented in (1) and (2), respectively

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{1}$$

$$\Omega(W-MSA) = 4hwC^2 + 2M^2hwC \tag{2}$$

where h and w represent the height and width of the feature map, C represents the number of channels, and M represents the size of the window.

2) SW-MSA Module: Due to the W-MSA module performing self-attention only in local windows, there is no connection between each window, which makes it challenging to capture the global features comprehensively. Therefore, Liu et al. [28] proposed the SW-MSA module to alleviate this drawback. The SW-MSA module window is offset by (M/2), (M/2) relative to the W-MSA module window. Within the Swin transformer, the W-MSA module alternates with the SW-MSA module to enlarge the network's receptive field.

C. Petty-Specialized Feature Aggregation

Compared to traditional optical images, SAR images tend to contain a larger number of small targets, which typically convey less information. Furthermore, due to the repeated downsample performed by the convolution and pooling layers, the pixel values corresponding to small targets in the feature map may gradually vanish, leading to more challenges in object detection. To address this issue, many existing object detection algorithms extract multiscale features via the backbone network. In this approach, the low-level feature maps possess higher resolution and more fine-grained features that enable accurate regression of detection coordinates. Therefore, utilizing low-level feature maps is more effective for predicting small objects. However, due to the limited semantic information in low-level feature maps, the classification ability of the model may be inadequate [60].

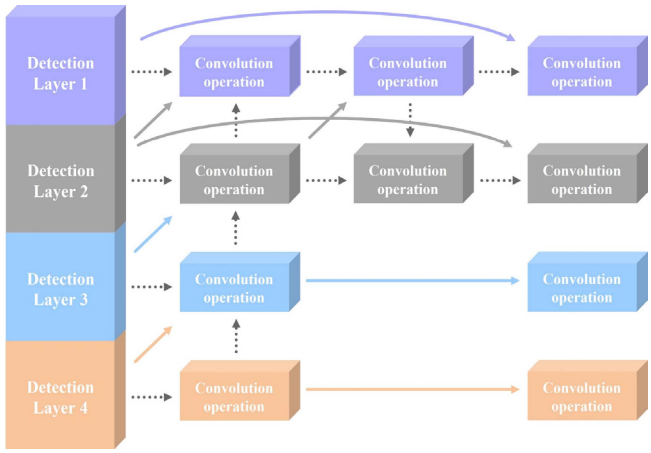


Fig. 2. Brief model structure of the PSFNet. PS-FPN structure in the neck feature fusion module of PSFNet. The dotted line represents the partial structure of the preserved PAN [40] module, whereas the solid line represents the innovation of this article in PS-FPN.

In addition, due to the lack of spatial information, the high-level feature map may not be able to predict the location of objects precisely. To address these issues mentioned above, many detection networks employ feature fusion techniques in the neck layer, such as FPN and PAN, to fuse feature maps at different scales and facilitate deep semantic information and shallow regression information. These fusion structures use top-down and bottom-up fusion paths to transfer and fuse semantic information from high-level feature maps and spatial location information from shallow feature maps, resulting in improved feature fusion in general object detection scenarios. However, in SAR images, the presence of speckle noise and a large number of small targets may hinder the effective extraction of feature information by general neck feature fusion networks.

Therefore, as shown in Fig. 2, we propose the PS-FPN module, which mainly consists of cross-layer feature aggregation and

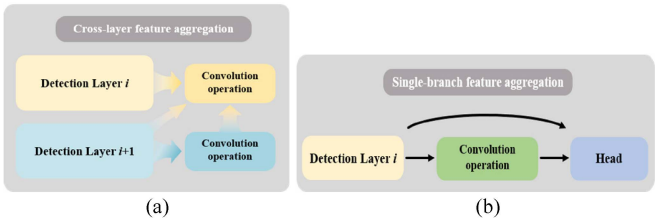


Fig. 3. Basic structure of the PS-FPN. (a) Cross-layer feature aggregation. (b) Single-branch feature aggregation.

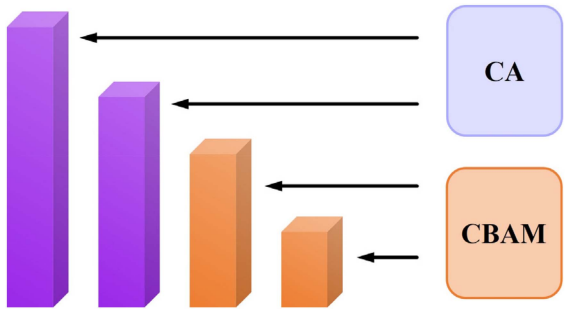


Fig. 4. Structure of the mix-attention.

single-branch feature aggregation. The structure of the cross-layer feature aggregation is shown in Fig. 3(a). This module retains the bottom-up transfer branch of the PAN module. In addition to the bottom-up feature information stitching from layer $i+1$ to layer i after feature extraction by convolution, we also add feature information stitching from layer $i+1$ to layer i before the convolution operation. These features extracted by convolution operation generally represent the original feature map. Still, they may lose some original details crucial for small targets with less feature information. Therefore, using the cross-layer feature aggregation module can better retain the detailed information of small targets and strengthen the feature extraction ability of small-sized and medium-sized targets. Similarly, the structure of the single-branch feature aggregation is shown in Fig. 3(b). In this structure, a residual is added to the tail of the neck module in the layer i. This can effectively retain the original feature details of each layer and prevent them from being covered by the stacked convolution layers and pooling layers. In addition, this approach helps to enrich the feature information of the detection branch and effectively prevents the degradation of the network [20].

The PS-FPN structure generally consists of four detection layers. We introduce a detection layer at the 4×4 downsample stage of the Swin transformer, enhancing the network's capability to detect small-sized and medium-sized targets in SAR images. This enhancement is achieved by utilizing the high adaptability of the receptive field in the shallow feature layers to effectively adapt to small target sizes. PS-FPN preserves the bottom-up feature fusion of the FPN architecture while introducing a top-down feature fusion between the 4×4 times downsample and the 8×8 times downsample. These approaches effectively flow information between the shallow feature layers and enrich the shallow feature information.

In a word, the cross-layer feature aggregation and single-branch feature aggregation structures play a vital role in the PS-FPN structure. We use two cross-layer feature aggregation and single-branch feature aggregation structures between detection layer 1 and detection layer 2, and one cross-layer feature aggregation structure between other deep detection layers. The shallow feature layer in our model is characterized by a large size feature map and small receptive field. It is primarily responsible for detecting small- and medium-sized targets in SAR images. In the PS-FPN framework, a bottom-up structure is utilized to transmit deep semantic information upward, ensuring the preservation of detailed features from deeper layers. This approach effectively addresses the problem of lacking shallow feature information related to small targets, thereby significantly improving overall performance. Furthermore, in order to address the potential model degradation caused by increasing the number of layers and complexity of the structure, the single-branch feature aggregation is incorporated in detection layer 1 and detection layer 2. It is worth mentioning that the proposed FPN-based structure for SAR images mainly focuses on enhancing the feature information of shallow feature layers 1 and 2. The aim of this approach is to emphasize the crucial feature information extracted by the Swin transformer in the shallow layers, thereby improving the detection ability of small- and medium-sized targets.

To keep the network lightweight, DSC [61] is used instead of convolution in PS-FPN to extract feature information more efficiently. DSC involves two processes: depthwise convolution (DW) and pointwise convolution (PW). DW convolution captures channel-specific spatial information, whereas PW convolution integrates cross-channel information and reduces computational complexity. They enable feature extraction and fusion within each channel, effectively reducing the parameters involved in the operation. Moreover, compared to other feature fusion modules, such as the PAN module, our approach significantly simplifies the module's processing of deep feature maps, resulting in minimal impact on SAR image detection accuracy while maintaining the model's lightweight.

D. Novel Attention Mechanism Strategy Mix-Attention

SAR images have a scattering imaging mechanism, which makes it difficult to effectively distinguish between background and targets. In addition, the presence of speckle noise further complicates object detection networks used for SAR image detection. To overcome these difficulties, we propose a new method that utilizes attention mechanisms to effectively suppress the network's interference from the noise and background in SAR images. Specifically, we introduce a novel attention mechanism mix-attention, which is tailored to the characteristics of different detection layers. Mix-attention is composed of two components, namely CA [62] and CBAM [63], both of which possess vital feature focusing capabilities and lightweight plug-and-play characteristics.

The structure of the mix-attention is shown in Fig. 4. To improve the detection accuracy of small objects, we use the

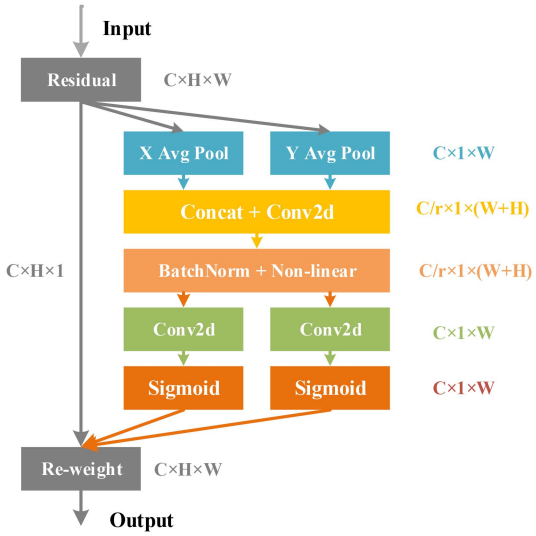


Fig. 5. Structure of the coordinate attention (CA).

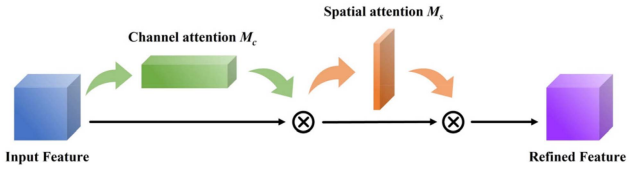


Fig. 6. Structure of the convolutional block attention module (CBAM).

CA attention mechanism on the first two shallow detection layers of the model. This approach helps to enhance the precise localization ability of the shallow feature layer. Next, we apply the CBAM attention mechanism on the two deep detection layers to compensate for the lack of spatial information. This process enhances the concentration and extraction of deep feature information, facilitating the transmission of more comprehensive feature details to the shallow layer. As a result, the detection of small targets is significantly strengthened.

In summary, to improve the position regression ability of the shallow layer for small targets, we use CA to fuse the target's position information into the feature information. In addition, the deep detection layer possesses abundant semantic information, a compact feature map size, and a wealth of channel information. Therefore, we use CBAM to enrich the feature information. These enhancements effectively improve the network's detection ability for SAR images while incurring almost no computational overhead. Following is a brief introduction to the structure of CA and CBAM.

1) CA: Recent research in mobile network design has shown that channel attention can significantly boost model performance. However, these approaches often neglect location information, which is critical for generating spatially selective attention maps. To address this issue, Hou et al. [62] proposed a CA that combines channel attention with location information. As is shown in Fig. 5, this technique captures remote dependencies along one direction and preserves precise location information along another. The resulting feature maps are then encoded into a pair of direction-aware and location-sensitive attention maps, respectively. These maps can be applied to the input feature maps to enhance the representation of the attention object. The formula is as follows:

$$z_{\mathrm{c}}^{\mathrm{h}}(h)=\frac{1}{W}\sum_{0\le i\le W}x_{\mathrm{c}}(h,i) \tag{3}$$

$$z_{\mathrm{c}}^{\mathrm{w}}(w)=\frac{1}{H}\sum_{0\le j\le H}x_{\mathrm{c}}(j,w) \tag{4}$$

$$f=\delta(F_1([z^{\mathrm{h}},z^{\mathrm{w}}])) \tag{5}$$

$$g^{\mathrm{h}}=\sigma(F_{\mathrm{h}}(f^{\mathrm{h}})) \tag{6}$$

$$g^{\mathrm{w}}=\sigma(F_{\mathrm{w}}(f^{\mathrm{w}})) \tag{7}$$

$$y_{\mathrm{c}}(i,j)=x_{\mathrm{c}}(i,j)\times g_{\mathrm{c}}^{\mathrm{h}}(i)\times g_{\mathrm{c}}^{\mathrm{w}}(j) \tag{8}$$

where x_{c} represents the input feature$F_{\mathrm{input}}\in R^{W\times H\times C}$, $z_{\mathrm{c}}^{\mathrm{w}}$ and $z_{\mathrm{c}}^{\mathrm{w}}$ represent the output result of AvgPool, F_1 represents concat and convolution operation,δ and σ represent BN and activation function, respectively, g^{h} andg^{w} represent focus results of the feature in x and y directions, respectively, and y_{c}represents the output result of the CA.

2) CBAM: CBAM is a simple and effective attention module for feedforward CNNs, as shown in Fig. 6. Given an intermediate feature map, CBAM infers the attention map along two independent dimensions: channel and space. The resulting attention map is then multiplied by the input feature map to modify the features adaptively. CBAM effectively combines spatial and channel information in the feature map, enhancing the network's detection performance. As a lightweight and versatile module, CBAM can be seamlessly integrated into any CNN architecture with minimal overhead. The formula for CBAM is as follows:

$$F'=M_{\mathrm{c}}(F)\times F \tag{9}$$

$$F''=M_{\mathrm{s}}(F')\times F' \tag{10}$$

$$M_{\mathrm{c}}(F)=\sigma(MLP(AvgPool(F)))+MLP(MaxPool(F)) \tag{11}$$

$$M_{\mathrm{s}}(F)=\sigma(f^{7\times 7}([AvgPool(F);MaxPool(F)])) \tag{12}$$

where F represents the input feature$F_{\mathrm{input}}\in R^{W\times H\times C}$, and M_{c} and M_{s} represent the spatial attention module and channel attention module, respectively.

IV. Experiments

In order to evaluate the performance of the proposed algorithm in SAR image object detection task, experiments were carried out on three SAR image datasets, i.e., MSAR-1.0 [43], SSDD [44], and SAR-Ship-Dataset [45]. As shown in Fig. 7, we list the number of each category in the three datasets. The detailed information about these datasets is described in Section IV-A. Section IV-B describes the relevant settings of the experiment. Section IV-C presents the evaluation criteria used in this article. Section IV-D provides a comprehensive analysis of the

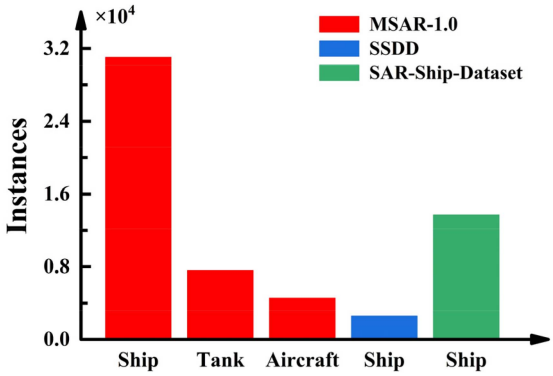


Fig. 7. Number of labels of each category on different datasets.

comparative results achieved on MSAR-1.0, SSDD, and SAR-Ship-Dataset. Section IV-E presents the results of the ablation experiment on MSAR-1.0.

A. Datasets Introduction and Processing

1) MSAR-1.0: The MSAR-1.0 dataset [43] comprises a total of 28449 images captured by the Hysi-1 satellite and the Gaofen-3 satellite. This dataset encompasses diverse scenarios such as airports, ports, nearshore areas, islands, offshore regions, urban areas, etc. In addition, the dataset has four label categories, namely aircraft, tank, bridge, and ship. Specifically, the dataset contains 1851 bridge instances, 39 858 ship instances, 12 319 tank instances, and 6368 aircraft instances. Most images of the MSAR-1.0 dataset have a resolution of 256×256 pixels, whereas bridge instances are presented at a higher resolution of 2048×2048 pixels. To specifically assess the detection performance of small- and medium-sized targets in SAR images, we apply a preprocessing step on the MSAR-1.0 by excluding the larger bridge targets. Subsequently, the dataset is divided into training, validation, and test sets following a ratio of 6:2:2.

2) SSDD and SAR-Ship-Dataset: The SSDD [44] includes 1160 images captured by SARs labeled with bounding boxes and a predefined category as ship. We adjust the resolution of the SSDD dataset to 416×416 pixels. The SAR-Ship-Dataset [45] includes 39 729 images that are captured by SARs and labeled with bounding boxes and a predefined category of ship. The resolution of the dataset is 256×256 pixels. To further assess the effectiveness of our method in detecting small and densely packed targets of SAR images, we perform a selection process on the SAR-Ship-Dataset. Specifically, we process this dataset exclusively comprising images that contain two or more ships, thereby focusing on challenging scenarios with dense target arrangements. The final dataset consists of 4918 images and labels. SSDD and SAR-Ship-Dataset are divided into the training set and validation set with the ratio of 8:2. In addition, the validation set is generated as the test set. The validation and test set are only used for model evaluation and do not participate in the model training process.

B. Implementation Details

The proposed method is implemented on Pytorch 1.10.0, CUDA 11.3 and CUDNN 8.2.1 with Intel(R) Core (TM) i7-12700 @ 2.10 GHz CPU and a NVIDIA GeForce RTX 3070Ti GPU. The operations of random flips, random scaling (between 0.6 and 1.3), crop, color dithering, and mosaic are used for data enhancement, and the Adam operation is used to optimize the overall goal. The learning rate of the network is set to 1e−3 for the first 200 iterations and 1e−4 for the last 100 iterations. The weight delay is 0.0005 and the momentum is 0.937.

C. Evaluation Criteria

In order to accurately quantify the detection performance of the proposed model on SAR images, this article employs the highest recognized evaluation index of multiple object detection tasks: mean average precision (mAP). AP is the process of calculating the average accuracy of 10 IoU thresholds from 0.5 to 0.95 in steps of 0.05 for each class. AP50 is the average accuracy calculated with an IoU threshold of 0.5. mAP is the average of AP of all categories, and mAP50 is the average accuracy of mAP calculated with the IoU threshold of 0.5. The calculation process of mAP is as follows:

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

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

$$mAP = \frac{1}{n}\sum\nolimits_{k=1}^{n} J(Precision, Recall)_k \tag{15}$$

where TP (true positive) represents real examples, FP represents false positive examples, FN represents false negative examples, n represents the number of categories, and $J(Precision, Recall)_k$ represents the average precision (AP) function.

D. Comparative Experiments

In order to verify the robustness of the proposed method, experiments were carried out on MSAR-1.0, SSDD, and SAR-Ship-Dataset with 256×256, 416×416, and 256×256 pixels, respectively. Table I compares the detection performance of the baseline model and the proposed algorithm on MSAR-1.0, SSDD, and SAR-Ship-Dataset, including mAP, Parameters, GFLOPS, and frames per second (FPS). The baseline model is YOLOv7 with Swin transformer as the backbone network. Meanwhile, the detection AP of both the baseline model and our algorithm for each category within the three datasets is listed in Table I.

As shown in Table I, compared with the experimental results of the baseline network on the MSAR-1.0 dataset, the detection AP of the proposed algorithm on aircraft, ship, and tank is increased by 23.2%, 4.8%, and 6.4%, respectively, and the total mAP is increased by 11.6%. The results demonstrate a substantial improvement in the detection accuracy of the proposed algorithm within the MSAR-1.0 dataset. Particularly noteworthy is the significantly higher accuracy growth observed for small objects such as aircraft and ships, as compared to larger objects

TABLE I
mAP (%) OF EACH CATEGORY OF OBJECTS IN VALIDATION SET OF MSAR-1.0, SSDD, AND SAR-SHIP-DATASET

Dataset	Method	mAP50 (%)	AP50 (%)			Parameters	GFLOPS	FPS	Input-size
			Aircraft	Ship	Tank				
MSAR-1.0	Baseline	60.9	17.6	78.1	87.2	52.2M	20.405G	24.7	256×256
	PSFNet	**72.5**	**40.8**	**82.9**	**93.6**	**40.2M**	**24.208G**	**23.5**	
SSDD	Baseline	90.8	/	90.8	/	52.2M	47.729G	15.4	416×416
	PSFNet	**92.7**		**92.7**		**40.2M**	**57.767G**	**14.3**	
SAR-Ship-Dataset	Baseline	87.9	/	87.9	/	52.2M	20.405G	24.7	256×256
	PSFNet	**90.9**		**90.9**		**40.2M**	**24.208G**	**23.4**	

The significance of bold values means the results of our proposed method.

TABLE II
COMPARISON RESULTS OF DIFFERENT METHODS ON MSAR-1.0

Method	Backbone	mAP50 (%)	AP50 (%)			Parameters	GFLOPS
			Aircraft	Ship	Tank		
YOLOv4 [15]	Darknet53	46.3	9.6	62.7	66.6	63.9M	22.636G
YOLOv5	CSP-DarkNet53	56.7	6.3	76.9	86.9	46.6M	18.284G
YOLOv7 [16]	CSP-DarkNet53	51.8	3.1	76.4	80.7	37.2M	16.824G
YOLOv8	CSP-DarkNet53	64.6	21.1	82.4	90.3	43.6M	26.466G
Retinanet [64]	Resnet50	47.3	0.1	60.4	81.5	55.4M	36.023G
SSD [17]	VGG16	46.2	6.7	63.4	68.3	23.9M	61.006G
EfficientDet [65]	Efficientnet	45.2	0.1	58.9	76.5	33.4M	10.431G
Faster Rcnn [13]	Resnet50	52.3	2.3	72.1	82.4	136.8M	184.890G
Centernet [52]	Resnet50	65.5	10.3	90.2	96.2	32.7M	17.554G
TPH-YOLOv5 [60]	CSP-DarkNet53	69.1	33.4	82.2	95.2	45.4M	51.267G
SGEA [6]	Resnet50	66.6	10.4	92.9	96.5	34.2M	17.574G
Our method	**Swin transformer**	**72.5**	**40.8**	**82.9**	**93.6**	**40.2M**	**24.208G**

The significance of bold values means the results of our proposed method.

such as tanks. Experimental results on SSDD and SAR-Ship-Dataset datasets show that the proposed algorithm improves the mAP by 1.9% and 3.0%, respectively. It further demonstrates the superior detection performance of the improved model on small objects.

In addition, after analyzing Table I, it is evident that our proposed network exhibits desirable lightweight performance. Specifically, there is almost no change in the amount of GFLOPS and FPS when compared to the baseline network, whereas the number of parameters is significantly reduced. This indicates that our proposed innovations in this network yield a highly efficient and effective performance.

Moreover, to comprehensively evaluate the effectiveness of the proposed algorithm and its detection capability on SAR images, we have performed comparative analyses with other state-of-the-art (SOTA) algorithms on the MSAR-1.0, SSDD, and SAR-Ship-Dataset.

1) Detection Results on MSAR-1.0: In Table II, we present a comparative analysis of the detection results achieved by the proposed algorithm and other SOTA algorithms on the MSAR-1.0. The evaluated algorithms include YOLOv4 [15], YOLOv5, YOLOv7 [16], YOLOv8, Retinanet [64], SSD [17], EfficientDet [65], Centernet [52], Faster R-CNN [13], TPH-YOLOv5 [60], and SGEA [6]. According to the results presented in Table II, the proposed method achieves the highest AP among other representative detectors. Specifically, the mAP of our approach reached 72.5%, which is a 20.7% improvement over YOLOv7. Moreover, our method achieves higher detection AP than YOLOv7 in each category, demonstrating the effectiveness of our innovations combined with YOLOv7 for SAR image detection. Furthermore, compared to the current SOTA algorithms, our method outperforms most algorithms in detecting medium-sized targets like ships and large targets like tanks, with similar precision to Centernet, TPH-YOLOv5, and SGEA. However, when it comes to detecting small targets like aircraft, PSFNet's detection precision surpasses other SOTA algorithms by a significant margin. Not only does it outperform Centernet by 7.4% in precision but it also exceeds other SOTA algorithms by more than 30%. This further demonstrates the unique advantage of PSFNet in detecting small- and medium-sized targets in SAR images. At the same time, the proposed method maintains a similar number of parameters and GFLOPS compared to these networks, providing excellent lightweight performance. Sacrificing lightweight performance for detection accuracy is avoided and a well-balanced model is achieved, which performs well in terms of both accuracy and efficiency of SAR object detection. It is worth noting that, to ensure a fair comparison, we only selected models with strong lightweight performance and did not include larger networks in the same model series.

In Fig. 8, we illustrate the detection results of the proposed method alongside SOTA models. The results showcase the superior detection performance of our method, particularly on small objects in SAR scenes. In the first row of the results, it is evident that PSFNet successfully detects all tanks, which are considered large-size targets in the MSAR-1.0, accurately regressing their coordinates. Conversely, YOLOv7 exhibits numerous false detections, and its regression of target coordinates is inaccurate. In the second row, we analyze the detection results

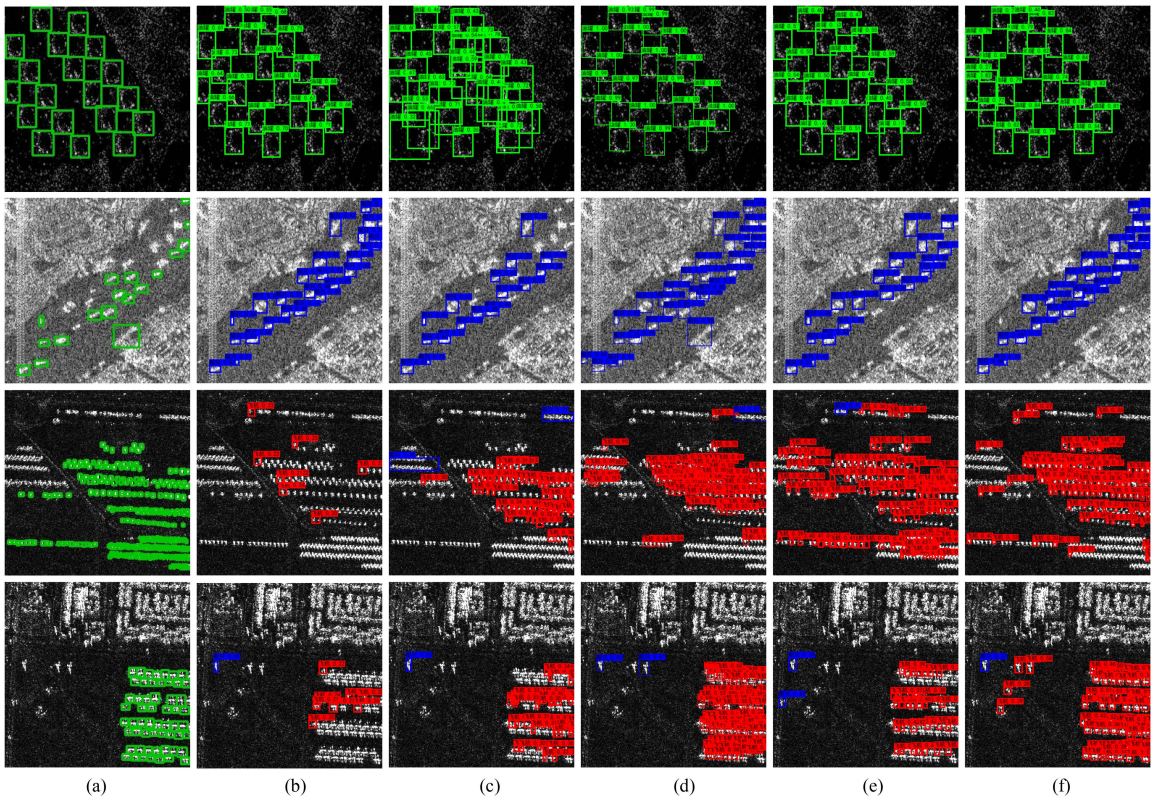


Fig. 8. Different model detection results based on the MSAR-1.0 dataset. (a) Ground truth. (b) Detection results of YOLOv5. (c) Detection results of YOLOv7. (d) Detection results of Faster RCNN. (e) Detection results of Centernet. (f) Detection results of our method. The number in the detection box represents the confidence of the target. The targets in the blue boxes are ships, tanks in green, and aircraft in red.

for medium-sized targets. It is worth noting that compared to other models, PSFNet achieves fewer false detections in this category. This superior performance can be attributed to the network's utilization of mix-attention strategy, which enables a more concentrated focus on targets being detected. In the third and fourth rows of Fig. 8, we present the detection results of our proposed method and compare them with SOTA algorithms for small-sized target ships. Our algorithm demonstrates superior performance in terms of detection accuracy, exhibiting fewer false detections and missed detection rates compared to other SOTA algorithms. These results highlight the effectiveness of our proposed method for small target detection in SAR images. Notably, YOLOv7 shows relatively poor performance in detecting small targets, with a significant number of ships going undetected. This further emphasizes the effectiveness of our innovative additions to the YOLOv7 network. To further demonstrate the robustness of our network's detection capabilities, a heatmap analysis is performed on MSAR-1.0 dataset. Heatmap divides the data into many small regions and assigns colors to represent the relative values or weights of each region, which enables a clear and intuitive representation of the data distribution. In Fig. 9, we present the heatmap analysis results of our model on the test set. The heatmap vividly demonstrates the excellent performance of our model in accurately locating small target aircraft. To assess the practicality of the proposed algorithm, we evaluate its detection performance on high-resolution SAR images. In Fig. 10, we demonstrate the detection results of

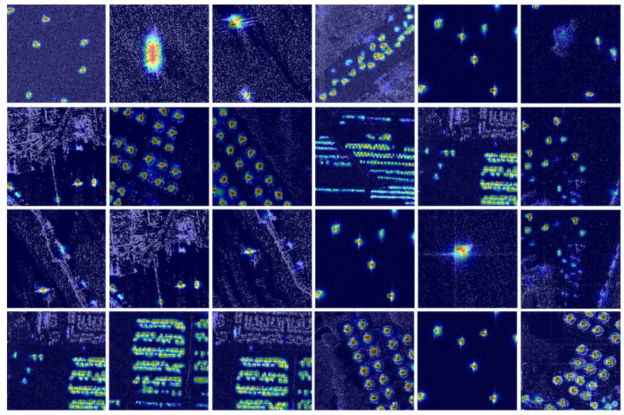


Fig. 9. Heatmap visualization of our mix-attention-based method on the MSAR-1.0 dataset.

our network on an image with 2048×2048 resolution from the MSAR-1.0 dataset. It is evident that our algorithm maintains its superior detection performance even on large-sized images. Partial detection results of PSFNet are presented in Fig. 11.

2) Detection Results on SSDD and SAR-Ship-Dataset: As the detection of small target ships remains a hot research topic in SAR images, we employed the SSDD to assess the detection capability of our algorithm on small target ships. Table III illustrates the performance comparison between the proposed method and several SOTA detectors, i.e., YOLOv4

TABLE III
COMPARISON RESULTS OF DIFFERENT METHODS ON SSDD AND SAR-SHIP-DATASET

Method	Backbone	SSDD		SAR-Ship-Dataset		Parameters
		mAP50 (%)	GFLOPS	mAP50	GFLOPS	
YOLOv4 [15]	Darknet53	84.5	59.967G	76.2	22.636G	63.9M
YOLOv5	CSP-DarkNet53	91.1	48.416G	90.4	18.284G	46.6M
YOLOv7 [16]	CSP-DarkNet53	91.3	44.425G	84.9	16.824G	37.2M
YOLOv8	CSP-DarkNet53	86.7	69.760G	89.7	26.466G	43.6M
Retinanet [64]	Resnet50	54.8	95.311G	71.2	36.023G	55.4M
SSD [17]	VGG16	62.8	115.787G	86.8	61.006G	23.9M
EfficientDet [65]	Efficientnet	87.3	41.128G	70.8	10.431G	33.4M
Faster Rcnn [13]	Resnet50	87.4	252.631G	79.7	184.890G	136.8M
Centernet [52]	Resnet50	83.3	46.354G	88.9	17.554G	32.7M
TPH-YOLOv5 [60]	CSP-DarkNet53	92.0	102.400 G	90.2	51.267G	45.4M
SGEA [6]	Resnet50	88.2	46.397G	90.3	17.574G	34.2M
Our method	**Swin transformer**	**92.7**	**57.767G**	**90.9**	**24.208G**	**40.2M**

The significance of bold values means the results of our proposed method.

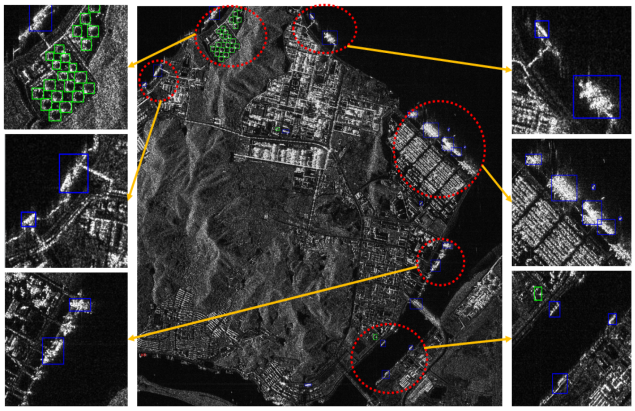


Fig. 10. Results in large-scale SAR images on MSAR-1.0. The targets in the blue boxes are ships, tanks in green, and aircraft in red.

[15], YOLOv5, YOLOv7 [16], YOLOv8, Retinanet [64], SSD [17], EfficientDet [65], Centernet [52], Faster RCNN [13], TPH-YOLOv5 [60], and SGEA [6], using the SSDD dataset. These experimental results demonstrate that the proposed algorithm achieves a mAP of 92.7%, surpassing the performance of SOTA methods.

To further validate our method's detection ability for small and dense targets in SAR images, we evaluated its performance on the SAR-Ship-Dataset. Table III presents a performance comparison of the proposed method with several SOTA detectors mentioned above on the SAR-Ship-Dataset. Experimental results show that the mAP of the proposed algorithm reaches 90.9%, outperforming the SOTA methods. Furthermore, Fig. 12 presents selected detection results on SSDD and SAR-Ship-Dataset, providing additional evidence of the effectiveness of our proposed method.

In summary, the performance of the proposed algorithm was evaluated using three different types of SAR image datasets. The results demonstrate that the algorithm achieves satisfactory detection performance on various types of SAR image targets, including general targets, small targets, densely arranged small targets, and super-resolution image targets. These findings provide strong evidence for the effectiveness of the proposed algorithm.

E. Ablation Study

To objectively assess the influence of the proposed innovations on the detection performance, ablation experiments were conducted on the MSAR-1.0 dataset under identical experimental conditions. Table IV presents the results of the ablation experiments, followed by a comprehensive analysis of the obtained results.

1) Effect of Swin Transformer Combined With YOLOv7: We propose a novel object detection network that combines Swin transformer and YOLOv7 in this article, aiming to enhance feature extraction and accelerate inference capabilities. As shown in Table IV, the mAP of the YOLOv7 network on the MSAR-1.0 dataset is 51.8%. However, after replacing the YOLOv7 backbone with Swin transformer, the mAP significantly improved, reaching 60.9%, which is a significant improvement of 9.1%. Furthermore, we observed that the detection accuracy (AP50) for small target aircraft in the MSAR-1.0 increased from 3.1% to 17.6%, showing a growth of 14.5%. This indicates that the combination of Swin Transformer and YOLOv7 can effectively enhance the SAR object detection accuracy, particularly for small targets. The enhanced detection capability is mainly attributed to the global attention mechanism of Swin Transformer, which enables better capture of target feature information and enhances the algorithm's feature extraction capability.

2) Effect of New Layer: To improve the detection accuracy of small targets in SAR images, a new detection layer was incorporated at the 4×4 downsample layer of the network's backbone. The results in Table IV clearly demonstrate the effectiveness of the added detection layer in improving the detection performance of small targets. Specifically, the proposed method achieves a significant 5.6% increase in mAP compared to the baseline network. Furthermore, the detection accuracy of small target aircraft increased by 9.9%. This highlights the crucial role of the additional layer in enhancing the accuracy of small object detection in SAR images.

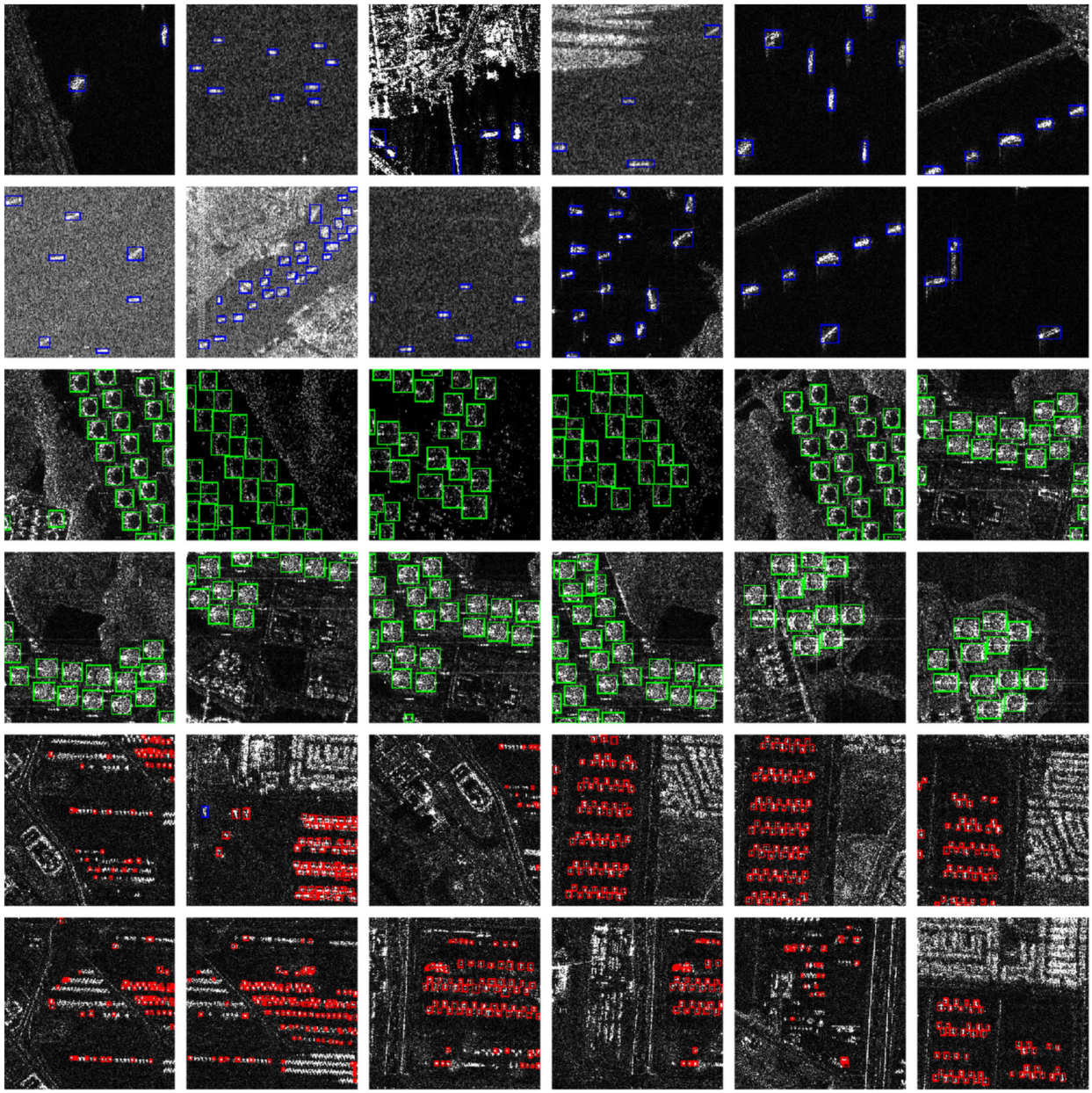


Fig. 11. Detection results of PSFNet on MSAR-1.0. The targets in the blue boxes are ships, tanks in green, and aircraft in red.

TABLE IV
ABLATION EXPERIMENT ON MSAR-1.0

YOLOv7 $_{PAN}$	Swin Transformer	New layer	FPN	PS-FPN	Mix-attention	mAP50 (%)	Aircraft (AP50 %)	Parameters	GFLOPS
√						51.8	3.1	37.2M	16.824G
√	√					60.9	17.6	52.2M	20.405G
√	√		√			61.2	20.2	45.0M	18.872G
√	√			√		64.3	22.5	46.3M	22.144G
√	√				√	62.5	20.3	52.9M	20.409G
√	√			√	√	65.5	25.2	46.9M	22.148G
√	√	√				66.5	27.5	52.8M	22.730G
√	√	√	√			65.9	25.4	45.2M	20.425G
√	√	√		√		70.5	40.0	39.5M	24.203G
√	√	√		√	√	**72.5**	**40.8**	**40.2M**	**24.208G**

The significance of bold values means the results of our proposed method.

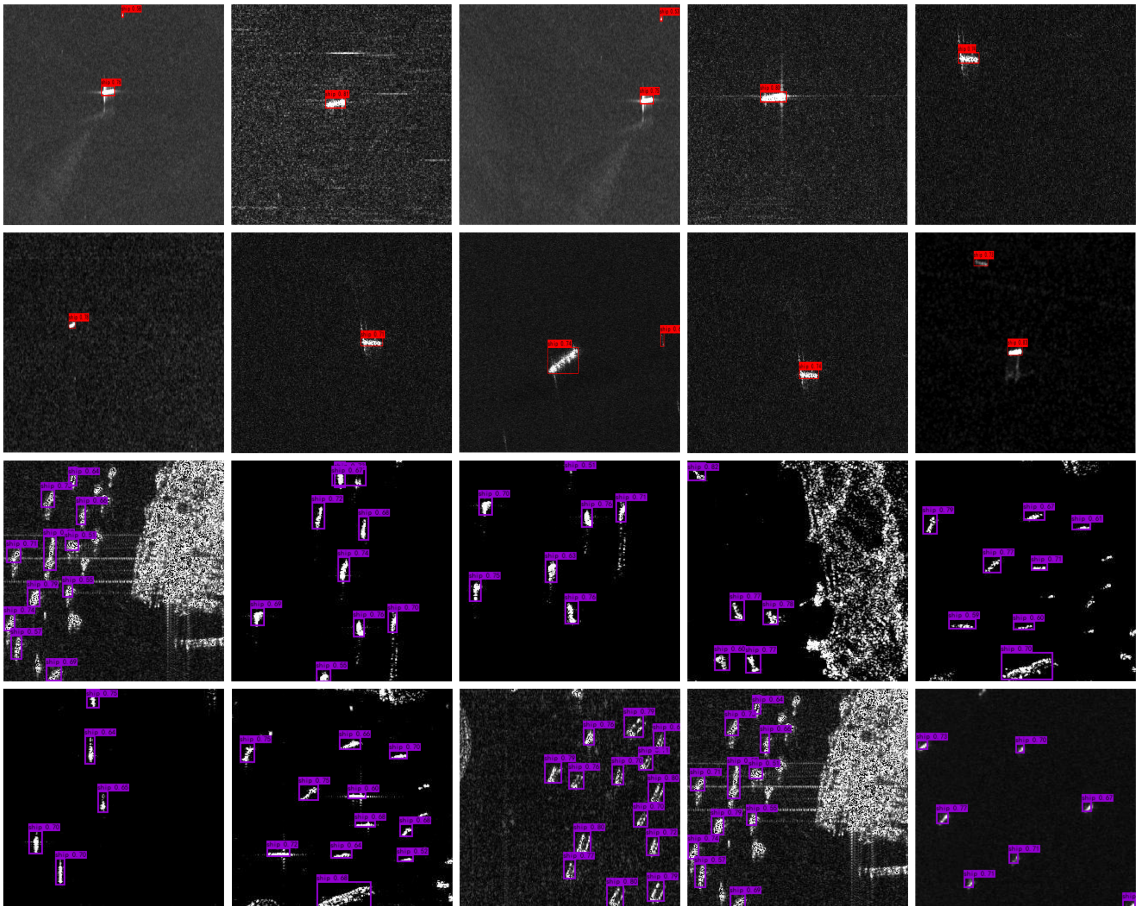


Fig. 12. Detection results of PSFNet on SSDD and SAR-Ship-Dataset. The number in the detection box represents the confidence of the target. The targets in the red boxes are ships in the SSDD and the targets in the purple boxes are ships in the SAR-Ship-Dataset.

3) Effect of PS-FPN: PS-FPN structure enables bottom-up propagation of deep feature information to shallow layers, resulting in enriched feature representations of small targets. To evaluate the effectiveness of PS-FPN, we conducted ablation experiments to compare the detection performance of the baseline network with and without the PS-FPN architecture. In addition, we compared the performance of the baseline network augmented with a new layer to that of the same network with PS-FPN incorporated. The experimental results demonstrate a notable improvement in detection accuracy after the incorporation of our proposed approach PS-FPN. Specifically, compared with the two networks without PS-FPN, the mAP increases by 1.8% and 4.0%, respectively. Given the need to keep the model lightweight, some modules within the PS-FPN architecture utilize DSCs in lieu of regular convolutions, which sacrifice a negligible loss of accuracy in exchange for a significant reduction in model parameters. As shown in Table IV, the parameters of the model with PS-FPN are significantly reduced by 12.5% compared with the baseline network, demonstrating its lightweight and efficiency.

To further validate the effectiveness of the PS-FPN module, this article conducted experiments and analyzed the results using FPN, PAN, and PS-FPN in ablation experiments. It is worth noting that the default feature fusion module in YOLOv7 is PAN, so the baseline network used PAN module. Both the baseline network and the baseline combined with the new layer network replaced PAN with FPN and PS-FPN. From Table IV, it can be seen that the model using PS-FPN achieved the highest mAP of 64.3% and 70.5%, and the AP for aircraft also outperformed the other two modules, reaching 22.5% and 40.0%. At the same time, the algorithm using the PS-FPN module did not bring significant changes in parameters and GFLOPS, remaining lightweight like FPN and PAN. This indicates that PS-FPN does not sacrifice model complexity for detection accuracy but rather improves network design performance while maintaining model complexity and enhancing SAR small target detection accuracy.

4) Effect of Mix-Attention: The mix-attention strategy is a technique used in SAR image detection to alleviate the interference of speckle noise and improve the detection accuracy of small targets. To assess the effectiveness of the mix-attention strategy, three experiments were conducted using different network architectures. Comparisons were made between networks equipped with this module and those without it to evaluate their performance. The results, shown in Table IV, demonstrate that after using the mix-attention strategy, the detection accuracy mAP was further improved, reaching 62.5%, 65.5%, and 72.0%, respectively. Furthermore, the results of our three mix-attention experiments indicate that the network

TABLE V
ATTENTION EXPERIMENT ON MSAR-1.0

Method	mAP 50 (%)	AP50 (%)			Parameters	GFLOPS	FPS
		Aircraft	Ship	Tank			
A (base-net)	70.5	39.6	79.2	92.7	39.5M	24.203G	24.5
B (CA+none)	71.0	40.2	80.1	92.7	39.5M	24.205G	23.8
C (none+CBAM)	68.5	30.2	82.5	92.7	40.2M	24.206G	23.9
D (CA+CA)	67.0	27.2	80.6	93.2	39.8M	24.212G	23.4
E (CBAM+CBAM)	68.5	33.7	80.3	92.1	40.3M	24.209G	23.5
F (SE+SE)	68.8	33.2	80.0	93.1	39.7M	24.206G	25.3
G (ECA+ECA)	65.8	24.8	80.5	92.1	39.5M	24.206G	23.8
Our method	**72.5**	**40.8**	**82.9**	**93.6**	**40.2M**	**24.208G**	**23.5**

The significance of bold values means the results of our proposed method.

incorporating mix-attention shows a minimal increase of less than 1% in the number of parameters and a negligible increase of 0.1% in GFLOPS. This finding provides evidence that mix-attention also delivers lightweight performance, which is highly desirable for efficient and effective network architectures.

To further assess the effectiveness of the mix-attention strategy, we conducted experiments on our proposed network, exploring various combinations of attention mechanisms. Specifically, we utilize the network architecture proposed in this article as the base-net for each experiment, excluding the mix-attention component. For brevity, we used A, B, C, D, E, F, G to denote the base-net network, with CA applied to the first two detection layers of the baseline network, CBAM applied to the last two detection layers, CA applied to all detection layers, CBAM applied to all detection layers, SE-Net (SE) [66] applied to all detection layers and ECA-Net (ECA) [67] applied to all detection layers. The results presented in Table V demonstrate that our integration of attention mechanisms achieves the highest detection accuracy, providing evidence for the superior performance of the mix-attention method compared to other combinations of attention mechanisms for SAR images. In addition, we observed that the mAP in method A outperformed the attention mechanism experiments in methods C, D, E, F, and G. This indicates that attention mechanisms should not be blindly added to the network, as it may not only fail to improve network performance but also potentially decrease the algorithm's detection accuracy. This indirectly demonstrates the rationality and effectiveness of using CA and CBAM in mix-attention for addressing issues related to small detection layer size and precise localization of small targets in shallow and deep feature focus, respectively.

V. CONCLUSION

Object detection in SAR images is a challenging task. In this article, a novel detection algorithm based on Swin transformer and YOLOv7 for SAR images is proposed. First, in order to enrich the semantic and spatial features of small targets, an innovative feature aggregation module PS-FPN is proposed. This module not only improves the semantic and spatial information of small targets but also strengthens the detailed information needed by small targets while maintaining lightweight structure of the network. Second, we propose a novel attention mechanism mix-attention to identify more attention regions in scenarios with dense objects. In addition, we add one more prediction head to extract shallow features that effectively preserve small targets' feature information. To verify the effectiveness of the proposed algorithm, extensive experiments are carried out on several challenging SAR image datasets. The experimental results demonstrate that our method outperforms other SOTA detectors in terms of detection accuracy and computation efficiency in various SAR scenarios. In summary, our approaches address these challenges of SAR image detection and provide promising results for real-world applications. However, the proposed algorithm requires more time in the training process and a larger dataset to maintain its detection advantage, a challenge we will try to solve in the future.

REFERENCES

[1] 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.

[2] M. Tello, C. Lopez-Martinez, and J. J. Mallorqui, "A novel algorithm for ship detection in SAR imagery based on the wavelet transform," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 201–205, Apr. 2005.

[3] N. Li, Q. Shen, L. Wang, Q. Wang, Z. Guo, and J. Zhao, "Optimal time selection for ISAR imaging of ship targets based on time-frequency analysis of multiple scatterers," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4017505, doi: 10.1109/LGRS.2021.3103915.

[4] N. Li, Z. Lv, Z. Guo, and J. Zhao, "Time-domain notch filtering method for pulse RFI mitigation in synthetic aperture radar," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4013805, doi: 10.1109/LGRS.2021.3077247.

[5] S.-W. Chen, X.-C. Cui, X.-S. Wang, and S.-P. Xiao, "Speckle-free SAR image ship detection," *IEEE Trans. Image Process.*, vol. 30, pp. 5969–5983, 2021, doi: 10.1109/TIP.2021.3089936.

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

[7] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2015.

[8] X. Qin, S. Zhou, H. Zou, and G. Gao, "A CFAR detection algorithm for generalized gamma distributed background in high-resolution SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 806–810, Jul. 2013.

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

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

[11] 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, 2014, pp. 580–587.

[12] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 28, pp. 91–99.

[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

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

[16] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[17] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[18] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713, doi: 10.1109/TGRS.2022.3160007.

[19] J. Chen et al., "DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2021, doi: 10.1109/JSTARS.2020.3037893.

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

[21] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.

[22] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8007805, doi: 10.1109/LGRS.2021.3056416.

[23] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.

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

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

[26] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[27] B. Wu et al., "Visual transformers: Token-based image representation and processing for computer vision," 2020, *arXiv:2006.03677*.

[28] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted Windows," 2021, *arXiv:2103.14030*.

[29] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *Computer Vision—ECCV 2020*, vol. 12373, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Glasgow, U.K.: Springer, 2020, pp. 323–339.

[30] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision—ECCV 2020*, vol. 12346, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Glasgow, U.K.: Springer, 2020, pp. 213–229.

[31] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[32] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to sequence perspective with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 6881–6890.

[33] M. Chen et al., "Generative pretraining from pixels," in *Proc. 37th Int. Conf. Mach. Learn.*, Jul. 2020, vol. 119, pp. 1691–1703.

[34] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12873–12883.

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

[36] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12299–12310.

[37] P. Wang, L. Wang, H. Leung, and G. Zhang, "Super-resolution mapping based on spatial–spectral correlation for spectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2256–2268, Mar. 2021.

[38] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3578–3587.

[39] 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.*, Jul. 2017, pp. 2117–2125.

[40] 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.*, Jun. 2018, pp. 8759–8768.

[41] 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.

[42] G. Ghiasi, T. Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7029–7038.

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

[44] 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.

[45] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sens.*, vol. 11, no. 7, Mar. 2019, Art. no. 765, doi: 10.3390/rs11070765.

[46] L. M. Novak, M. C. Burl, and W. W. Irving, "Optimal polarimetric processing for enhanced target detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 29, no. 1, pp. 234–244, Jan. 1993.

[47] J. Ai et al., "Robust CFAR ship detector based on bilateral-trimmed-statistics of complex ocean scenes in SAR imagery: A closed-form solution," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 3, pp. 1872–1890, Jun. 2021.

[48] W. Huo, Y. Huang, J. Pei, Q. Zhang, Q. Gu, and J. Yang, "Ship detection from ocean SAR image based on local contrast variance weighted information entropy," *Sensors*, vol. 18, no. 4, 2018, Art. no. 1196.

[49] Z. Shi, X. Yu, Z. Jiang, and B. Li, "Ship detection in high-resolution optical imagery based on anomaly detector and local shape feature," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4511–4523, Aug. 2014.

[50] C. Chen, C. He, C. Hu, H. Pei, and L. Jiao, "A deep neural network based on an attention mechanism for SAR ship detection in multiscale and complex scenarios," *IEEE Access*, vol. 7, pp. 104848–104863, 2019.

[51] J. Ai, R. Tian, Q. Luo, J. Jin, and B. Tang, "Multiscale rotation-invariant haar-like feature integrated CNN-based ship detection algorithm of multiple-target environment in SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10070–10087, Dec. 2019.

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

[53] Z. Sun et al., "An anchor-free detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7799–7816, 2021, doi: 10.1109/JSTARS.2021.3099483.

[54] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 860, doi: 10.3390/rs9080860.

[55] 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, 2021, Art. no. 4209, doi: 10.3390/rs13214209.

[56] Z. Zhou et al., "HRLE-SARDet: A lightweight SAR target detection algorithm based on hybrid representation learning enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5203922, doi: 10.1109/TGRS.2023.3251694.

[57] X. Ma, K. Ji, B. Xiong, L. Zhang, S. Feng, and G. Kuang, "Light-YOLOv4: An edge-device oriented target detection method for remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10808–10820, 2021, doi: 10.1109/JSTARS.2021.3120009.

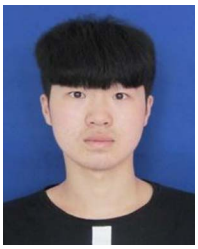
[58] Z. Zhou et al., "FSODS: A lightweight metalearning method for few-shot object detection on SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5232217, doi: 10.1109/TGRS.2022.3192996.

[59] 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, 2023, doi: 10.1109/JSTARS.2022.3221784.

[60] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Montreal, BC, Canada, 2021, pp. 2778–2788.

[61] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 1800–1807.

[62] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 13713–13722.
[63] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. IEEE Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
[64] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.
[65] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. Comput. Vis. Pattern Recognit.*, 2020, pp. 10778–10787.
[66] 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. 2020.
[67] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11531–11539.



Peng Zhou received the B.S. degree in electronic information engineering from Changzhou Institute of Technology, Changzhou, China, in 2021. He is currently working toward the master's degree in information and communication engineering with the School of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China.

His research interests include image processing and object detection in remote sensing images.

Peng Wang (Senior Member, IEEE) received the B.E. degree in microelectronics and the Ph.D. degree in information and communications engineering from the College of Information and Communications Engineering, Harbin Engineering University, Harbin, China, in 2012 and 2018, respectively.

He is currently an Associate Professor and Doctoral Supervisor with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He is also a Hong Kong Scholar with the Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong. In 2016, he was a Visiting Ph.D. Student with the Grenoble Images Parole Signals Automatics Laboratory, Grenoble Institute of Technology, Saint Martin d'Hères, France. He has authored two books and more than 50 articles. His research interests include remote-sensing imagery processing and machine learning.

Dr. Wang is a Reviewer of more than 20 international journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Remote Sensing of Environment*, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

Jie Cao received the B.S. degree in electronic engineering from Nanjing University of Science and Technology, Nanjing, China, in 1983, and the M.S. degree in communication and electronic system from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, in 1986.

He has long been engaged in the technical research work of UAV system, remote control and telemetry, image tracking and recognition, image processing, and other aspects. He has authored/coauthored more than 30 academic papers. He is currently a Second-Level Professor with NUAA.

Mr. Jie was the recipient of more than 10 national and provincial science and technology awards.

Daiyin Zhu (Member, IEEE) received the B.S. degree in electronic engineering from Southeast University, Nanjing, China, in 1996, and the M.S. and Ph.D. degrees in electronics from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, in 1998 and 2002, respectively.

From 1998 to 1999, he was a Guest Scientist with the Institute of Radio Frequency Technology, German Aerospace Center (DLR), Oberpfaffenhofen, Germany, where he was in the field of synthetic aperture radar (SAR) interferometry. In 1998, he joined the Department of Electronic Engineering, NUAA, where he is currently a Professor. He has developed algorithms for several operational airborne SAR systems. His research interests include radar imaging algorithms, SAR ground moving target indication, SAR/ISAR autofocus techniques, SAR interferometry, and multiple-input multiple-output SAR signal processing.

Qiyuan Yin received the B.S. degree in communication engineering from Tongda College of Nanjing University of Posts and Telecommunications, Nanjing, China, in 2021. He is currently working toward the master's degree in information and communication engineering with the School of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing.

His research interests include image processing and object location.

Jiming Lv was born in Dongtai City, Jiangsu Province, China. Since 2021, he has been working toward the Ph.D. degree in information and communication engineering with the Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing, China.

His research interests include deep learning and synthetic aperture radar image processing.

Ping Chen received the Ph.D. degree in signal and information processing from the North University of China, Taiyuan, China, in 2011.

He is currently a Professor with the Departments of Information and Communication Engineering, North University of China, and the Director of the Center for Shanxi Key Laboratory of Signal Capturing and Processing. He has undertaken projects funded by a variety of grants and authored/coauthored more than 80 academic articles in the research areas of image processing, image recognition, and X-ray CT imaging.

Yongshi Jie received the Ph.D. degree in signal and information processing from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2021.

He is currently an Engineer with the Beijing Institute of Space Mechanics and Electricity, China Academy of Space Technology, Beijing. He has finished 10 articles. His research interests include remote sensing image information extraction and deep learning.

Cheng Jiang received the B.S. degree in applied physics and the Ph.D. degree in optical engineering from the College of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing, China, in 2007 and 2013, respectively.

He is currently a Professor with the Beijing Institute of Space Mechanics and Electricity, China Academy of Space Technology, Beijing. He has undertaken projects funded by a variety of grants and authored/coauthored more than 30 academic articles in the research areas of remote sensing data processing and satellite-based application.