


Feature Enhancement Pyramid and Shallow Feature Reconstruction Network for SAR Ship Detection

Lin Bai, *Member, IEEE*, Cheng Yao , Zhen Ye , Dongling Xue , Xiangyuan Lin , and Meng Hui, *Member, IEEE*

Abstract—Recently, convolutional neural network based methods have been studied for ship detection in optical remote sensing images. However, it is challenging to apply them to microwave synthetic aperture radar (SAR) images. First, most of the regions in the inshore scene include scattered spots and noises, which dramatically interfere with ship detection. Besides, SAR ship images contain ship targets of different sizes, especially small ships with dense distribution. Unfortunately, small ships have fewer distinguishing features making it difficult to be detected. In this article, we propose a novel SAR ship detection network called feature enhanced pyramid and shallow feature reconstruction network (FEPS-Net) to solve the above problems. We design a feature enhancement pyramid, which includes a spatial enhancement module to enhance spatial position information and suppress background noise, and the feature alignment module to solve the problem of feature misalignment during feature fusion. Additionally, to solve the problem of small ship detection in SAR ship images, we design a shallow feature reconstruction module to extract semantic information from small ships. The effectiveness of the proposed network for SAR ship detection is demonstrated by experiments on two publicly available datasets: SAR ship detection dataset and high-resolution SAR images dataset. The experimental results show that the proposed FEPS-Net has advantages in SAR ship detection over the current state-of-the-art methods.

Index Terms—Deep learning, feature enhancement pyramid (FEP), SAR ship detection, shallow feature reconstruction (SFR), synthetic aperture radar (SAR).

I. INTRODUCTION

SYNTHETIC aperture radar (SAR) is an active Earth observation system installed on aircraft, satellites, spacecraft, and other flight platforms to observe the Earth around the clock and all-weather, getting more and more attention. With the successful operation of TerraSARX, RADARSAT-2, Sentinel-1, Gaofen 3, and other satellites in orbit, SAR images have been widely used in disaster monitoring, environmental surveying, crop estimation, and so on [1], [2], [3], [4], [5], [6]. With the

rapid progress of radar technology in recent years, the resolution of SAR images has become higher and higher, which makes it possible to detect and identify marine targets with high precision using SAR images.

SAR ship detection mainly employed manual feature approaches in the past. Considering SAR imaging mechanism, most of the traditional detection methods are designed from the perspective of signal processing, such as constant false alarm rate (CFAR) based methods [7], [8], [9], [10], [11], [12]. The CFAR methods mainly use the statistical distribution of false alarm rate and background clutter adjusting thresholds adaptively to detect target regions. For example, G^0 distribution and generalized gamma distribution are used as statistical models of clutter [13], [14]. Besides, Tello et al. [15] proposed a wavelet transform-based detection method to interpret the information through wavelet coefficients by exploiting the differences in the statistical behavior of ships and their surroundings. The CFAR-based SAR ship detection usually detect targets by setting a suitable threshold. However, when complex bright objects exist in background, especially for inshore scenes with many interfering scattered spots and noises, CFAR-based strategies report many false alarms. In addition, these methods cannot achieve end-to-end detection due to inconvenient parameter adjustment, which limits the detection accuracy and efficiency, especially in changing environments.

Recently, convolution neural networks (CNNs) have attracted much attention because of their powerful feature extraction capabilities. CNN-based object detection methods have been developed into two main types. One is based on two-stage detector, which is a coarse-to-fine process that includes region proposal and subsequent classification and regression. The other is based on one-stage detector without region proposal part. Most of the existing two-stage detectors are based on region proposal networks (RPNs) that generate category-independent region proposals in the first stage and then perform classification and regression in terms of these regions in the second stage. Girshick et al. [16] first introduced CNN into object detection by selective search [17] to obtain region proposals and then proposed R-CNN. Faster R-CNN [18] proposed by Ren et al. became the main architecture of the two-stage detector, which proposed a nearly costless RPN. Since the one-stage detector does not have a RPN, its computational cost is lower than that of two-stage detector, and its detection efficiency is higher. As a one-stage detector, YOLO [19] treat detection as a regression problem. SSD [20] uses multiscale feature mapping on different layers to detect objects. In [21], Lin et al. analyzed the problem

Manuscript received 6 October 2022; revised 20 November 2022 and 13 December 2022; accepted 14 December 2022. Date of current version 9 January 2023. This work was supported in part by the Key Research and Development Program of ShaanXi Province under Grant 2020GY-060, in part by the Xi'an Science and Technology Project under Grant 2020KJRC0126 and Grant 202012, and in part by the National Key Research and Development Program of China under Grant 2020YFC1512002. (*Corresponding author: Zhen Ye.*)

The authors are with the School of Electronics and Control Engineering, Chang'an University, Xi'an 710064, China (e-mail: linbai@chd.edu.cn; 2020232038@chd.edu.cn; yezhen525@126.com; 2020232090@chd.edu.cn; 2020132070@chd.edu.cn; ximeng@chd.edu.cn).

The source code of this method can be found at <https://github.com/so-bright/FEPS-Net>.

Digital Object Identifier 10.1109/JSTARS.2022.3230859

of class imbalance comprehensively and proposed a focal loss and a classical one-stage network (RetinaNet), which surpasses the two-stage detector in detection accuracy.

Due to the success of the CNN-based object detection methods, researchers employed them for SAR ship detection. Initially, some studies combined CNN with traditional methods. Kang et al. [22] combine traditional methods with deep learning using the bounding box generated by CNN networks as guard windows for CFAR. An et al. [23] first implemented the sea-land separation with the help of the fully convolutional network and then modeled the distribution of sea clutter. Recently, with the expansion of SAR ship datasets, many deep learning based methods have been proposed, which further improve the detection accuracy and efficiency. To improve the multiscale detection performance, some researchers take the perspective of network structural connection. Jiao et al. [24] used the dense connection to extract features at different scales. Kang et al. [25] fused deep semantic and shallow resolution features by extracting contextual features of the region of interest. Due to a large amount of clutter noises in SAR images, some studies use attention mechanisms to suppress noises. Zhao et al. [26] constructed a feature pyramid by using receptive fields block (RFB) and convolution block attention module (CBAM). To achieve fast detection, Zhang et al. [27] proposed a fast detection network using multiple feature extraction modules and lightweight strategies. Lin et al. [28] proposed SER faster R-CNN, which uses a multiscale feature cascade strategy to improve the quality of shared features and minimize redundant features through SE mechanism and rank modification, thereby improving detection performance. To address the multiscale ship detection problem, Cui et al. [29] proposed a dense attention pyramid network, which combines the salient features with global features to improve detection performance. Besides, Li et al. [30] proposed a multidimensional deep learning network considering the features from spatial domain and frequency domain, respectively. Since the setting of anchors in anchor-based detectors directly affects the detection performance, some recent studies introduced anchor-free based detectors. Fu et al. [31] proposed FBR-Net, which significantly improved the detection performance by feature balancing and refinement. To suppress the interference of noise, Cui et al. [32] introduced CenterNet based on a spatial random grouping attention mechanism. To reduce the occurrence of false alarms, Ma et al. [33] proposed a detection method based on keypoint estimation and attention mechanism to improve detection accuracies of multiscale ships. Hu et al. [34] proposed a novel SAR ship detection network based on feature balance by constructing local attention and nonlocal attention. In ShipDeNet-20 [35], a novel SAR ship detector is built from scratch with only 20 layers, which is lightweight and suitable for hardware transplantation. After the recent appearance of the popular visual transformer (ViT), some transformer models have been introduced into SAR ship detection, considering the advantages of ViT in establishing long-range dependencies. Xia et al. [36] proposed a ViT architecture, called CRTransSar, to enhance context learning by combining transformer and CNN. Li et al. [37] introduced the Swin transformer as the backbone in cascade R-CNN to improve feature extraction ability and

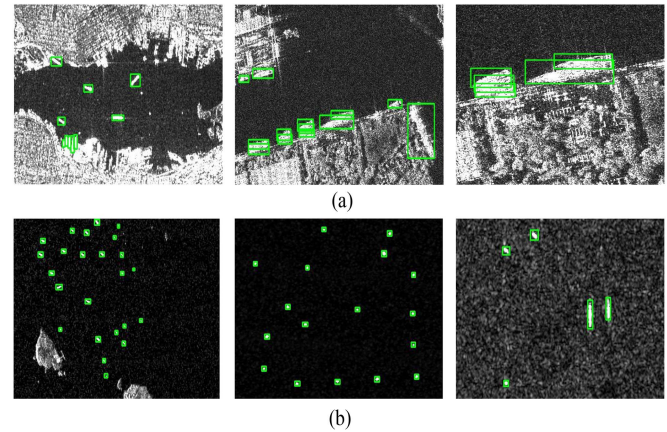


Fig. 1. Some complex samples in SAR ship images. (a) Complex inshore scenes with dense arrangement ships. (b) Offshore scenes with small ships.

proposed a feature fusion module to optimize the feature fusion capability of the feature pyramid. Although the ViT has achieved great success in computer vision, there are still problems, such as high training costs and insufficient performance with less data. However, the objects in remote sensing images (such as ships in SAR images) have the characteristics of multiscales, occlusions, and sparse target distribution, which also make the direct use of the ViT model not as effective as in the field of computer vision.

These methods mentioned above use different technical means to promote the development of SAR ship detection. However, SAR ship detection is still a challenging task due to the complexity of the SAR imaging mechanism. For example, in the inshore scene, the surroundings of ships may produce scattering interference as shown in Fig. 1(a). Similar characteristics exist between ship targets and their surroundings, which will obstruct the acquisition of position information of ship targets and lead to false alarms. More importantly, an amount of small ship targets are spread in SAR ship images, bring more difficulties for accurate target detection as shown in Fig. 1(b). Taking the SSDD [38] dataset as an example, its small ship targets (area $< 32^2$ pixels) account for more than 60% of all ship targets. Due to the different receptive fields and resolutions on different layers of the convolutional network, the small target feature information will decrease with the decrease of feature image resolution. For SAR ship detection, the omission of small ships information will deteriorate the final detection results significantly. To solve the multiscale detection problem, the feature pyramid network (FPN) [39] structures are widely used in CNN-based detection methods. It solves the problem of detecting different scale targets by lateral connectivity and bottom-up feature fusion without additional computation. In addition, multiscale feature fusion oriented methods, such as ASPP [40] and RFB [41], improve the feature representation by expanding the size of the convolution kernel to obtain larger receptive fields and fusing multiscale contexts to obtain fine-grained features. FPN-based methods are different from the above strategies in that training and prediction are performed independently at different feature layers. It fuses deep-level semantic information with shallow-level semantic information through a top-down path step by step. This kind

of approaches are developed recently, such as PANet [42], BiFPN [43], NAS-FPN [44], and many more. Unfortunately, FPN-based methods suffer from the following shortcomings. First, fusion by simple upsampling and lateral connection may lead to some problems, such as spatial information mismatch and feature misalignment. Second, there are semantic gaps among different layers, so fusing the features from different layers directly may reduce the multiscale representation capability. The scattering and blurring characteristics around ship targets in SAR images exacerbate these shortcomings, which confuses ship boundary information and make the ship localization inaccurate. Besides, the presence of lots of small ships poses a great difficulty for SAR ship detection. In the field of object detection, using feature maps with high resolution to obtain richer features of small objects is one of the basic approaches to solve such problems. Fu et al. [45] use deconvolution to get high-resolution feature maps as a way to detect small objects. Jeong et al. [46] fuse features with different scales by deconvolution and pooling techniques. However, the semantic information of the feature maps obtained using these approaches is relatively weak. Since the ship features in SAR images are similar to the surrounding noise features, the ship features with weaker semantics generate more false alarms in this case. To address the above problems and improve the detection accuracies, we propose a feature-enhanced pyramid and shallow feature reconstruction network (FEPS-Net). In FEPS-Net, we construct an enhanced feature pyramid structure that uses a spatial attention mechanism to suppress scattered spots and noises. To solve the feature misalignment problem, a learning offset is used to change the convolutional sampling position in the bottom-up feature fusion process of the pyramid network. Furthermore, a new feature map with rich semantic and feature information is reconstructed through feature fusion strategy to capture the features of small ships. The main contributions of this article are as follows.

- 1) In this article, an FEPS-Net is proposed for SAR ship detection. The network effectively solves the negative impact caused by SAR image scattering noises and significantly improves the detection accuracy for small ships.
- 2) A feature enhancement pyramid (FEP) is designed, in which the spatial enhancement module (SEM) is used to reduce the influence of scattering noises on feature extraction. In the bottom-up feature fusion process, the features after upsampling are aligned by the feature alignment module (FAM) to improve the localization accuracy of ships.
- 3) For effectively detecting ships with different scales, especially small size ships, a shallow feature reconstruction (SFR) module is proposed to reconstruct shallow feature layers in the backbone. The SFR module is beneficial for semantic information extraction and position description, thus significantly improving the ability of detecting ships, especially small ships.

We have conducted extensive experiments on two datasets opened recently, named SSDD [38] and HRSID [47]. Through comparative experiments, the proposed method has better detection performance than several existing methods. The results of ablation experiments demonstrate that the proposed FPE

and SFR modules have distinct contributions to accurate ship detection.

The remainder of this article is organized as follows. The framework is described in Section II. In Section III, the experimental results and analysis are presented. Section IV gives a comprehensive discussion on the experimental results. Finally, Section V summarizes with concluding remarks.

II. METHODOLOGY

In this section, we describe the proposed method and the implementation details of each module. The overall architecture of our method is shown in Fig. 2 and consists of three parts, backbone, neck, and head. First, we present the overall architecture of the network. Next, the FEP structure and the SFR modules are described in detail. Finally, we describe the loss function used in this article.

A. Overall Architecture

We use ResNet-50 [48] as the backbone for feature extraction in FEPS-Net. Multilevel feature maps of C_3 , C_4 , and C_5 are outputs from the last three stages of the ResNet-50, respectively. For accurate detection, we have made two improvements to the neck components. On the one hand, we design an FEP consisting of an SEM and an FAM. The SEM is used to enhance the spatial location information and suppress the background noises, and the FAM is used to mitigate the contextual misalignment problem during feature fusion. On the other hand, we designed an SFR module to accurately detect small ships. In head section, the feature maps output from different feature layers is trained and predicted. Specifically, we used five pyramidal feature layers ($R'_2, P'_3, P'_4, P'_5, P'_6$), where R'_2 is the reconstructed shallow feature layer and P'_6 is obtained by a stride-2 3×3 convolution on C_5 . As FEPS-Net is anchor-based, the anchor points are set by using sliding windows on each point of the feature maps from each layer. Three different anchor boxes with varying area sizes are generated for each point. Each box is further divided into three aspect ratios with a scale of $[0.5, 1, 2]$, which means that *nine* different anchor boxes are generated at each point to cover the object. Also, each anchor will correspond to a one-hot vector of the number of k (number of object species) class categories and a 4-D regression vector. These two sets of prediction vectors are obtained by the classification branch and the regression branch, where the classification branch is used to predict the category probability of each anchor at each position, and the regression branch is used to predict the offset between each anchor and the ground truth at each position. Moreover, since the layers have similar semantics, the feature map parameters of the head at different scales are shared, which is beneficial to the model establishment.

B. Feature Enhancement Pyramid

The FPN [39] is mainly designed to solve the multiscale problem in object detection and improve detection performance by lateral connection and bottom-up feature fusion without increasing computational effort. The bottom-up fusion in FPN

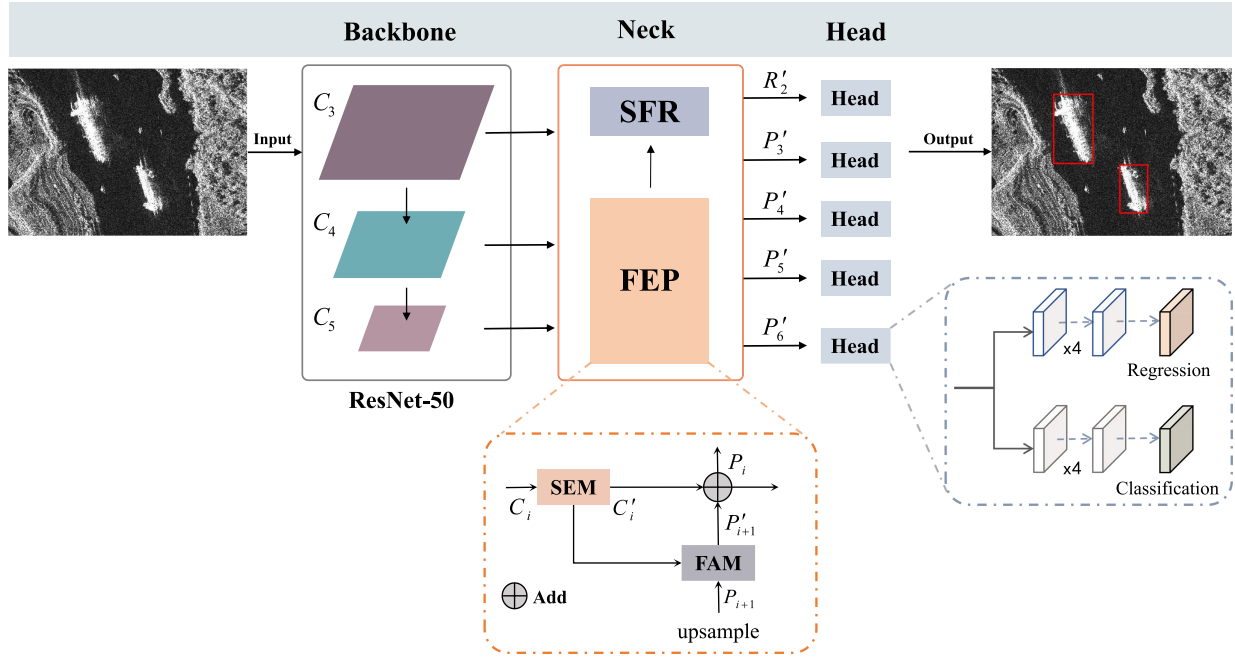


Fig. 2. Overall architecture of FEPS-Net. It mainly includes three parts: 1) feature extraction part with ResNet50 as backbone; 2) the neck part composed of FEP and SFR; and 3) parallel detection head.

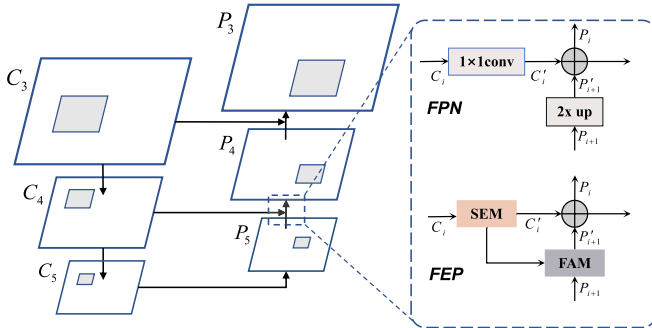


Fig. 3. Comparison of the structure of FPN and FEP.

is an indispensable part of the FPN because deep semantic information is fused with shallow features to obtain a richer feature representation.

However, the bottom-up feature fusion in FPN can result in feature misalignment or spatial information mismatch, which will lead to inaccurate object localization. Moreover, due to the influence of scattering noises in the background of SAR ships, the mixed information in feature maps will aggravate the problem of inaccurate localization and, thus, lead to the wrong prediction. To address this problem, we propose an FEP structure shown as Fig. 3. The FEP structure contains two modules, namely the SEM and the FAM, respectively. The SEM is used for noise suppression around the ship and gives more attention to the features of the ship. The FAM is used to adjust the sampling position of convolutional kernels so that the current feature map is aligned with the upper layer features. Then, deep semantic information is combined with shallow feature

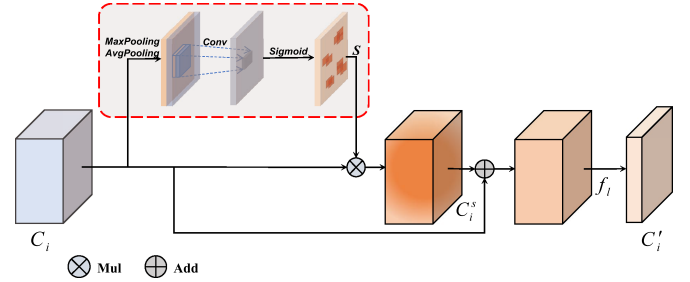


Fig. 4. Structure of SEM.

for feature enhancement by fusing feature maps obtained from SEM and FAM.

1) *Spatial Enhancement Module*: An SEM is explored to suppress the noise interference and highlight the spatial position information of ship targets. The structure of SEM is shown in Fig. 4.

First, given the feature tensor C_i . To extract the maximum and average information of the feature map, maximum pooling and average pooling are performed in the channel dimension, respectively. Then, the spatial position information of ship targets is modeled using sigmoid activation function for convolution with 7×7 kernel size. Finally, the weight map $S \in \mathbb{R}^{H_i \times W_i \times 1}$ is scaled to match channel number of original feature maps C_i to obtain the spatial enhanced feature maps C_i^s . The process can be described as follows:

$$f_s = \sigma(\text{Conv}_{7 \times 7}[\text{MaxPooling}(x); \text{AvgPooling}(x)]) \quad (1)$$

$$S = f_s(C_i) \quad (2)$$

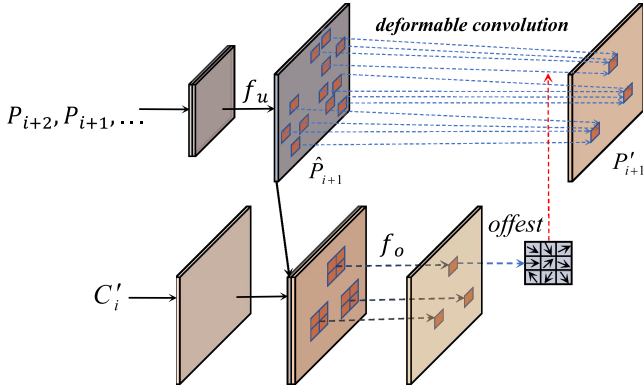


Fig. 5. Structure of FAM.

where MaxPooling and AvgPooling represent the maximum pooling and average pooling, respectively, and σ represents a sigmoid activation function. S represents the attention weight to adjust the feature representation.

Moreover, a skip connection is added between the original input C_i and the enhanced feature maps C_i^s , which prevents the information in the original feature maps from being overamplified or suppressed. The output feature map C'_i is obtained by channel compression. This process can be summarized as

$$C'_i = f_l(C_i + C_i^s) \quad (3)$$

where f_l denotes 1×1 convolution operation. Eventually, the feature map outputs by SEM are used for feature fusion and learning offset, respectively.

2) *Feature Alignment Module*: To alleviate the problems of feature misalignment and spatial mismatch, we propose an FAM. Importantly, we used more informative multilevel features for feature fusion.

As shown in the Fig. 5, the feature map \hat{P}_{i+1} is obtained by upsampling multilevel features. Aiming richer spatial information, \hat{P}_{i+1} is concatenated with shallow feature C'_i following feature alignment. Offsets are obtained using the position information of C'_i and \hat{P}_{i+1} , and each offset value is regarded as the distance between corresponding points of \hat{P}_{i+1} and C'_i . The feature alignment process can be described as

$$\hat{P}_{i+1} = f_u[p_{i+1}, p_{i+2}, \dots] \quad (4)$$

$$\Delta_i = f_o[C'_i, \hat{P}_{i+1}] \quad (5)$$

$$P'_{i+1} = f_{DCN}[\hat{P}_{i+1}, \Delta_i] \quad (6)$$

where the f_u includes upsampling, concatenate, and channel compression operations. The f_o is a convolution operation, where the offset (Δ_i) between \hat{P}_{i+1} and C'_i is learned by convolution. f_{DCN} is employed to align \hat{P}_{i+1} and Δ_i . In the above process, feature alignment functions are implemented by a deformable convolution [49], which can be described as

$$y(p) = \sum_{n=1}^N w_n \cdot x(p + p_n). \quad (7)$$

TABLE I
INFLUENCE OF EACH MODULE IN OUR METHOD FOR SSDD DATASET

FEP	SFR	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
×	×	0.517	0.866	0.543	0.411	0.652	0.709
✓	×	0.569	0.914	0.632	0.499	0.673	0.696
×	✓	0.588	0.952	0.677	0.552	0.651	0.705
✓	✓	0.599	0.960	0.675	0.551	0.682	0.706

TABLE II
ABLATION STUDIES OF FEP

FEP		AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
FAM	SEM						
×	×	0.517	0.866	0.543	0.411	0.652	0.709
✓	×	0.540	0.897	0.580	0.443	0.673	0.689
×	✓	0.545	0.904	0.611	0.476	0.648	0.745
✓	✓	0.569	0.914	0.632	0.499	0.673	0.696

TABLE III
EFFECTS OF SAM AND SEM BY COMPARING WITH THE BASELINE

Method	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
Baseline	0.517	0.866	0.543	0.411	0.652	0.709
+ SAM	0.525	0.892	0.557	0.447	0.644	0.705
+ SEM	0.545	0.904	0.611	0.476	0.648	0.745

TABLE IV
EFFECTS OF C₂ AND SFR BY COMPARING WITH THE BASELINE

Method	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
Baseline	0.517	0.866	0.543	0.411	0.652	0.709
+ C ₂	0.569	0.932	0.625	0.521	0.650	0.675
+ SFR	0.588	0.952	0.677	0.552	0.651	0.705

First, define an input feature map $C_i \in \mathbb{R}^{H_i \times W_i \times C}$ with the output $y(p)$ after convolution, as shown in (7), where N is the sampling number (i.e., N is 9 using a 3×3 convolution kernel). w_n and $p_n \in \{(-1, -1), (-1, 0), \dots, (1, 1)\}$ denote the N_{th} weight and the prespecified offset, respectively.

$$y(p) = \sum_{n=1}^N w_n \cdot x(p + p_n + \Delta p_n). \quad (8)$$

Besides, as in (8), the deformable convolution adaptively learns other offsets $\{\Delta p_n | n = 1, \dots, N\}$ for different sampling positions, where Δp_n is a tuple that can be represented as (h, w) , with $h \in (-H_i, H_i)$ and $w \in (-W_i, W_i)$. Specifically, the position deviation between C'_i and \hat{P}_{i+1} is considered to be the reference of offsets. The deformable convolution uses the obtained offsets to change the sampling position of the convolution kernel and then align \hat{P}_{i+1} .

C. Shallow Feature Reconstruction

Different feature layers contain different semantic and position information for CNN-based object detection networks. More semantic information is contained in deep features, while the position details are present in the shallow feature maps.

TABLE V
DETECTION ACCURACIES OF THE BASELINE AND OURS IN OFFSHORE AND INSHORE SCENES

Scene	Method	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
Offshore	Baseline	0.599	0.962	0.643	0.512	0.712	0.830
	Our	0.645	0.985	0.747	0.604	0.721	0.800
Inshore	Baseline	0.348	0.663	0.340	0.231	0.495	0.590
	Our	0.471	0.858	0.489	0.407	0.559	0.629

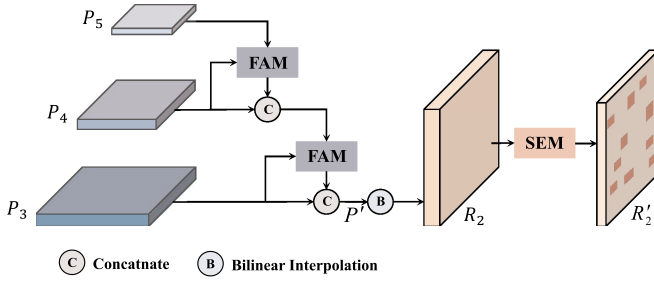


Fig. 6. Structure of SFR.

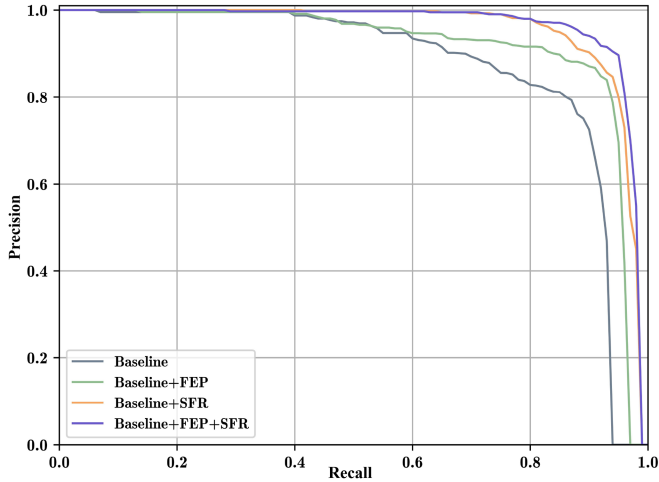


Fig. 7. PR curves of different module improvements.

For most SAR ship images, the available features of small ships are limit or even outright lost, which makes detection very difficult. Although the deep features have stronger semantic information, the positions of small ships are still located hardly. In contrast, the shallow features contain rich position information in favor of small object detection.

The shallow features with high resolution and rich semantic information will contribute to small ship detection for SAR images. Obtaining shallow features directly from backbone network may lead semantic information weak and introduce a large amount of background noises since the backbone network performs limit convolution operations on the original SAR images. The proposed SFR module is shown in Fig. 6. Multilevel features with powerful semantic information are extracted by FEP, and shallow feature maps are reconstructed after feature alignment and spatial information enhancement. The detailed

procedure is as follows. First, P_5 is aligned with P_4 by FAM and concatenated to obtain P'_4 . Then, P_3' is obtained in the same way. Finally, the reconstructed feature map R'_2 is obtained by bilinear interpolation of P'_3 and enhancement of spatial information by SEM. Since the feature maps obtained by upsampling with nearest neighbor interpolation are coarse, we use bilinear interpolation for upsampling to make the feature maps smoother. Since upsampling may bring some noise interference, we use the SEM module after upsampling to reduce the noise interference. The process can be formulated as

$$P'_4 = f_c[P_4, f_{\text{FAM}}(P_5, P_4)] \quad (9)$$

$$P'_3 = f_c[P_3, f_{\text{FAM}}(P'_4, P_3)] \quad (10)$$

$$R'_2 = f_{\text{SEM}}[f_b(P'_3)] \quad (11)$$

where P'_4 and P'_3 represent the transition values in the process of SFR, R'_2 represents the final shallow feature map, and f_c and f_b represent the concatenation and bilinear interpolation operations, respectively. The f_{FAM} and f_{SEM} represent the operations of FAM and SEM, respectively.

After feature fusion by SFR, the reconstructed shallow feature map has stronger abilities of semantic representation and detailed position description, which is more conducive to SAR ship detection.

D. Loss Function

The proposed FEPS-Net is optimized by a multitasking loss, whose function can be described as

$$L = \frac{1}{N_{\text{pos}}} L_{\text{cls}} + \frac{\lambda}{N_{\text{pos}}} L_{\text{reg}} \quad (12)$$

where L_{cls} and L_{reg} are classification loss and regression loss, respectively. N_{pos} is the number of positive samples. Besides, λ is weighting coefficient. To solve the positive and negative sample imbalance problem, focal loss as an improved cross entropy loss is introduced as

$$L_{\text{cls}} = FL(p_t) = -(1 - p_t)^\gamma \log(p_t) \quad (13)$$

where

$$p_t = \begin{cases} p_t, & \text{if } y = 1 \\ 1 - p_t, & \text{otherwise} \end{cases} \quad (14)$$

The regression loss is calculated by using SmoothL1 loss, which can be formulated as

$$\text{Smooth}_{L1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \quad (15)$$

III. EXPERIMENTAL DETAILS

A. Datasets and Settings

We conducted extensive experiments on two datasets: the SSDD and the HRSID. We used the updated version of SSDD released by Zhang et al. [38]. The data were collected from RadarSat-2, TerraSAR-X, and Sentinel-1 sensors with four modes of polarization, such as HH, HV, VV, and VH. The resolution range of the images is from 1 to 15 m, the width

TABLE VI
COMPARISON OF DIFFERENT CNN-BASED METHODS FOR SSDD DATASET

Method	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l	Params(M)	Inference Time(ms)
YOLOv3 [53]	0.487	0.886	0.480	0.443	0.586	0.526	61.52	14.5
RetinaNet [21]	0.517	0.866	0.543	0.411	0.652	0.709	36.10	22.4
SSD300 [20]	0.524	0.919	0.527	0.460	0.625	0.609	23.75	13.9
Faster R-CNN [18]	0.544	0.852	0.611	0.472	0.661	0.674	41.12	21.5
FoveaBox [54]	0.520	0.865	0.586	0.436	0.650	0.620	36.01	23.0
Libra R-CNN [55]	0.547	0.849	0.620	0.464	0.680	0.686	41.39	26.8
Deformable DETR [56]	0.557	0.922	0.616	0.488	0.682	0.682	40.00	40.5
Cascade R-CNN [57]	0.551	0.881	0.603	0.497	0.646	0.619	68.93	32.0
HRSDNet [47]	0.557	0.907	0.603	0.467	0.675	0.725	37.20	47.6
FCOS [58]	0.560	0.919	0.617	0.493	0.671	0.635	31.84	19.8
FBR-Net [31]	-	0.941	0.591	-	-	-	32.50	40.1
FEPS-Net	0.599	0.960	0.675	0.551	0.682	0.706	37.31	31.7

TABLE VII
COMPARISON OF DIFFERENT CNN-BASED METHODS FOR HRSID DATASET

Method	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l	Params(M)	Inference Time(ms)
SSD300 [20]	0.426	0.719	0.457	0.424	0.564	0.133	23.75	13.8
RetinaNet [21]	0.536	0.801	0.594	0.546	0.600	0.221	36.10	39.2
Faster R-CNN [18]	0.560	0.791	0.636	0.566	0.606	0.134	41.12	43.5
FoveaBox [54]	0.549	0.794	0.619	0.557	0.626	0.309	36.01	41.3
Libra R-CNN [55]	0.556	0.774	0.636	0.560	0.614	0.162	41.39	49.5
Deformable DETR [56]	0.458	0.719	0.516	0.457	0.525	0.101	40.00	72.4
HRSDNet [47]	0.573	0.818	0.634	0.583	0.615	0.244	37.20	65.4
YOLOv3 [53]	0.578	0.826	0.654	0.601	0.524	0.047	61.52	14.1
FCOS [58]	0.589	0.870	0.659	0.607	0.634	0.252	31.84	37.3
Cascade R-CNN [57]	0.591	0.805	0.671	0.599	0.632	0.201	68.93	53.6
FEPS-Net	0.657	0.907	0.743	0.668	0.652	0.316	37.31	94.4

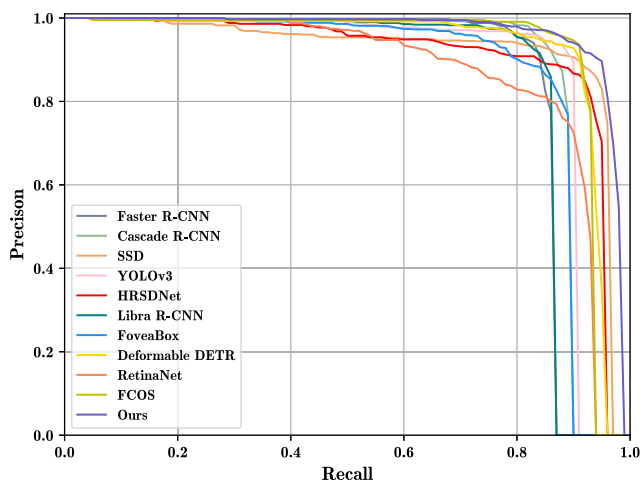


Fig. 8. PR curves of different methods on SSDD.

range of the images is from 214 to 668 pixels, and the height range is from 160 to 526 pixels, where the scale distribution of objects is [1:1, 1:2, 2:1]. In our experiments, the image input

sizes were all resized to 448×448 pixels. The dataset has 1160 images with 2456 ships totally, of which 928 images are used for training and 232 images are used for testing. The HRSID dataset released by Wei et al. [47] was also employed for experiments. The dataset was derived from Sentinel-1B, TerraSAR-X, and TanDEM-X, consisted of 5604 images cropped from multiple large scene images with 800×800 pixels, where 3642 images were used for training and 1962 images were used for testing.

All experiments were implemented by Pytorch and executed on an Nvidia Geforce GTX 2080Ti GPU. To be fair and more convenient for comparison with other methods, we utilized the MMDetection toolbox [50] to unify the experimental benchmarks. We used stochastic gradient descent to optimize the network, training with 50 epochs using 8 images per batch. The initial learning rate was set to 0.001, which was multiplied by 0.1 at the 35th and 45th epochs. The weight decay and momentum are 0.0001 and 0.9, respectively. We selected ResNet-50 [48] pretrained on ImageNet [51] as the backbone of the proposed network and the other parts of the convolutional layers were initialized in the same way as the baseline network.

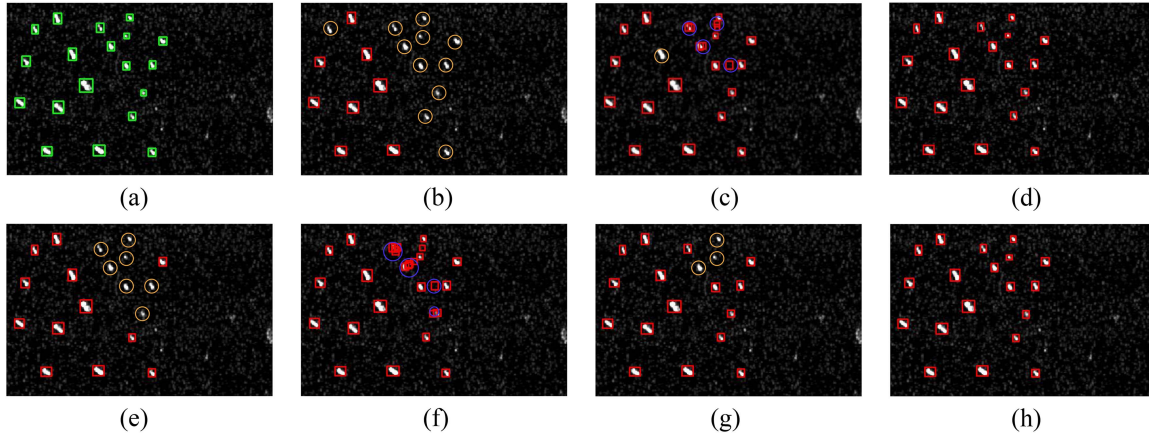


Fig. 9. Comparison of the detection results by different methods for small ships scene of SSDD dataset. The yellow and blue circles in the figure represent missing ships and false alarms, respectively. (a) Ground truth. (b) Result of the Faster R-CNN. (c) Result of the HRSDNet. (d) Result of the YOLOv3. (e) Result of the Cascade R-CNN. (f) Result of the RetinaNet. (g) Result of the FCOS. (h) Our result.

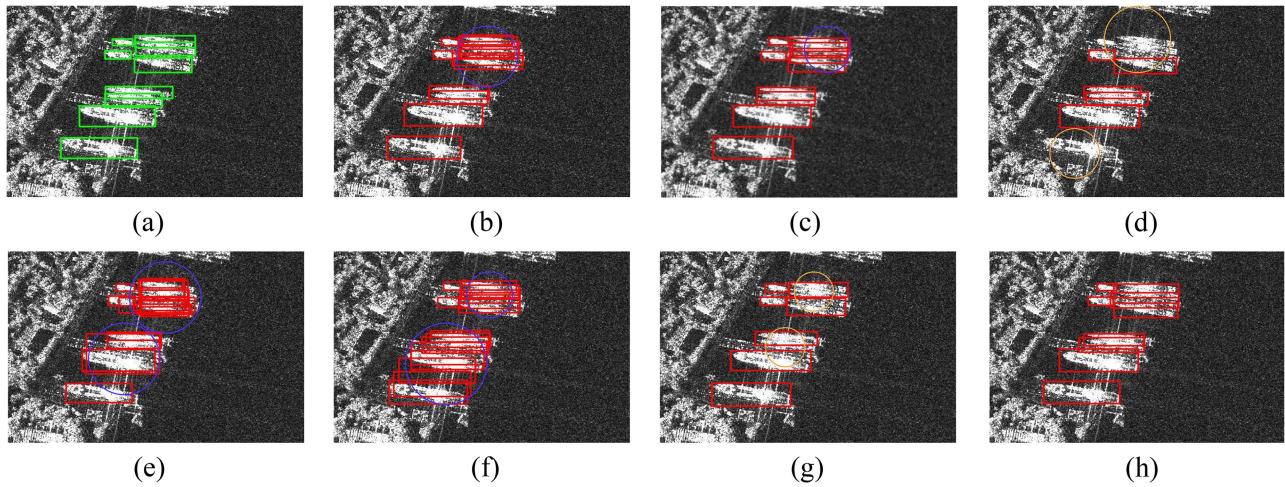


Fig. 10. Comparison of the detection results by different methods for densely arranged ship scene of SSDD dataset. The definitions are the same as those in Fig 11.

B. Evaluation Metric

To evaluate the performance of all methods comprehensively, we used AP, AP₅₀, AP₇₅, AP_s, AP_m, and AP_l as evaluation metrics, which are the same as the metrics defined on the COCO dataset [52]. Here, the AP is average precision as IoU = 0.5 : 0.05 : 0.95; AP₅₀ is the metric for the case of IoU = 0.5; AP₇₅ is the metric for the case of IoU = 0.75, which is a more stringent evaluation metric that better reflects location accuracy. AP_s, AP_m, and AP_l are metrics to evaluate the ability of the proposed network in detecting objects with different sizes. They represent the AP scores for small (area < 32² pixels), medium (32² < area < 64² pixels), and large (area > 64² pixels) objects according to the area occupation defined in parentheses. In experiments, AP₅₀ is primary metric and other metrics are used as references. The precision and recall are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (16)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (17)$$

where TP (true positives), FP (false positives), and FN (false negatives) refer to the number of correct detections, false alarms, and missing targets, respectively. The AP is defined as

$$\text{AP} = \int_0^1 P(R) dR \quad (18)$$

where $P(R)$ is a curve of precision–recall. The AP metric is used to evaluate the comprehensive performance of the model.

C. Evaluations of the Proposed Method

In this section, the effectiveness of the proposed method is evaluated and analyzed, using RetinaNet [21] as the baseline network. At the same time, we conducted experiments in two different scenes (inshore and offshore) and analyzed the detection performance by different metrics.

1) *Ablation Study*: For FEPS-Net, we design two modules (FEP and SFR) to enhance the detection performance of the network. To analyze the effectiveness of each one, we conducted a series of ablation studies with SSDD dataset. To ensure

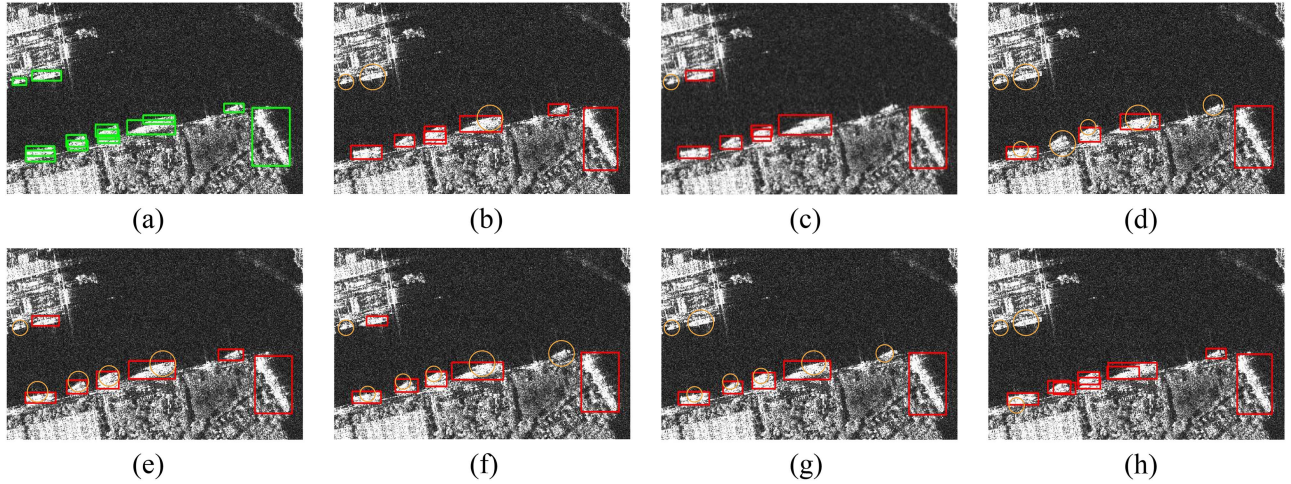


Fig. 11. Comparison of the detection results by different methods for inshore interference scene of SSDD dataset. The definitions are the same as those in Fig. 11.

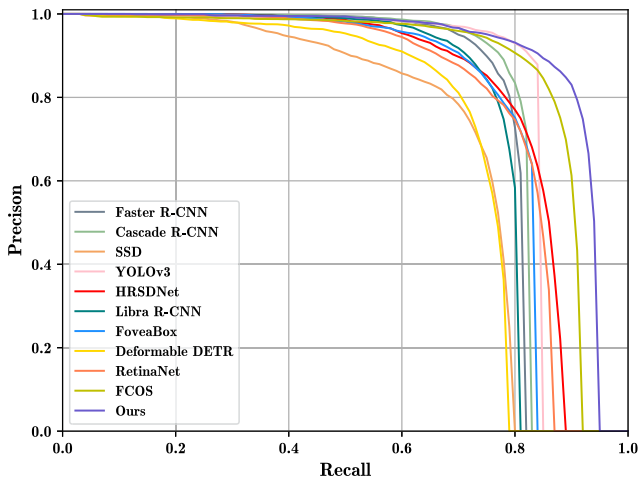


Fig. 12. PR curves of different methods on HRSID.

fairness, the parameters of all methods are set to the same. As overall quantitative comparison, Table I illustrates that both FEP and SFR contribute to the detection accuracies in different degrees. Compared with the baseline network (without FEP and SFR), the proposed network results in gains of 8.2% and 9.4% for AP and $AP_{0.5}$, respectively. Especially for AP_s , the detection accuracy of the proposed approach is 14% higher than baseline network, which means the detection accuracy significantly improved for small ships. Fig. 7 indicates the corresponding precision–recall (PR) curves. The figure provides a more visual indication of the gains from each module. In the following sections, we analyze the effect of each module in detail.

a) Effect of FEP: Since the FEP structure contains two submodules, we further performed an ablation study to analyze the effectiveness of these submodules and gave the experimental results presented in Table II. It can be seen that the detection accuracy of the strategy with SEM module is 2.8%, 3.8%, and 6.8% higher than the strategy without SEM module for AP, AP_{50} , and AP_{75} , respectively. AP_{75} is a more stringent metric for

target localization. From the significant gain brought to AP_{75} , SEM can suppress the noise around the ship and focus on the ship features to improve the localization accuracy. The SEM inspired by the spatial attention module (SAM) in CBAM. The SEM optimizes its structure by adding skip connections between the original input and the enhanced feature maps. It prevents the features with negative impact from being overamplified or the features with positive impact from being oversuppressed. Table III lists the effects of SEM and SAM compared with the baseline method. FAM also brings significant improvements, increasing 2.3%, 3.1%, and 2.1% in AP, AP_{50} , and AP_m , respectively. From the results of AP_m , FAM effectively improves the performance of detecting medium-sized ships, which demonstrates the importance of feature alignment in the bottom-up multiscale feature fusion process of the feature pyramid. Under the combined effect of SEM and FAM, AP, AP_{50} , AP_{75} , AP_s , and AP_m are improved by 5.2%, 4.8%, 8.9%, 8.8%, and 2.1%, respectively. From these results, it can be seen that the FEP consisting of SEM and FAM significantly improves ship detection performance. Especially, the gain of AP_{75} indicates that FEP can locate ship targets more accurately. The accuracies of small and medium ship detection are reflected by AP_s and AP_m , which also benefit obviously by FEP module.

b) Effect of SFR: The extensive distribution of small objects in SAR ship images poses a great challenge to accurate detection. In deep CNNs, shallow features are more beneficial for detecting small ships because shallow features have higher resolution and thus contain more detailed position feature information. To analyze the effectiveness of shallow features, we conduct experiments with the shallow feature map C_2 directly output by backbone and the shallow feature map reconstructed by SFR module, respectively. The detection results are given in Table IV. From the results, the use of shallow features obtained in two different ways (using C_2 and using SFR) improved the AP by 5.2% and 7.1%, respectively, which indicates that shallow features can significantly improve the detection performance of SAR ship images. In particular, for small ship detection, AP_s gains 11% and 14.1% by using C_2 and SFR, respectively.

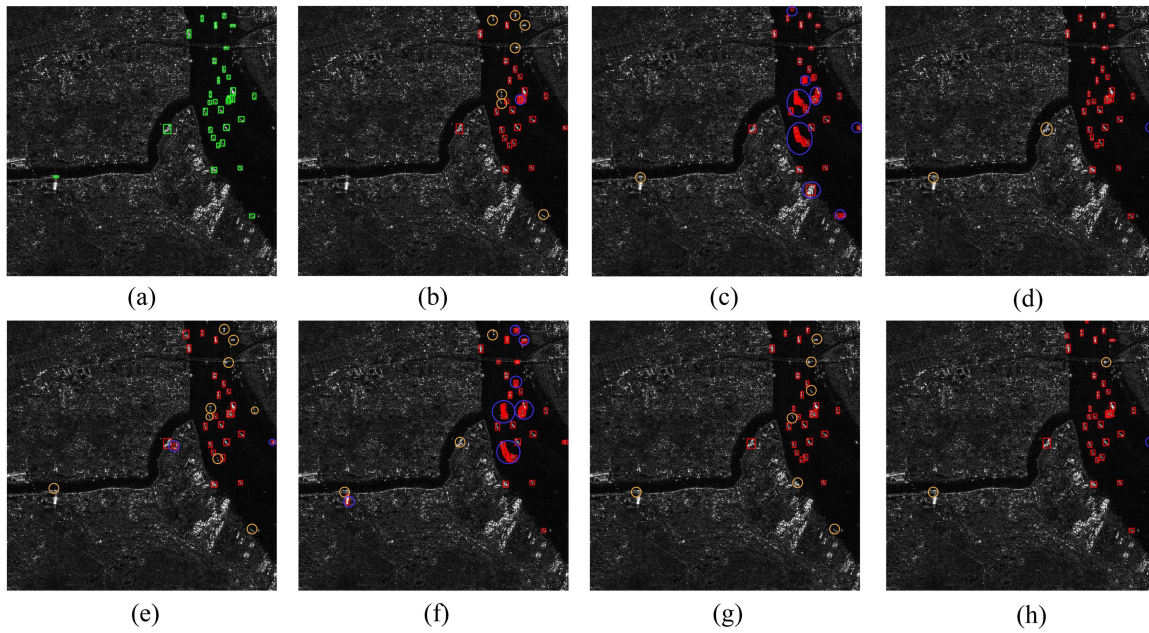


Fig. 13. Comparison of the detection results by different methods for inland complex interference scene of HRSID dataset. The yellow and blue circles in the figure represent missing ships and false alarms, respectively. (a) Ground truth. (b) Result of the Faster R-CNN. (c) Result of the HRSDNet. (d) Result of the YOLOv3. (e) Result of the Cascade R-CNN. (f) Result of the RetinaNet. (g) Result of the FCOS. (h) Our result.

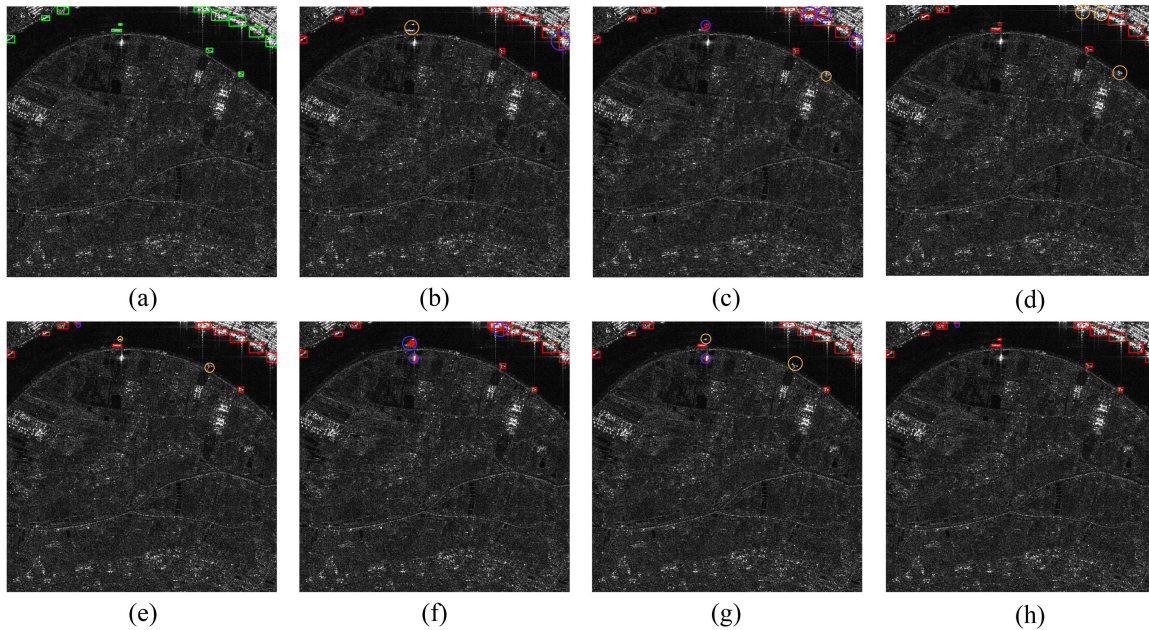


Fig. 14. Comparison of the detection results by different methods for inshore interference scene of HRSID dataset. The definitions are the same as those in Fig 15.

Directly inducing C^2 through the backbone network is the simplest way to obtain shallow feature maps. Unfortunately, the semantic information of the obtained feature maps is weak and accompanied by a large amount of background noise, which may cause false detection. The SFR supplements the semantic information of the shallow features and performs feature alignment and spatial information enhancement during feature fusion, which further enriches the semantic information. According to

Table IV, the detection accuracies using SFR improves 1.9%, 2.0%, 5.2%, and 3.1% over that of using C^2 for AP , AP_{50} , AP_{75} , and AP_s , respectively. This proves that the shallow feature maps obtained by SFR have richer semantic information. Finally, compared to the baseline, SFR obtained significant improvements of 7.1%, 8.6%, and 13.4% for AP , AP_{50} , and AP_{75} , respectively. In particular, AP_s obtains 14.1% improvement in detecting small ships. These results indicate that SFR can significantly improve

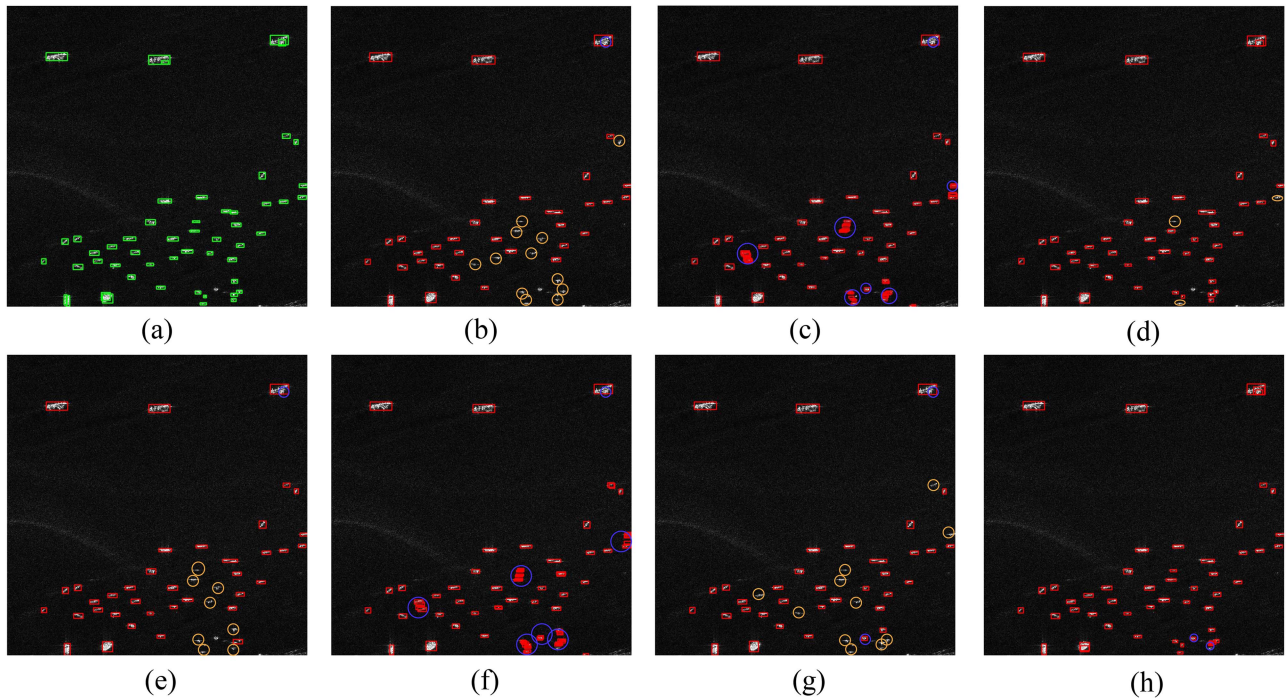


Fig. 15. Comparison of the detection results by different methods for a small ship scene for HRSID dataset. The definitions are the same as those in Fig 15.

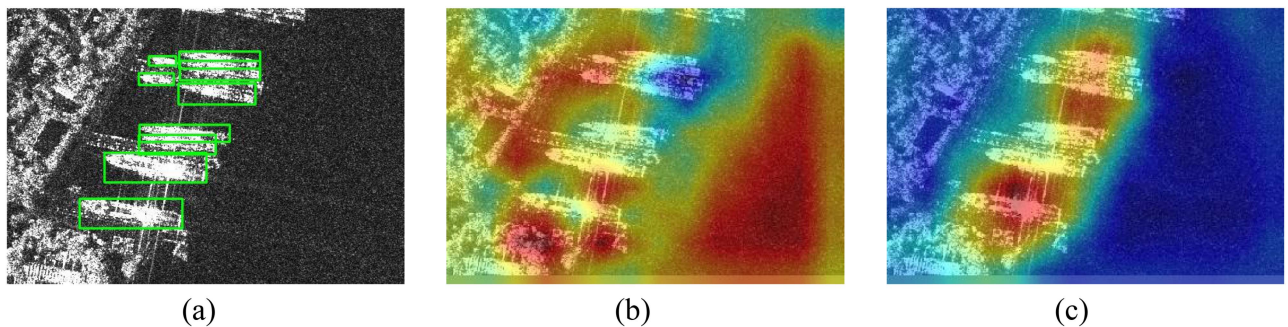


Fig. 16. Some visualization results. (a) Ground truth. (b) Visualization of feature map from FPN. (c) Visualization of feature map from FEP.

the detection performance of the proposed network, especially for small ship detection.

D. Comparison With Other Methods

In this section, the performance of the proposed FEPS-Net and several state-of-the-art methods is evaluated with two typical datasets, including SSDD dataset and HRSID dataset. These existing methods include SSD [20], YOLOv3 [53], Faster R-CNN [18], Cascade R-CNN [57], RetinaNet [21], FoveaBox [54], Libra R-CNN [55], FCOS [58], Deformable DETR [56], HRSDNet [47], and FBR-Net [31]. Table VI lists the detection results of all above methods for SSDD dataset and the corresponding PR curves are shown as Fig. 8 for a more intuitive comparison of the performance of each method. It can be seen from Table VI that the FEPS-Net transcends other methods in AP, AP₅₀, and AP₇₅. In particular, AP₇₅ of ours shows a significant improvement compared to the suboptimal FCOS,

improving by 5.8%, which means that more accurate localization and higher quality prediction boxes can be obtained by ours. Additionally, the proposed method has significant advantages on AP_s, which is mainly benefitted by SFR module for more detailed location description and more robust semantic information. Figs. 9–11 show an intuitive analysis of detection results. The yellow and blue circles in the figures represent missing ships and false alarms, respectively. It can be seen that the FEPS-Net produces fewer omissions and higher quality prediction boxes in the scenes of small ships, dense ship arrangement, and in-shore interference. Besides, some ship targets in Fig. 11 are not detected, which may be due to the strong inshore interference in the scene and the lack of such training samples, so the model cannot learn such features well. In contrast, the performance of other methods for SAR ship detection is not satisfactory in these complex scenes.

To verify the robustness and generalization ability of the proposed method, we conducted comparative experiments on

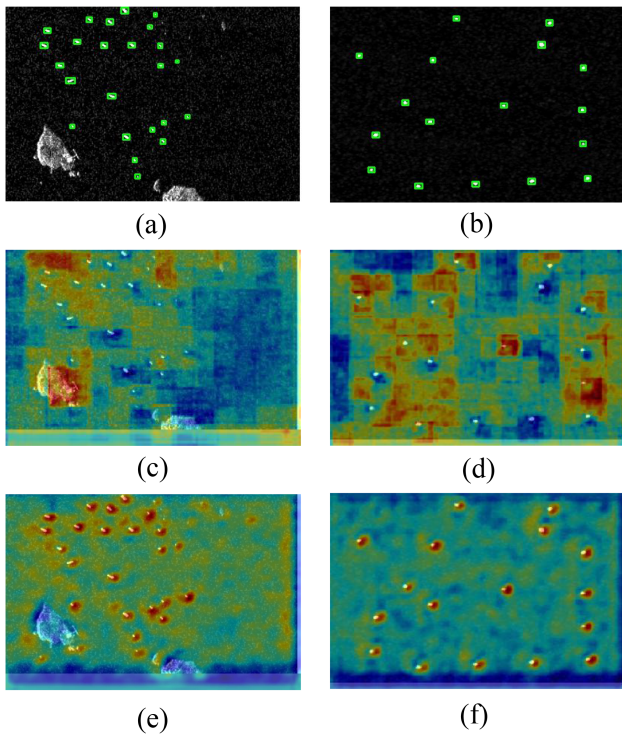


Fig. 17. Some visualization results. (a) and (b) Ground truth. (c) and (d) Results of $C2$. (e) and (f) Results of SFR.

the HRSID dataset. As given in Table VII, the proposed method possesses the best performance in all accuracy metrics compared with the other methods, where AP, AP₅₀, and AP₇₅ are 65.7%, 90.7%, and 74.3%, respectively. In particular, the detection accuracies of the proposed method improve 7.2% and 6.1% for AP₇₅ and AP_s compared with that of the suboptimal Cascade R-CNN and FCOS, which indicates that the proposed method still has the advantages of accurate ship localization and small ship detection in high-resolution scenes. The corresponding PR curves are shown in Fig. 12. As can be seen from the figure, FEPS-Net always has the highest detection precision rate under different recall rates. This means that FEPS-Net detects the largest number of real ships among all detected targets. The detection results of HRSID dataset are shown in Figs. 13–15. In Figs. 13 and 14, land occupies most of the area and contains many interferers whose features are similar to that of ships, making ship detection more difficult. The detection results show that the other methods produce many missed detections and misidentify some land interferers as ship targets. In contrast, FEPS-Net has fewer misses and false alarms in the above cases. In a scenario with a large number of small ships, shown as Fig. 15, other existing methods produce missed detection to varying degrees, while FEPS-Net shows good performance in small ship detection.

IV. DISCUSSION

In this section, we performed additional experiments for different scenes and discussed the experimental results comprehensively. For evaluated detection performance of the proposed FEPS-Net, FEP and SFR were analyzed according to ablation and visualization experiments.

A. Feature Enhancement Pyramid

The proposed FEP is mainly designed to suppress scattering noise in SAR ship images and align contextual features in the feature fusion process, which improves ship localization accuracy. The effectiveness of FEP is fully verified by the experimental results given in Table II. To demonstrate the effect of ship localization intuitively using FEP, the visualized feature maps are shown in Fig. 16. It is obvious that the ship features extracted by FEP are more complete and clearer than by FPN under the interference of inshore strong scattering noise. This is because the FEP structure contains two important components. One of them is SEM, which can effectively reduce the effect of noise and enhance the representation of ship features. Besides, FAM is used to alleviate the problem of inaccurate target localization, and this module plays a key role in the feature fusion process, where aligned features can be used to predict boundaries better. As given in Table II, the method combined SEM with FAM yields outstanding performance for SAR ship detection.

B. Shallow Feature Reconstruction

At present, one of the pressing challenges is how to accurately detect a large number of small ships in SAR ship images. From the analysis of the effectiveness of SFR in ablation study, it is an effective method to detect small ships by shallow features. However, there are a lot of noises and weak semantic information in the shallow features, which makes it difficult to distinguish the real ships and aggravates the false alarm phenomenon. Compared with extracting shallow features directly by the backbone, reconstructing shallow features by SFR will obtain more semantic information because SFR fuses multilevel deep features using feature alignment and spatial enhancement. The semantic information of shallow features directly affects the accuracy of discriminating real ships, which is especially important in the case of complex backgrounds for SAR images. Fig. 17 gives a visualization of the feature maps obtained by the two strategies mentioned above. Fig. 17(c) and (d) shows that more noises appear in the $C2$ feature map, which significantly impacts the detection of small ships. In contrast, the feature maps obtained by the SFR module clearly give more attention to small ships, as shown in Fig. 17(e) and (f).

C. Evaluation in Different Scenes

SAR ship detection involves in the inshore and offshore scenes. In each of these two cases, we evaluated the detection performance of the proposed and baseline methods. In the inshore scene, the land noise interference is relatively large, and the ship targets are more densely distributed. Due to the use of SEM and FAM in FEPS-NET, the noise interference can be suppressed effectively, and the boundary features of ships can be accurately characterized. From Table V, it can be seen that the AP, AP₅₀, and AP₇₅ of the proposed method are improved by 12.3%, 19.5%, and 14.9%, respectively. In the offshore scene, a large number of small ships are distributed sparsely, which increases the difficulty of ship detection. From the AP_s values given in Table V, it can be known that the proposed SFR module is beneficial for small ship detection, which indicates

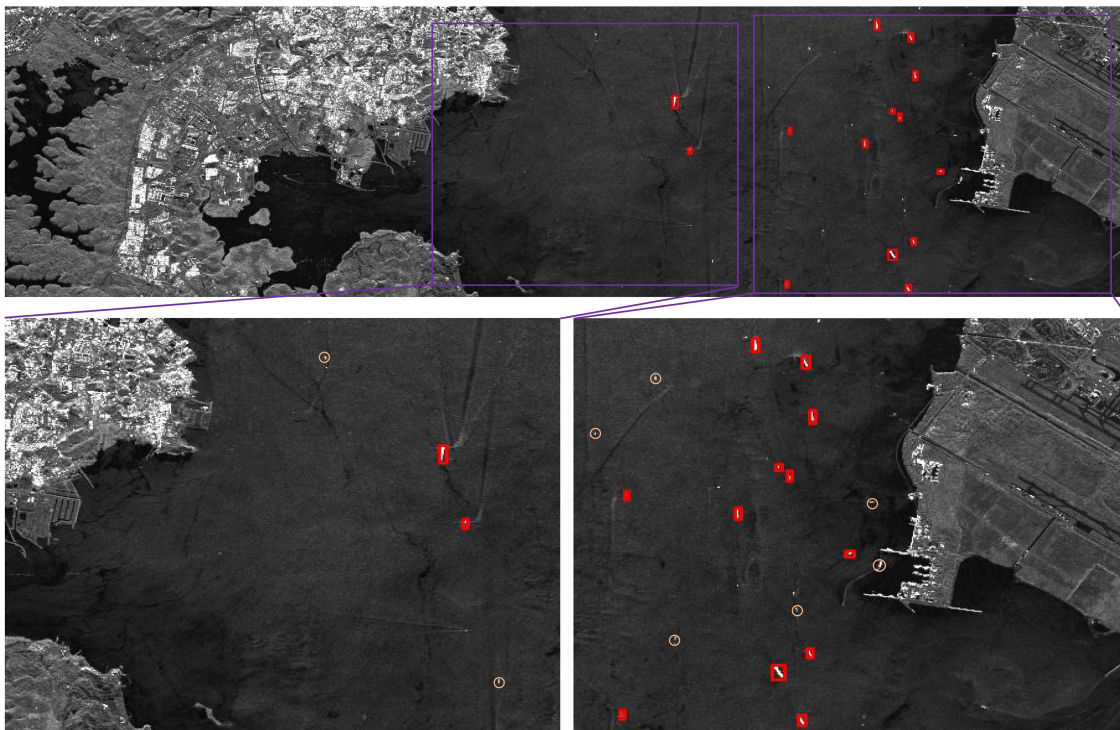


Fig. 18. Detection result of SAR images in large scene. The red box indicates detection results, and the yellow circle indicates missed ship.

that the module can play an active role in semantic information extraction and small object localization.

D. Verification on Large Scene SAR Image

To verify the generalization of our method, we downloaded a large-scene SAR image from TerraSAR-X on the open website [59]. According to the test results in Fig. 18, our method performs well on large-scene SAR images, most of the apparent ship targets in the images were correctly detected, indicating that our method has high generalization performance and can obtain satisfactory results in different SAR scenes.

V. CONCLUSION

In this article, we propose an end-to-end detection network, namely FEPS-Net, to improve the performance of SAR ship detection with significant scattering noise and small ships. In FEPS-Net, we construct an FEP module by embedding spatial attention and feature alignment into the feature pyramid to reduce the effect of scattering noise and improve the ship localization accuracy. In addition, we propose an SFR module for reconstructing shallow features with richer semantic information and position description to improve the detection accuracy of small ships. Compared with other methods, FEPS-Net shows outstanding detection performance (especially for small ship detection) under complex scenes, such as inshore scene with strong noise interference. Due to shallow features with a higher resolution, this may increase the computational costs of the proposed method. In future work, we will further explore lightweight methods for SAR ship detection.

REFERENCES

- [1] E. Nezry, M. Leysen, and G. D. de Grandi, "Speckle and scene spatial statistical estimators for SAR image filtering and texture analysis: Some applications to agriculture, forestry, and point targets detection," in *Synthetic Aperture Radar and Passive Microwave Sensing*, vol. 2584. Bellingham, WA, USA: SPIE, 1995, pp. 110–120.
- [2] A. T. Manninen and L. M. H. Ulander, "Forestry parameter retrieval from texture in CARABAS VHF-band SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 12, pp. 2622–2633, Dec. 2001.
- [3] H. McNairn and B. Brisco, "The application of C-band polarimetric SAR for agriculture: A review," *Can. J. Remote Sens.*, vol. 30, no. 3, pp. 525–542, 2004.
- [4] M. W. Lang and E. S. Kasichke, "Using C-band synthetic aperture radar data to monitor forested wetland hydrology in Maryland's coastal plain, USA," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 2, pp. 535–546, Feb. 2008.
- [5] P. Paillou, S. Lopez, T. Farr, and A. Rosenqvist, "Mapping subsurface geology in Sahara using L-band SAR: First results from the ALOS/PALSAR imaging radar," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 3, no. 4, pp. 632–636, Dec. 2010.
- [6] S.-E. Park, Y. Yamaguchi, and D.-J. Kim, "Polarimetric SAR remote sensing of the 2011 Tohoku earthquake using ALOS/PALSAR," *Remote Sens. Environ.*, vol. 132, pp. 212–220, 2013.
- [7] F. C. Robey, D. R. Fuhrmann, E. J. Kelly, and R. Nitzberg, "A CFAR adaptive matched filter detector," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 28, no. 1, pp. 208–216, Jan. 1992.
- [8] 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.
- [9] H. Dai, L. Du, Y. Wang, and Z. Wang, "A modified CFAR algorithm based on object proposals for ship target detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1925–1929, Dec. 2016.
- [10] T. Li, Z. Liu, R. Xie, and L. Ran, "An improved superpixel-level CFAR detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 184–194, Jan. 2017.
- [11] C. Wang, F. Bi, W. Zhang, and L. Chen, "An intensity-space domain CFAR method for ship detection in HR SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 529–533, Apr. 2017.

- [12] W. Yu, Y. Wang, H. Liu, and J. He, "Superpixel-based CFAR target detection for high-resolution SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 730–734, May 2016.
- [13] G. Gao, L. Liu, L. Zhao, G. Shi, and G. Kuang, "An adaptive and fast CFAR algorithm based on automatic censoring for target detection in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 6, pp. 1685–1697, Jun. 2009.
- [14] 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.
- [15] 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.
- [16] 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.*, 2014, pp. 580–587.
- [17] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.
- [18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
- [20] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
- [21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
- [22] M. Kang, X. Leng, Z. Lin, and K. Ji, "A modified faster R-CNN based on CFAR algorithm for SAR ship detection," in *Proc. IEEE Int. Workshop Remote Sens. with Intell. Process.*, 2017, pp. 1–4.
- [23] Q. An, Z. Pan, and H. You, "Ship detection in Gaofen-3 SAR images based on sea clutter distribution analysis and deep convolutional neural network," *Sensors*, vol. 18, no. 2, 2018, Art. no. 334.
- [24] J. Jiao et al., "A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection," *IEEE Access*, vol. 6, pp. 20881–20892, 2018.
- [25] 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.
- [26] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, 2020.
- [27] T. Zhang, X. Zhang, J. Shi, and S. Wei, "HyperLi-Net: A hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, no. 12, pp. 123–153, 2020.
- [28] 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.
- [29] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Nov. 2019.
- [30] D. Li, Q. Liang, H. Liu, Q. Liu, H. Liu, and G. Liao, "A novel multidimensional domain deep learning network for SAR ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5203213.
- [31] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1331–1344, Feb. 2021.
- [32] 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.
- [33] X. Ma, S. Hou, Y. Wang, J. Wang, and H. Wang, "Multiscale and dense ship detection in SAR images based on key-point estimation and attention mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5221111.
- [34] Q. Hu, S. Hu, and S. Liu, "BANet: A balance attention network for anchor-free ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5222212.
- [35] T. Zhang and X. Zhang, "ShipDeNet-20: An only 20 convolution layers and <1-mb lightweight SAR ship detector," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 7, pp. 1234–1238, Jul. 2021.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.
- [40] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [41] S. Liu et al., "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 385–400.
- [42] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.
- [43] 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. 10 778–10 787.
- [44] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7029–7038.
- [45] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.
- [46] J. Jeong, H. Park, and N. Kwak, "Enhancement of SSD by concatenating feature maps for object detection," 2017, *arXiv:1705.09587*.
- [47] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.
- [48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [49] 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. 9300–9308.
- [50] K. Chen et al., "MMDetection: Open mmlab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
- [51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [52] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [53] J. Redmon and A. Farhadi, "YOLOV3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [54] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyond anchor-based object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020.
- [55] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 821–830.
- [56] 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*.
- [57] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
- [58] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9626–9635.
- [59] Accessed: Nov. 17, 2022. [Online]. Available: <https://www.intelligence-airbusds.com/imagery/sample-imagery/>



Lin Bai (Member, IEEE) received the B.S. degree in electronic information science and technology from Northwest University, Xi'an, China, in 2003, the M.S. degree in electronic science and technology from Xidian University, Xi'an, in 2006, and the Ph.D. degree in signal and information processing from Northwestern Polytechnical University, Xi'an, in 2010.

He is currently an Associate Professor with the School of Electronic and Control Engineering, Chang'an University, Xi'an. His research interests include machine learning and remote sensing image

processing.



Dongling Xue received the B.S. degree from the North University of China, Taiyuan, China, in 2020. She is currently working toward the M.S. degree with the School of Electronics and Control Engineering, Chang'an University, Xi'an, China.

Her research interests include computer vision and remote sensing images change detection.



Cheng Yao received the B.S. degree from Xi'an University, Xi'an, China, in 2020. He is currently working toward the M.S. degree with the School of Electronics and Control Engineering, Chang'an University, Xi'an, China.

His research interests include computer vision and remote sensing images object detection and recognition.



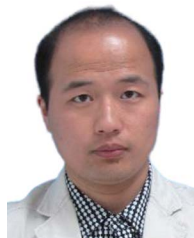
Xiangyuan Lin received the B.S. degree in 2020 from Chang'an University, Xi'an, China, where he is currently working toward the M.S. degree with the School of Electronics and Control Engineering.

His research interests include computer vision and remote sensing images semantic segmentation.



Zhen Ye received the B.S., M.S., and Ph.D. degrees in information and communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2007, 2010, and 2015, respectively.

She spent one year as an Exchange Student with Mississippi State University, Mississippi State, MS, USA. She is currently an Associate Professor with the School of Electronics and Control Engineering, Chang'an University, Xi'an. Her research interests include hyperspectral image analysis, pattern recognition, and machine learning.



Meng Hui (Member, IEEE) received the B.S. and M.S. degrees in mechanical engineering and the Ph.D. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2004, 2007, and 2011, respectively.

He is currently an Associate Professor with the School of Electronic and Control Engineering, Chang'an University, Xi'an. His research interests include the areas of nonlinear dynamics and memristive circuits and corresponding applications.