A Hierarchical Feature Fusion and Attention Network for Automatic Ship Detection From SAR Images

Qianqian Mao, Yinwei Li, *Member, IEEE*, and Yiming Zhu

***Abstract*—Automatic ship target detection technique is a typical and meaningful application for synthetic aperture radar (SAR) image interpretation. Nevertheless, the detection of ship targets within SAR imagery is encumbered by several detracting elements, including obscured outlines, varying dimensions, and elaborate backgrounds, which collectively render the identification process challenging. Existing methodologies for discerning ship targets prove inadequate in effectively navigating these complications. Therefore, we propose a new deep neural network to automatically detect ship target from SAR images, which is named as hierarchical feature fusion and attention network (HFFANet). HFFANet is based on CSPDarknet, the backbone network of YOLOX, and adaptive feature fusion and attention (AFFA) module is innovated to enhance feature extraction. In AFFA, adaptive multilevel feature fusion module is proposed to achieve effective multilevel feature adaptive fusion to better extract target contours and suppress background clutter to reduce false alarms, and enhanced residual coordinate attention module is also proposed to enhance spatial location information and embed it into channel features in the channel layer. The experiments on SAR ship dataset have been conducted, and the mean average precision of HFFANet is 98.53%. Compared with the classical networks, the experimental results show that our model not only achieves the optimal balance in precision and recall, but also achieves the optimal calculation cost.**

***Index Terms*—Attention mechanism, deep learning, multilevel feature fusion, ship target detection, synthetic aperture radar (SAR) image interpretation.**

I. Introduction

SYNTHETIC aperture radar (SAR) is an active Earth observation imaging system that is characterized by high resolution and can effectively identify camouflage and penetrate occlusions [1]. With its unique advantages, SAR images have been widely used for disaster monitoring, environmental monitoring, crop estimation and mapping [2]. Ship detection from SAR images is a research hotspot because of its important applications in both military and civil affairs [3]. However, due to the characteristics of SAR imaging such as long wavelength, discrete target and background noise interference, the features of ships in SAR images are composed of many scattering points, and small ships are easily ignored, which reduces detection rate [4]. In addition, the complex background and surrounding facilities around the ships near the pier are easily identified as the same target, which further increases the difficulty of ship detection in SAR images. Therefore, effective ship target detection methods for SAR image interpretation are crucial [5].

The field of artificial intelligence has become popular in recent years, and the rapid development of artificial intelligence is due to the application of deep learning algorithms [6], which has higher classification accuracy, avoids complex manual extraction of features, and greatly reduces workload, resulting in good generalization performance and good adaptability to complex environments. Nowadays, deep learning-based target detection techniques are divided into two main categories, namely two-stage and one-stage methods. Two-stage methods first obtain the candidate anchor and then classify the candidate anchor to find a more accurate location. The classical algorithms of the two-stage are given in [7], [8], and [9], which usually have higher accuracy but are slower. The one-stage methods do not need to get the candidate anchor and directly generate the class and location coordinate values of object, and the final result can be obtained directly after a single detection. The classical algorithms of one stage are given [10], [11], and [12], which in the past was thought to be faster but less accurate. But with the development of one stage series of algorithms, the accuracy of one stage algorithm has long exceeded that of two-stage algorithm under the condition of fast speed.

Although deep learning technology has also achieved great success in SAR image target detection [13], the ship target in SAR image has some interference factors such as blurred contour, different size, and complex background, which brings unique challenges to the application of deep learning technology in ship target detection from SAR images [14]. Therefore, this article proposes a new end-to-end deep neural network for automatic ship target detection from SAR images. CSPDarknet, the backbone network of YOLOX [12], the first network in the YOLO family that applies the anchor free model, is adopted

Manuscript received 31 March 2024; revised 18 June 2024; accepted 20 July 2024. Date of publication 31 July 2024; date of current version 15 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61988102 and Grant 12105177, in part by the Natural Science Foundation of Shanghai under Grant 21ZR1444300, and in part by Basic Strengthening Program Technical Area Fund under Grant 2023-JCJQ-JJ-0108. *(Corresponding author: Yinwei Li.)*

Qianqian Mao is with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: 212230413@st.usst.edu.cn).

Yinwei Li is with the Terahertz Technology Innovation Research Institute, Terahertz Spectrum and Imaging Technology Cooperative Innovation Center, University of Shanghai for Science and Technology, Shanghai 200093, China, also with the School of Intelligent Emergency Management, University of Shanghai for Science and Technology, Shanghai 200093, China, and also with the Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 200092, China (e-mail: liyw@usst.edu.cn).

Yiming Zhu is with the Terahertz Technology Innovation Research Institute, Terahertz Spectrum and Imaging Technology Cooperative Innovation Center, University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: ymzhu@usst.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3435989

as the baseline network to reduce both the time consumption and the arithmetic power. To solve the problems of complex interference and lack of adaptive feature extraction and fusion methods in SAR images, an adaptive feature fusion and attention (AFFA) module is proposed, which combines bidirectional feature pyramid network (BiFPN) [10], enhanced residual coordinate attention (ERCA) module and adaptive multilevel feature fusion (AMFF) module. This makes the network pay more attention to the important space in the global semantics, improve the detection accuracy and save arithmetic power.

The main contributions of this article are summarized as follows.

1) A new effective end-to-end deep neural network for automatic SAR images ship target detection, namely, hierarchical feature fusion and attention network (HFFANet), is proposed based on adaptive feature fusion and attention. HFFANet achieves fast, automatic, and high-precision ship target detection results through more discriminative features extraction and effective multilevel feature fusion.
2) The AFFA module is developed by combining domain knowledge from SAR multiscale analysis with improved attention mechanism in deep learning. In AFFA, BiFPN fuses high- and low-level features to make the model better fuse features and weight useful information; ERCA is applied to shallow features, which makes the model focus on spatial features of small targets; AMFF performs adaptive multilevel feature fusion to make the model pay more attention to target features and suppress background interference. Therefore, AFFA module can enhance feature extraction and fusion to improve the detection accuracy of multiscale ship targets in complex backgrounds.

The rest of this article is organized as follows. Section II is the problem statement. Section III describes the proposed network in detail. Section IV indicates the experiment results and the performance assessment of the proposed network. Finally, Section V concludes this article.

II. Related Work

In the past decades, many scholars have conducted research on ship target detection using SAR images. In traditional methods, various image features are frequently used. Structural feature-based detection algorithm has superior accuracy and robustness [15], [16], [17]. But its complexity is high and the application scenario is limited. Grayscale feature-based detection algorithm is based on constant false alarm rate [18], [19], which has good performance in weak target environments, but has more false alarm detections. Texture features-based detection algorithm can achieve good performance in single backgrounds [20], [21], [22], [23]. But these methods are sensitive to speckle noise and complex backgrounds and complex preprocessing is usually mandatory, inevitably resulting in low accuracy, high false detection rate, poor robustness, and time-consuming.

Deep learning techniques have been employed for automatic targets detection from SAR imagery [24]. Compared with traditional algorithms, it has the advantage of high accuracy, no manual feature extraction, and better robustness and generalization performance. In deep learning target detection task, the performance of target detection is closely related to the structure of network. Many network improvement methods have been proposed for different application scenarios [25], [26], [27]. The biggest difficulty of ship targets detection in SAR images is that most ships are small in size, and the high background complexity leads to the already small ships being more easily confused with background noise. In addition, there are also many large-scale ship targets in SAR images, and the detection effect will be greatly reduced if only small-scale targets are considered and large-scale targets are ignored. Therefore, in the view of the unique characteristics of SAR images, it becomes quite necessary to investigate the following challenges to enhance the performance of ship target detection in the research domain of SAR image analytics.

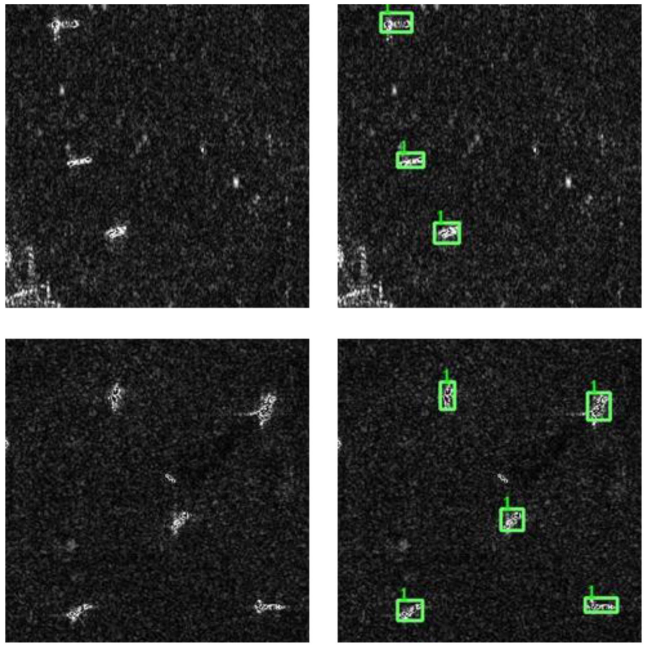

Fig. 1. SAR image slices with small target ships.

A. Challenges of Small Target Ships Detection in SAR Images

SAR image slices of small target ships are shown in Fig. 1. On the left is the original image and on the right is the slice graph labeled with ground truth. As you can see, some ship target sizes in SAR images are very small relative to slices. In deep networks, the characteristic information of small target ships is easily lost after multiple pool processing. Therefore, this feature makes small target ship become the difficulty of SAR ship detection.

B. Challenges of SAR Target Ships Detection With Background Interference

Fig. 2 shows SAR ship image slices with complex background interference. On the left is the original image, and on the right is the image marked with the actual ship frame. As can be seen, the contours of SAR ship targets will be aliased with the surrounding background when coasts or islands appear in the SAR images.

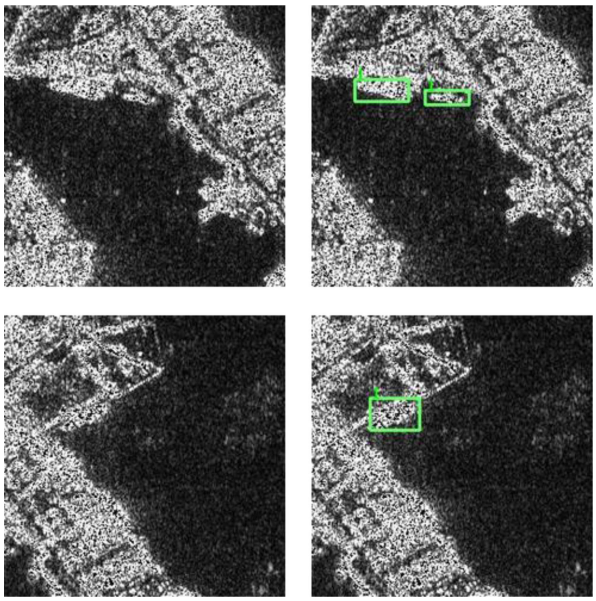

Fig. 2. SAR ship image slices with complex background interference.

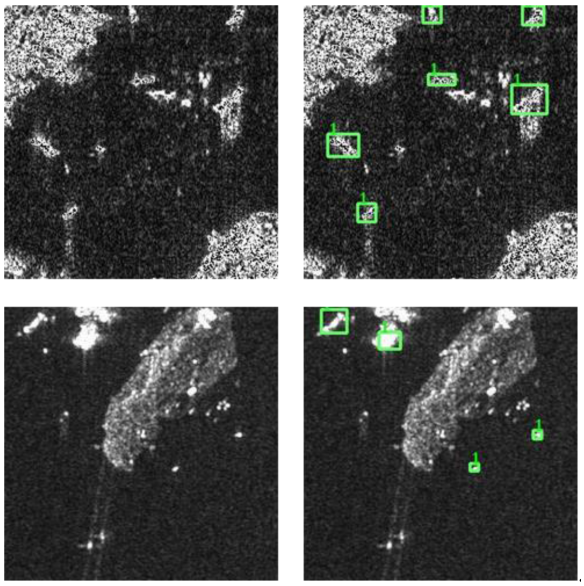

Fig. 3. SAR multiscale target ships with background interference.

This situation will make detection difficulty and easily lead to misjudgment.

C. Challenge of SAR Multiscale Target Ship Detection Under Background Interference

Another type of slice of SAR dataset is multiscale target ships in complex scenes. This type of target is the most difficult to detect and is highly sensitive to interference noise, islands, coasts, etc. And the size of the ships is very different, which has a great influence on the performance of the model. The slice of multiscale SAR ship image under complex background interference is shown in Fig. 3. On the left is the original drawing, and on the right is an image marking the actual frame of the ship. It can be seen that in SAR images with complex background interference, multiscale ship targets are easily confused by background and clutter interference, resulting in false positives and missed detection.

In order to improve the multiscale target detection performance of SAR ship images, it is necessary to retain the feature information of small targets in different scale feature layers and detect the features of large targets well. In [28], a feature pyramid network structure is proposed to achieve multiscale feature fusion, and it is applied to ship target detection [29]. In [30], [31], and [32], the authors introduce the attention mechanism in multi-scale feature fusion target detection algorithm. In [33] and [34], the authors improve SAR multiscale target detection. Chen et al. [35], apply adaptive feature fusion to bridge detection. All these methods improve the multiscale target detection capability of SAR from different angles.

Some researchers have also studied the performance of SAR ship detection under complex background. In [36], a multiresolution SAR target detection algorithm based on region convolutional neural network (R-CNN) under complex backgrounds is proposed. Li et al. [37], propose an anti-jamming model for SAR target detection based on SSD to improve anti-jamming capability. Fu and Wang [38], propose an SSD-based nearshore SAR target detection algorithm, which is effective for a large number of land scenes.

In addition, saving arithmetic power while ensuring network accuracy is also an important consideration of target detection algorithms. Chang et al. [39], propose a real-time target detection system based on YOLOv2, which reduces the computation time while improving the accuracy. Zhang et al. [40], propose an improved algorithm based on YOLOv3, which replaces the backbone feature extraction network with Darknet-19, thus reducing the computational effort. Liu et al. [41], propose the receptive field block (RPF) structure, and introduce extended convolution into the RPF structure to ensure the detection speed of the model.

From the above analysis, it can be seen that the difficulties of ship detection in SAR images are many small-size targets, strong background interference and large ship size difference. While some researchers have begun to investigate these issues, as far as we know, they have only improved one of these factors. At the same time, these improved networks do not involve computational costs. To address these issues, we combine SAR multiscale analysis with an improved attention mechanism in deep learning, called adaptive feature fusion and attention. Under the premise of ensuring the speed of model detection, the framework realizes the high-precision automatic detection of ships from SAR images.

III. Methodology

A. Hierarchical Feature Fusion and Attention Networks (HFFANet)

To realize high-precision SAR ship target detection, an efficient end-to-end target detection framework HFFANet is proposed in this article, as shown in Fig. 4. HFFANet

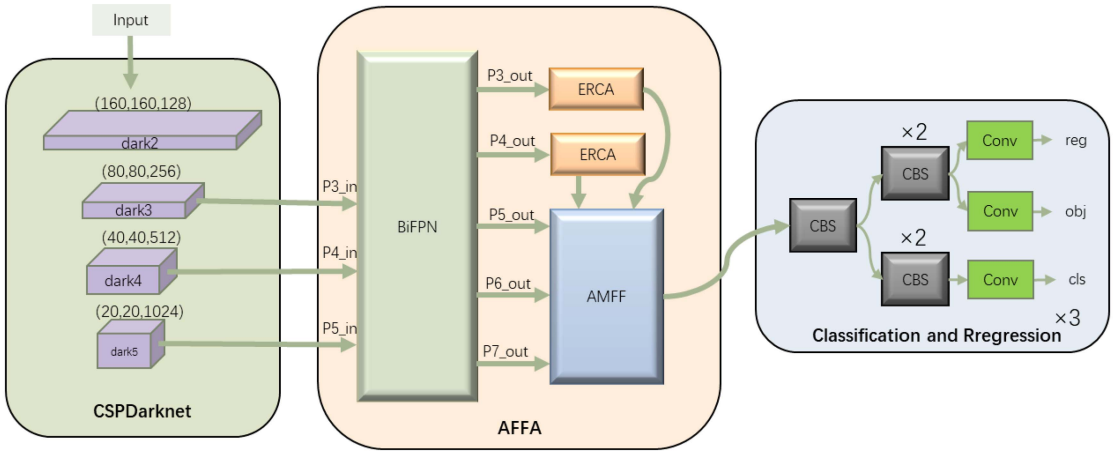

Fig. 4. Framework of HFFANet.

consists of backbone feature extraction network CSPDarknet, enhanced feature extraction network AFFA, and classification and regression module. In the network, CSPDarknet of YOLOX backbone network is adopted as the baseline and AFFA module combines BiFPN and ERCA with AMFF to enhance the spatial information in the extracted high-resolution feature layer and reduce the influence of background features.

The SAR image is input into the backbone network for initial feature extraction, and three feature maps of different sizes are output. Then, these feature maps are fed to AFFA module for enhanced feature extraction and background interference feature suppression. In AFFA, BiFPN is introduced to weight important features layers [10]. ERCA is a coordinated attention mechanism proposed in this article, which combines the coord attention module and residual connectivity approach [42], [43]. It makes the network pay more attention to the spatial position of target point while avoiding overfitting, and can locate the target position in the image more accurately. By using it to suppress the interference of background scattering information in shallow feature layer, the network makes the network can pay more attention to the details of small targets. Deep features refer to features obtained by further downsampling shallow features. For example, in the CSPDarknet architecture, "dark4" is considered a shallow feature relative to "dark5," but a deep feature relative to "dark3." In addition, it helps the network more accurately identify ship images with small targets and complex backgrounds. AMFF feature pairs can be well combined with shallow and deep features, enabling the network to adaptively choose to focus on important spatial information. The classification and regression module performs classification and regression on the multilevel feature layers of the AMFF output, and finally output the prediction results.

B. CSPDarknet

With the development of deep learning field, many lighter, faster, and better networks have been created for us to choose from and improve. YOLOX cleverly blends the best advances in various fields of target detection with the YOLO family of algorithms, providing significant performance gains while maintaining the YOLO family's typically efficient inference speed. YOLOX uses YOLOv3 and Darknet53 as baseline for mature improvements. The entire backbone of YOLOX is a CSPDarknet composed of residuals, and uses a spatial pyramid pooling (SPP) structure to improve the network's perceptual domain by maximizing feature extraction for different pool kernel sizes.

Considering the excellent feature extraction capability of CSPDarknet, we choose CSPDarknet as the baseline. The input image is first subjected to CSPDarknet for initial feature extraction, and the extracted features layer is the feature set of the input image. In CSPDarknet, we obtain three feature layers, namely "dark3," "dark4," and "dark5." These three feature layers are located in the middle layer, lower middle layer, and bottom layer of the main CSPDarknet.

C. Adaptive Feature Fusion and Attention

To enhance and fuse the features extracted from the backbone network effectively, AFFA module is proposed. It consists of three components: BiFPN, ERCA, and AMFF. The input features are fused by BiFPN and then the shallow feature layers are processed by ERCA. Finally, AMFF is carried out for BiFPN's deep feature layers and ERCA's output features.

1) Bidirectional Feature Pyramid Network: Each level feature contains distinctive image information. High-level features have a larger perceptual range and stronger semantic feature representation, but the relative resolution is low, and the detail perception ability is poor; low-level features have a smaller perceptual range and weaker semantic representation, but relatively high resolution and strong detail perception ability. Therefore, to improve the accuracy of the network, fusion of high- and low-level features is a good approach. In previous multiscale feature fusion structures, such as the PANet used by YOLOX, the emphasis is often placed on the fusion of features at different scales, ignoring the fact that different scale features actually have different importance. Some features contribute more to the network, while others contribute less. To address this problem,

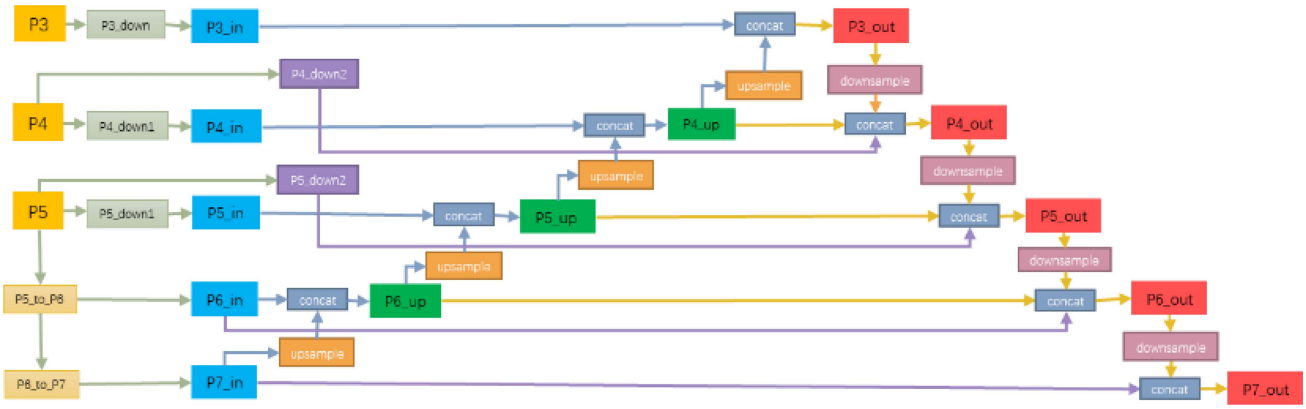

Fig. 5. BiFPN structure.

BiFPN introduces additional weights for features at different scales, enabling the network to learn the importance of each input feature through both top-down and bottom-up multiscale feature fusion. Therefore, we use BiFPN to improve the network's ability to distinguish the importance of feature levels and suppress unwanted information by placing different weights on the feature levels.

BiFPN is a structure that facilitates fast fusion of multiscale features. In AFFA, BiFPN is used only for initial feature fusion. Thus, this article only uses BiFPN once to achieve the expected effect, reducing the amount of computation and improving the speed of feature fusion. Its structure is shown in Fig. 5, where upsampling and downsampling respectively serve to increase and decrease the resolution of feature maps. Upsampling uses interpolation methods, while downsampling uses max pooling methods [44].

The BiFPN structure has three inputs: P3, P4, and P5. These three inputs are "dark3," "dark4," and "dark5," respectively for the features that should be extracted from the backbone. The inputs for P6 and P7 are generated after two downsamples of P5. As shown in Fig. 5, input features are superimposed on the feature layer through bottom-up upsampling, and then superimposed on the feature layer by top-down down-sampling. Because different feature inputs have different resolutions after upsampling or downsampling, they contribute differently to the output. It is necessary to learn by weighting each input feature layer to determine which input features are important and thus which input feature layers are more focused. Therefore, bidirectional cross-scale connection and fast normalization fusion are used. After fast normalization fusion, each normalized weight is between 0 and 1.

2) Enhanced Residual Coordinate Attention: In the output feature layer of BiFPN, different feature layers have different effects on ship target detection. Among them, the shallow feature layer has smaller perception field and is more suitable for acquiring small target features. Therefore, to improve the detection accuracy of small target ships under complex background, we introduce an attention mechanism in the shallow feature layers P3_out and P4_out of BiFPN output.

Squeeze-and-excitation (SE) is a classical channel attention mechanism that enhances important information in channels by weighting different channels using global average pooling and a fully connected layer [45], as shown in Fig. 6(a). SE can improve network performance with less computing burden. However, the global pool establishes interchannel connections by compressing global information, causing the structure to pay too much attention to inter-channel information and lose location information. For the ship target detection task, the importance of location information is self-evident. To solve this problem, inspired by CA mechanism, this article proposes a new ERCA mechanism, which can embed position information into the channel to obtain the position information of SAR ship in complex environment. Its structure is shown in Fig. 6(b).

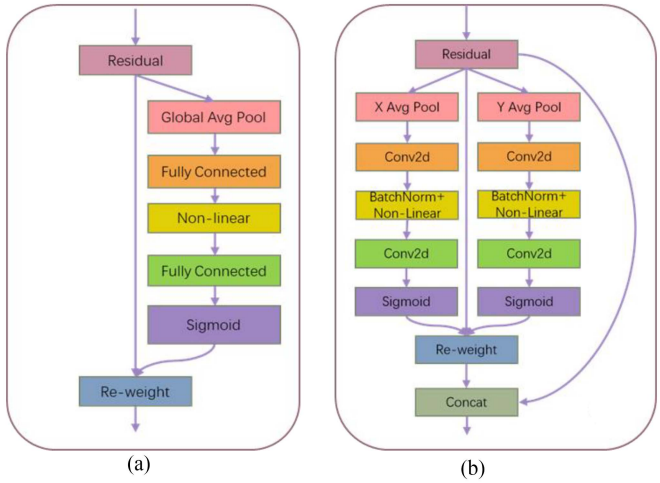

Fig. 6. Structure of attentional mechanisms. (a) SE structure. (b) ERCA structure.

To enhance the spatial location information, the input two-dimensional global pool is first divided into two one-dimensional (1-D) global averaging pools, that is, the height and width of the image are separated and encoded. Two 1-D global averaging pools can extract horizontal and vertical features, respectively. The output feature of BiFPN is x with the dimension of $W \times H \times C$. First, ERCA encodes channel by channel along horizontal and vertical coordinate directions, generating 1-D feature maps in both directions. In this case, 1-D feature maps in two directions

not only have global perceptual field, but also have global feature and position information. The two 1-D global averaging pools are defined as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{1}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j \le H} x_c(j, w) \tag{2}$$

where $z_c^h(h)$ and $z_c^w(w)$ are output feature maps in vertical direction and horizontal direction, respectively.

The obtained bidirectional feature maps not only retain the remote dependence relationship between feature maps, but also retain the accurate position information in the spatial position, which helps the network to predict the target position better. Considering that the background of SAR ship images is complex and the feature maps in different directions are quite different, the convolution of the two directional feature maps is not conducive to the location features acquisition of the model. Unlike the CA mechanism, ERCA has two branches in both directions, processing features in each direction. This structure can better acquire the perceptual field in both directions. For SAR ship target detection, ERCA structure further enhances the features of both directions, emphasizes the position information of ship target, and inhibits the interference of strong scattering features in the background. After convolution, batch normalization and nonlinearization, two 1-D feature maps are obtained as follows:

$$a^h = \delta\left(\mathrm{Bn}\left(F\left(z_c^h(h)\right)\right)\right) \tag{3}$$

$$a^w = \delta\left(\mathrm{Bn}\left(F\left(z_c^w(w)\right)\right)\right) \tag{4}$$

where $F(\cdot)$ is the convolutional layer, $\mathrm{Bn}(\cdot)$ is the batch normalization, and $\delta(\cdot)$ is the nonlinear activation layer.

After activating the results of (3) and (4) by convolution, the enhanced feature map is obtained by multiplying the input feature map $x_c(i, j)$. Finally, inspired by the design of residual module, the residual connection is introduced into ERCA. That is, the final feature map $y_c(i, j)$ is generated by adding the enhanced feature map to the input feature map $x_c(i, j)$. The formula is as follows:

$$y_c(i, j) = x_c(i, j) + x_c(i, j) \times \sigma\left(F\left(a^h\right)\right) \times \sigma\left(F\left(a^w\right)\right). \tag{5}$$

From the above analysis, it can be seen that ERCA is not only simple in structure, but also pays more attention to the details of small target, which can reduce false alarms and missing detection of small ship.

3) Adaptive Multilevel Feature Fusion: BiFPN and ERCA output features solve the problem that small target ships in SAR images are easy to be wrongly detected, but there are still many problems in ship detection under complex background, such as ship size difference. To solve these problems, this article designs the AMFF structure, which consists of multilevel feature fusion and adaptive spatial feature fusion (ASFF), as shown in Fig. 7 AMFF extracts five feature maps of different sizes from ERCA and BiFPN and inputs them into the ASFF module for further processing.

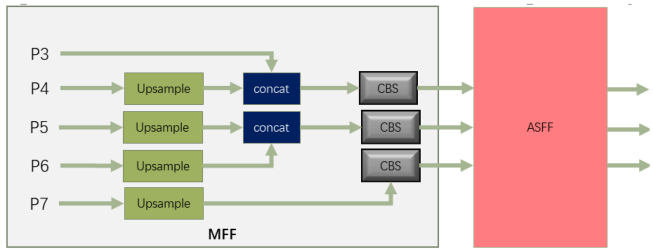

Fig. 7. AMFF module structure.

Fig. 8. CBS structure.

a) Multilevel feature fusion: To suppress the complex background interference, the multilevel features of the input are fused. Since larger scale feature maps can provide more detail for smaller targets, large-scale feature maps are produced by upsampling input feature maps to further improve the detection performance of small targets in complex backgrounds. In addition, in order not to increase computational burden, we use CBS structure to halve the number of channels. Fig. 8 gives the CBS structure, which consists of a convolutional layer, a batch normalization layer and an activation layer.

In the CBS structure, S represents the activation function "SiLU" [46], and the activation functions in the rest of the network structure are the default activation functions of the original network YOLOX. Compared with other classical activation functions, SiLU activation function is better at handling overfitting problems and is suitable for processing complex SAR images. Assuming that the input image is x and the output is SiLU, the calculation formula is as follows:

$$\mathrm{SiLU} = x \cdot \frac{1}{1 + e^{-x}}. \tag{6}$$

As shown in Fig. 7, the five input feature maps are named P3, P4, P5, P6, and P7. In order to obtain large-scale feature maps, P4 is superimposed with shallow feature P3 after double upsampling, P5 and P6 are superimposed after double upsampling, respectively, and P7 is double upsampled. The process of feature superposition is also the process of feature fusion. The superposed features and the up-sampled P7 are respectively input to the CBS module for convolution, batch normalization and activation. The features of CBS processing are the fusion of shallow features and deep features. Among them, the size of the convolution kernel in CBS is 1×1, which is used to adjust the number of channels to facilitate subsequent processing. Finally, they are input into the ASFF module for adaptive feature fusion.

b) Adaptive spatial feature fusion: In general, feature fusion can enhance feature extraction, but it can also bring problems. A target may appear at the same location in different feature layers. In this case, feature fusion is to take a certain feature layer as a positive sample and other feature layers as a negative sample, which is easy to lead to inconsistent positive feature effects. Here, positive samples and negative samples represent correctly identified samples and incorrectly identified

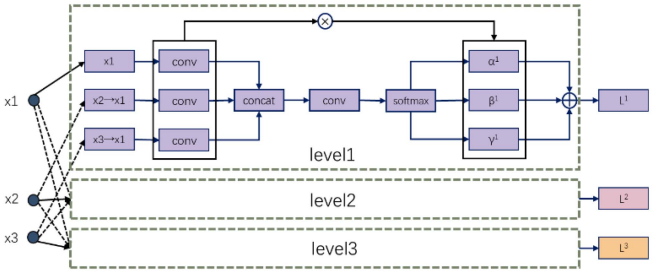

Fig. 9. Structure of ASFF module.

samples, respectively. Especially when the feature map contains objects of different sizes, the inconsistency of positive effects of different scale feature fusion will be further exacerbated because large-size objects are usually associated with small-scale feature maps and small-size objects are associated with large-scale feature maps. To solve the above problem, ASFF is introduced to perform adaptive fusion on the output of multilevel feature fusion. The purpose of ASFF is to make the network pay more attention to the features of different scales, filter the background interference information, and select the useful ship target information. Fig. 9 gives the structure of ASFF module.

D. Classification and Regression

The output feature layer of ASFF is sent into the classification and regression module for target detection. The detection results consist of classification (Cls), confidence (Obj), and regression (Reg). The classification part is called the classification branch, and the confidence and regression part are called the regression branch. In the classification branch, if the confidence value is higher than the threshold, the sample is considered positive and the category is marked. In the regression branch, the distance between each pixel in the ground truth and the four edges of the ground truth is taken as the regression target value for training. Because the size of ships varies greatly, in order to make the network have better generalization performance when detecting ships, it is necessary to use detection modules with different input sizes. In this article, three detection modules with the same structure but different input feature scales are used as regression and classification modules, as shown in Fig. 10.

E. Training

HFFANet is an anchor-free model [47] with much less computational burden of training compared to anchoring models. The loss function is used to compare the predicted and expected outputs of the model and to find the direction of optimization. In addition, to optimize loss function descent, Mini-Batch-SGD is introduced [48].

1) Loss Function: The loss function is used to estimate the difference between the predicted and ground truth of a model and is a non-negative real number. The smaller the loss function, the closer the predicted value is to the ground truth, and the more accurate the model.

The output anchor frames of classification and regression module are used for preliminary screening. For each anchor

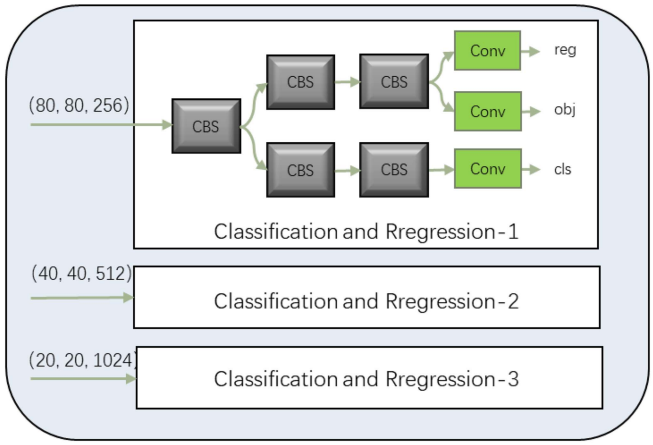

Fig. 10. Classification and regression module.

frame, the IoU is calculated as follows:

$$\text{IoU} = \frac{A \cap B}{A \cup B} \tag{7}$$

where A is the anchor frame, and B is the ground truth.

By setting the IoU threshold, the anchor frame is initially screened and the prediction box is obtained. The output predicted box is then finely filtering using simple optimized transport allocation [49]. For finely filtered prediction box, the loss function is calculated as follows:

$$\text{loss} = l_{\text{Reg}} + l_{\text{Obj}} + l_{\text{Cls}} \tag{8}$$

where l_{Reg} is the regression loss, l_{Obj} is the confidence loss, and l_{Cls} is the classification loss.

2) Gradient Descent: When training the neural network, set a random value for the weight, and then slowly this random weight value is slowly backpropagated later in the training process to make it close to the ideal weights. The gradient descent method is used to update weight in each iteration training. The weight updating formula is as follows:

$$w_{q+1} = w_q - lr \cdot \partial \frac{\text{loss}}{\partial w_q} \tag{9}$$

where w_q is the weight value of the q iteration, w_{q+1} is the weight value of the $q+1$ iteration, and lr is the learning rate.

In addition, the mini-batch-SGD method is used to accelerate the training speed and make the network converge faster [48]. Finally, the weights are obtained until the loss function converges to a local minimum.

IV. Results

A. Datasets

To verify the performance of the proposed algorithm, this article conducts the experiments using multisource and multiscale SAR ship datasets [50]. The dataset contains nearly 40 000 slices of SAR ship image from the Gaofen-3 and ESA Sentinel-1 satellites. Table I shows the SAR parameter information of the dataset. These SAR images are obtained by working with many different modes of SAR sensors, and the obtained images are not

TABLE I
SAR PARAMETER INFORMATION FOR THE DATASET

Sensor	Imaging Mode	Resolution Rg.×Az.(m)	Swath (km)	Incident angle(°)	Polarization	Number of images
GF-3	UFS	3 × 3	30	20–50	Single	12
GF-3	FS1	5 × 5	50	19–50	Dual	10
GF-3	QPSI	8 × 8	30	20–41	Full	5
GF-3	FSII	10 × 10	100	19–50	Dual	15
GF-3	QPSII	25 × 25	40	20–38	Full	5
Sentinel-1	SM	1.7 × 4.3 to 3.6 × 4.9	80	20–45	Dual	49
Sentinel-1	IW	20 × 22	250	29–46	Dual	10

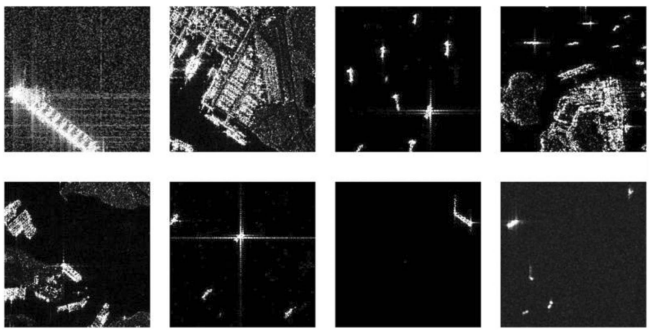

Fig. 11. Sample dataset.

consistent in size and background of the ships. The scenarios in the dataset include ports, offshore, and islands, and target types include common ship targets such as cruise ships, bulk carriers, large container ships and fishing vessels, as shown in Fig. 11. Because the dataset contains almost all imaging scenarios and ships of different types and sizes, it can satisfy the network's verification of generalization performance.

B. Parameter Settings

The experimental environment used in this experiment was built under Windows. All comparison experimental networks were implemented based on the pytorch framework and trained using the NVIDIA GeForce RTX 3090 GPU synchrotron. In order to ensure that the generalization performance of the algorithm is not affected by the data set sequence, 10% of the images are randomly selected as the test set, and the remaining images are divided into the training set and the verification set according to the ratio of 9:1. The initial learning rate (lr) for all networks is set to 0.01 and then automatically decays with training. Set the batch size to 8.

C. Evaluation Metrics

Mean average precision (mAP) is a measure of network model's performance in predicting target locations and categories, which is used here to measure the accuracy of the model. To get the mAP, you need to get the precision and recall

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

where TP is the number of ships correctly detected by the network., FP is the number of false alarm ships, and FN is the number of missing ships.

Here, set a hyperparameter score threshold that directly affects the values of precision and recall. By setting different scoring thresholds, different precision and recall values can be obtained. Generally, as precision increases, recall decreases, and it is crucial to balance precision and recall rates. Therefore, F1 is introduced to measure the model's ability to balance precision and recall. F1 is the harmonic average of precision and recall, calculated by the following formula:

$$F1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{12}$$

The larger the F1, the better the model's balance of precision and recall.

After precision and recall of all scoring thresholds are obtained, precision–recall curves are drawn. The area under the curve is the average precision (AP). The larger the AP, the better the network performance. Since there is only one type of target in the dataset, in this case AP is the mAP we need. For the sake of computational convenience and to save computational resources, we employ the discrete summation method to calculate the mAP. The interval of recall is set to a fixed value of 0.1, and under this condition, the formula for calculating AP is as follows:

$$\text{AP} = \sum_{n} (R_n - R_{n-1}) P_n \tag{13}$$

where R_n and $P_n$are the recall and the maximum precision at the nth recall point, respectively, and $(R_n - R_{n-1})$ represents the interval of recall.

In order to more fully demonstrate our network accuracy while maintaining computational speed, this article introduces floating-point operations (FLOPs), number of parameters (Params) and inference speed (Inference time). FLOPs are the number of floating-point operations performed per second. Params is the total number of parameters that need to be trained in model training to measure the model size. Inference time refers to the time required for a neural network to make a forward propagation. Therefore, comparing these three aspects of the network allows a more complete evaluation of the network's performance.

Since the convolution kernel we are using has the same length and width, we will call them both k. H_{out} and W_{out} are the length and width of the output feature layer, respectively, and $H_{\text{out}} * W_{\text{out}}$ is the size of the output feature layer. C_{in} and C_{out} are the number of channels in the input and output feature layers, respectively, then $k^2 * C_{in} * C_{\text{out}}$ is the number of convolution parameters. Therefore, the formula for calculating FLOPs is as follows:

$$\text{FLOPs} = 2k^2 * H_{\text{out}} * W_{\text{out}} * C_{\text{in}} * C_{\text{out}}. \tag{14}$$

TABLE II
COMPUTATIONAL COST OF EACH NETWORK

Network	FLOPs(M)	Params(M)	Inference time(ms)
Faster-RCNN	1478444.97	136.69	136.77
SSD	486046.93	**23.61**	44.49
YOLOX	199259.32	54.15	38.97
HFFANet	**161408.09**	40.24	**38.13**

The bold values highlight the best performing data in that column.

TABLE III
DETECTION ACCURACY OF EACH MODEL

Network	Precision (%)	Recall (%)	F1	mAP (%)
Faster-RCNN	52.31	**96.05**	0.68	92.37
SSD	94.55	90.38	0.92	96.46
YOLOX	**95.94**	88.77	0.92	97.16
HFFANet	94.88	95.36	**0.95**	**98.53**

The bold values highlight the best performing data in that column.

Params is obtained as follows:

$$\text{Params} = k^2 * C_{\text{in}} * C_{\text{out}}. \tag{15}$$

Inference time is calculated by dividing the forward time of the model by the batch size, i.e.

$$\text{Inference time} = \frac{\text{forward time}}{\text{batch size}}. \tag{16}$$

The forward time is calculated as follows:

$$\text{farward time} = \left(\frac{H_{\text{in}}}{\text{stride}}\right)^2 * C_{\text{in}} * C_{\text{out}} * k * \frac{\text{batch size}}{\text{GPU frequency}} \tag{17}$$

where H_{in} is the height of input image, stride is the step size of model, and GPU frequency is the frequency of GPU.

D. Analysis of the Experiment Results

To demonstrate the effectiveness of our proposed network, we experimentally compare HFFANet with some classical one- or two-stage object detection algorithms, such as Faster-RCNN [8], SSD [11], and YOLOX [12]. To ensure the rigor of the experimental comparison, all networks are implemented using the same environment configuration in the PyCharm software, using the official default parameters and code. In addition, the training set, validation set, test set, training method, and train epochs used in the experiment are all the same. Due to the potential impact of differences in feature detail captured by varying input sizes on the accuracy of the final experiment, we have accordingly adjusted all comparative networks to match the input image size of 256×256 used in our HFFANet model to ensure a fair comparison. Table II shows the computational cost of each network.

It can be seen from Table II that Faster-RCNN, as a two-stage algorithm, has the worst performance in terms of FLOPs, Params, and Inference time. Although SSD have the fewest Params, it has about three times the FLOPs of HFFANet, and the inference time is also slower than the other two one-stage models. While the YOLOX's FLOPs is only about a third of SSD's FLOPs, it is still not as good as the HFFANet. Our HFFANet is optimal in FLOPs and inference time, and the Params is also within the acceptable range. Therefore, our model is optimal in terms of computational cost.

In order to better display the balance and expressiveness of precision and recall, score threshold = 0.5 is usually set to obtain Precision, Recall, and F1, as shown in Table III.

As can be seen from Table III, although the recall of Faster-RCNN reaches 96.05%, its precision is only 52.31%, which makes its F1 only 0.68, and thus its mAP is also the lowest. While the SSD reached 94.55% precision, it does not take recall into account. YOLOX's precision is the highest of the four networks, but recall is only 88.77%. HFFANet strikes the best balance between precision and recall. Therefore, the F1 value of HFFANet is the largest. In addition, the mAP of our network is the highest, which proves that the overall performance of our model is also the best. From the above two tables, it can be seen that our model not only performs best in terms of calculational cost, but also in the model mAP.

To more intuitively show how our network improves under different challenges of ship detection, as described in Section II, we list three sets of images for comparison: small ships under normal background, ships of similar size under complex background, and multiscale ships under complex backgrounds. Three images are selected from each group for comparison. In the following three sets of comparison images, the green rectangular box represents the ground truth, the yellow rectangular box represents the detected object, and the red and orange ellipse represent false and missed alarm, respectively.

1) Analysis of Small Ships Detection Results Under Normal Background: Fig. 12(a) is the labeled SAR images of small ships detection under normal background, including 4, 12, and 6 ships, respectively. Fig. 12(b) shows the ship detection results of Faster-RCNN. As can be seen from Fig. 12(b) that there are 12 false detections, including 8 false alarms and 4 missed ships. It indicates that Faster-RCNN can easily confuse the small ship target with the background in small ships detection. Fig. 12(c) depicts the detection results of ship by SSD.

We can see 15 wrong detections, in which the number of false alarms decreases by 5 to 3 compared with Faster-RCNN, but the number of missed ships increases by 8 to 12. The results indicate that SSD is easy to miss small targets in the detection process, resulting in poor detection effect. Fig. 12(d) demonstrates the detection results of ships by YOLOX. There are four false alarms and two missed ships, which is better than Faster-RCNN and SSD. Fig. 12(e) illustrates the detection results of ships by our proposed network. All ship targets are detected with only three false alarms. Small ships are easily confused with the background, resulting in considerable false alarms and missed ships. It can be seen from the overall detection results that HFFANet is more capable of handling small ship detection than Faster R-CNN, SSD, and YOLOX.

2) Analysis of Similar Size Ship Detection Results Under Complex Background: Fig. 13(a) is the labeled SAR image of ships of similar size under complex background, including

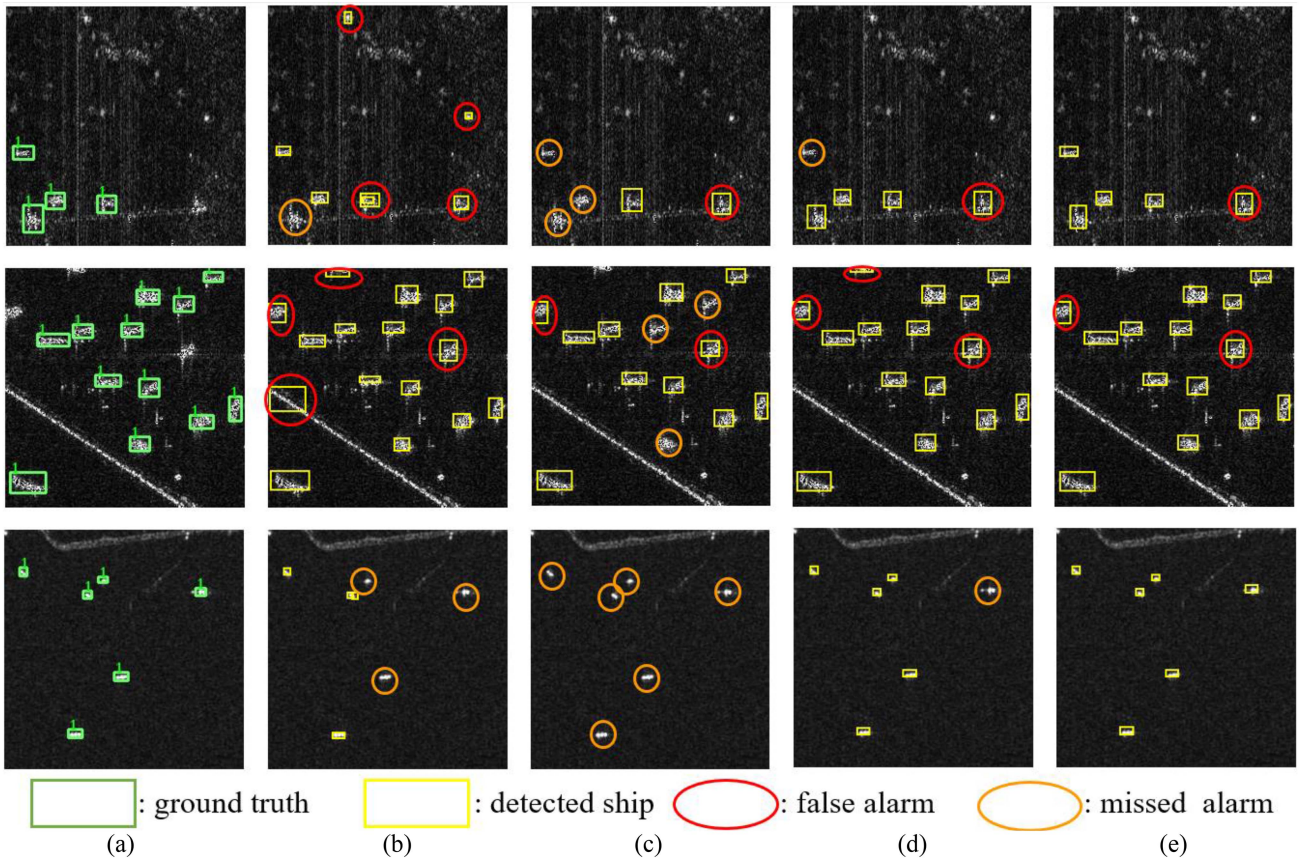

Fig. 12. Labeled SAR images and detected results of small ships under normal background by different networks. (a) Labeled SAR images. (b) Detected results of ships by Faster-RCNN. (c) Detected results of ships by SSD. (d) Detected results of ships by YOLOX. (e) Detected results of ships by HFFANet.

five, one, and four ships, respectively. Fig. 13(b) depicts the ship detection results of Faster-RCNN. There are eight false detections, including six false alarms and two missed ships. Specifically, from the middle image of Fig. 13(b), even if the true ship target is detected, its prediction frame is significantly larger than the ground truth. Fig. 13(c) demonstrates the detection results of ships by SSD. We found that there are one false alarm and one missed ship, which is far superior to Faster-RCNN. In addition, the size of the correctly detected target prediction box in the middle image is also closer to the ground truth. Fig. 13(d) illustrates the detection results of ships by YOLOX. We can see four false alarms and one missed ship, which is better than Faster-RCNN, but worse than SSD. Comparing Fig. 13(b) and (d), we notice that there is one common false alarm and one common missed ship, which is caused by the confusion between the ship and the dock. Fig. 13(e) shows the detection results of ships by HFFANet, from which it can be seen that all ships are detected, with only one false alarm. At the same time, the size of the correctly detected target prediction box in the middle image is also closer to the ground truth. In the ship target detection under complex background, it is easy to mistake the background for ship target, so as to miss the true ship target on the coast. Especially, it can be seen from the overall detection results that slice edges of SAR image are more easily detected by errors. However, HFFANet can distinguish between ships and backgrounds better than the Faster-RCNN, SSD and YOLOX.

3) Analysis of Multiscale Ship Detection Results Under Complex Background: Fig. 14(a) demonstrates the labeled SAR images of multiscale ships under complex backgrounds, including five, seven, and five ships, respectively. Fig. 14(b) shows the ship detection results of Faster-RCNN. We can see twelve false detections, including five false alarms and seven missed ships. Fig. 14(c) depicts the detection results of ship by SSD. Only eleven ships were absent, accounting for 64.70% of all ships. Fig. 14(d) is the ship detection results of YOLOX. We can see three false alarms and three missed ships, which is better than Faster-RCNN and SSD. Fig. 14(e) illustrates the detection results of ships by HFFANet, from which it can be seen that there are two false alarms and two missed ships. According to the overall detection results, Faster-RCNN and SSD are almost ineffective in multi-scale target detection under complex background. Although YOLOX is much better than Faster-RCNN and SSD, it is still worse than our proposed network.

4) Results Comparison Under Different Training Iteration Number: To ensure the rigor of the experimental comparison, the training of all networks in the experiment has no pretraining weight. Because of the difference of networks, different networks will converge under different training iteration number. Generally speaking, when the network reaches a certain of training iteration number, the loss function converges to the minimum value, and the network accuracy does not increase with the increase of the training iteration number. The loss function

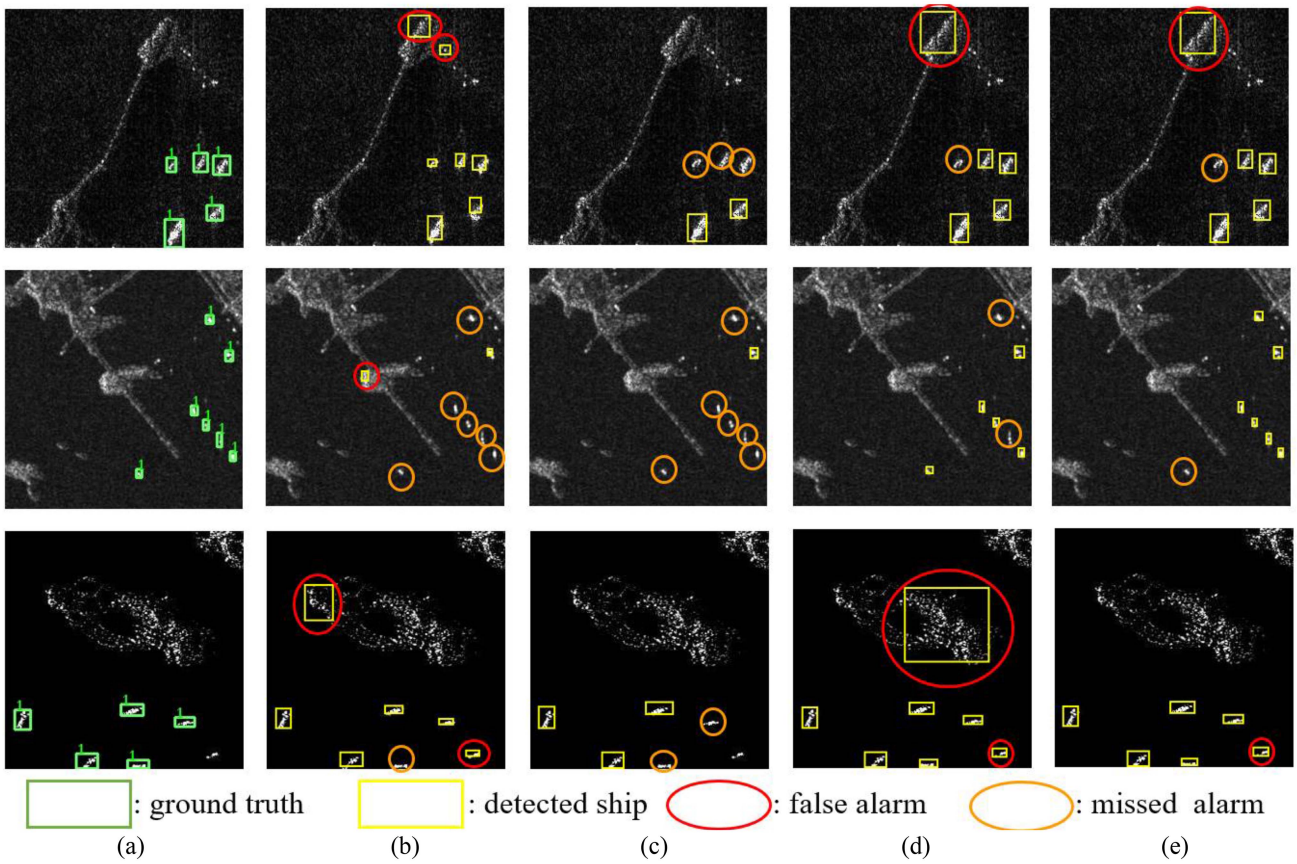

Fig. 13. Labeled SAR images and detected results of similar size ships under complex background by different networks. (a) Labeled SAR images. (b) Detected results of ships by Faster-RCNN. (c) Detected results of ships by SSD. (d) Detected results of ships by YOLOX. (e) Detected results of ships by HFFANet.

of HFFANet converges to its minimum value after 120 training iterations. Therefore, to compare the effect of the number of training iterations on network's mAP, all the networks are trained for 60, 80, 100, and 120 times, respectively.

According to Fig. 15 we can see that Faster-RCNN does not converge at the 60th iteration, with mAP of 82.40%, and the other three networks are basically stable, proving that the convergence rate of Faster-RCNN is the slowest. Faster-RCNN gradually stabilizes after 80 iterations with mAP of 91.22% and reaches 92.37% accuracy after 120 iterations, but it is still far lower than the other three networks. The mAP of SSD is consistent, reaching its peak at the 100th iteration with 96.52% and slightly decreasing to 96.46% at 120 iterations. YOLOX shows a gradual increase, with the highest mAP of 97.19% at the 100th iteration and a slight decrease to 97.16% at the 120th iteration. HFFANet outperforms all with the highest mAP at each iteration milestone, showing 97.05% at 60 epochs and steadily rising to an impressive 98.53% by the 120th epoch.

Therefore, the end result shows that HFFANet not only converges faster but also achieves a higher mAP in fewer iterations. When the training iteration reaches full convergence, our proposed network still performs the best, with the mAP reaching 98.53%, which is 2.01% and 1.34% higher than SSD and YOLOX at the 100th iteration, respectively.

5) Results Comparison Under Different Training Set Percentages: The division of the training set and the validation set also affects the network accuracy. To further demonstrate the robustness of HFFANet under different training set percentages, the dataset is repartitioned and the training set percentages are adjusted to 30%, 50%, 70%, and 90%. At this time, the repartitioned datasets are fed into the networks for testing. The experimental results are shown in Fig. 16. As can be seen from Fig. 16, the mAP of Faster-RCNN increases rapidly with the increase of training sets percentage, which also indicates the low generalization ability of Faster-RCNN. The SSD's mAP is basically stable except for 70% of the training set ratio, but its maximum is only 96.69%. YOLOX's mAP on all training set percentage is basically the same, indicating that YOLOX has good generalization performance, but its maximum value is only 97.53%. HFFANet is based on YOLOX and also inherits its good generalization, with a maximum value of 98.53%, the highest of all networks.

V. Discussion

Compared with Faster-RCNN, SSD, and YOLOX, HFFANet is guaranteed in terms of the number of parameters, the amount of computation, and the speed of inference. HFFANet achieved the best balance between precision and recall, with mAP achieving the highest 98.53%. From the detection results of the three groups of images, it can be seen that Faster-RCNN, SSD, and YOLOX have more false positives and missed positives than HFFANet. In addition, the comparison of results under different training iterations and percentage of training sets shows that

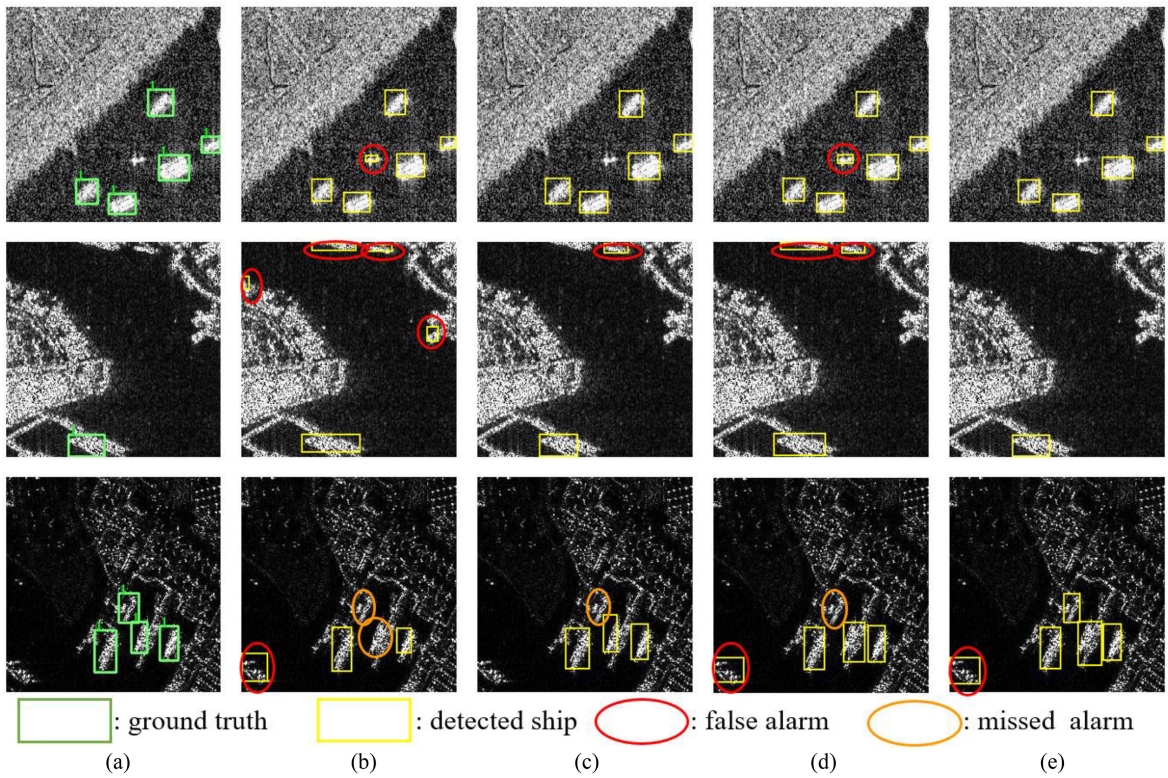

Fig. 14. Labeled SAR images and detected results of multiscale ship under complex backgrounds by different networks. (a) Labeled SAR images. (b) Detected results of ships by Faster-RCNN. (c) Detected results of ships by SSD. (d) Detected results of ships by YOLOX. (e) Detected results of ships by HFFANet.

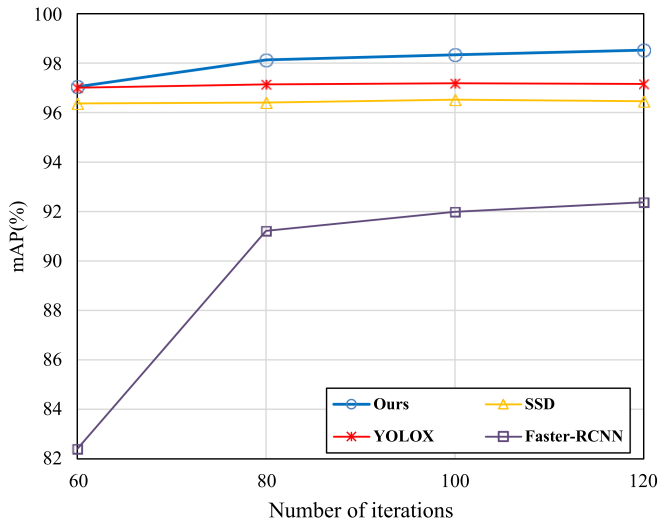

Fig. 15. Network mAP under different number of iterations.

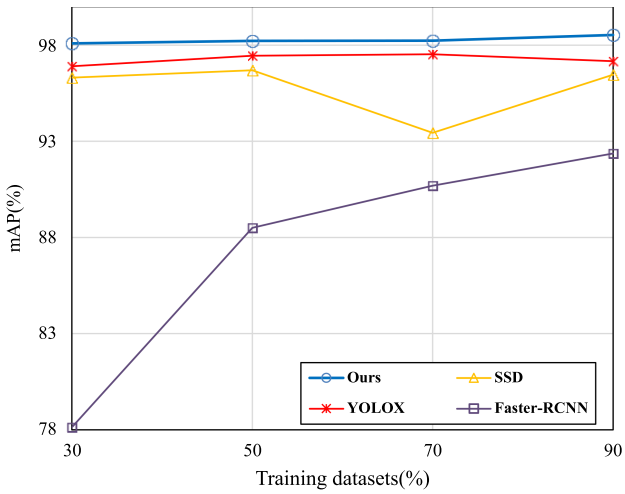

Fig. 16. Network mAP under different training set percentages.

the generalization effect of HFFANet is better than that of the other three networks. Therefore, the experiment proves that our proposed network is successful.

However, in the comparison experiments, we find a phenomenon that is contrary to the common perception. That is, Faster-RCNN, which has the highest computational cost, has the lowest accuracy. The reasons for this phenomenon can be explained from the following two aspects.

First, from the perspective of network structure, Faster-RCNN is a classic two-stage network, while other networks in the comparison experiment are all one-stage networks. For one-stage network, there is no procedure for generating regional proposals. The class probability and position coordinates of the target are directly generated by the feature mapping, and the final detection result can be directly obtained by a single detection. So, one-stage network takes less computation. The two-stage network performs target detection in two stages. In the first stage, regional proposals are obtained. In the second stage, the localization is refined and classified. So, two-stage network is more computationally intensive

Second, the Faster-RCNN series uses only the last layer of the convolutional network. However, the feature mapping of the last layer of convolutional networks is often too small.

This makes subsequent detection and regression unsatisfactory. Even some small objects have no feature points on the final convolution layer. Therefore, Faster-RCNN performs poorly in small target detection.

VI. Conclusion

The biggest difficulty in automatic detection of ship targets from SAR images is that ship scales vary greatly, and most ships are small in size, which makes ship targets more susceptible to background and clutter interference in highly complex environments. To solve these difficulties, HFFANet end-to-end neural network is proposed. HFFANet has good multiscale target automatic detection performance, which is mainly due to two new structures proposed in this article. The combination of ERCA and AMFF structures can effectively extract multiscale features of fusion and suppress background interference.

HFFANet can extract effective features of multiscale ships from SAR images and distinguish between ships and background disturbances, thereby improving feature maps. In addition, because HFFANet is robust against targets with complex background, it can also be used to detect other SAR targets, such as aircraft, buildings, and vehicles. The proposed HFFANet closely combines neural network with SAR image analysis and accelerates the research in the field of SAR intelligent target detection.

References

[1] M. Villano, G. Krieger, K. P. Papathanassiou, and A. Moreira, "Monitoring dynamic processes on the earth's surface using synthetic aperture radar," in *Proc. IEEE Int. Conf. Environ. Eng.*, 2018, pp. 1–5, doi: 10.1109/EE1.2018.8385251.

[2] H. Sportouche, F. Tupin, and L. Denise, "Extraction and three-dimensional reconstruction of isolated buildings in urban scenes from high-resolution optical and SAR spaceborne images," *IEEE Geosci. Remote Sens. Lett.*, vol. 49, no. 10, pp. 3932–3946, Oct. 2011.

[3] C. Wang, S. Jiang, H. Zhang, F. Wu, and B. Zhang, "Ship detection for high-resolution SAR images based on feature analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 119–123, Jan. 2014, doi: 10.1109/LGRS.2013.2248118.

[4] P. Zhuo, X. Zhan, and Y. Wang, "Ship detection in SAR images based on shearlet features," in *Proc. Image Sig. Process. Remote Sens.*, 2018, p. 57, doi: 10.1117/12.2325310.

[5] H. Wang, S. Liu, Y. Lv, and S. Li, "Scattering information fusion network for oriented ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Oct. 2023, Art. no. 4013105, doi: 10.1109/LGRS.2023.3324933.

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

[7] G. Ross, "Fast R-CNN," in *Proc. ECCV*, 2015, pp. 1440–1448.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020, doi: 10.1109/TPAMI.2018.2844175.

[10] 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. 10778–10787.

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

[12] Z. Wang, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," Jul. 2021, doi: 10.48550/arXiv.2107.08430.

[13] A. El Housseini, A. Toumi, and A. Khenchaf, "Deep Learning for target recognition from SAR images," in *Proc. Seminar Detection Sys. Archi. Tech.*, 2017, pp. 1–5, doi: 10.1109/DAT.2017.7889171.

[14] L. Sen, X. Fu, and J. Dong, "Improved ship detection algorithm based on YOLOX for SAR outline enhancement image," *Remote Sens*, vol. 14, no. 16, 2022, doi: 10.3390/rs14164070.

[15] D. Gu and X. Xu, "Multi-feature extraction of ships from SAR images," in *Proc. Int. Cong. Sig. Process.*, 2013, pp. 454–458, doi: 10.1109/CISP.2013.6744037.

[16] Z. Sun, B. Xiong, X. Leng, and K. Ji, "Geometric structure feature extraction of ship target in SAR image based on fine segmentation," *Chin. J. Radio Sci.*, vol. 35, no. 4. pp. 585–593, 2020, doi: 10.13443/j.cjors.2020041402.

[17] Y. Xing and X. Qiu, "Geometric structure feature extraction of ship target in high-resolution SAR image," *J. Sig. Process.*, vol. 32, no. 4, pp. 424–429, 2016.

[18] J. Ai, Z. Cao, Y. Ma, Z. Wang, F. Wang, and J. Jin, "An improved bilateral CFAR ship detection algorithm for SAR image in complex environment," *J. Radars*, vol. 10, no. 4, pp. 499–515, 2021.

[19] J. Chang, Q. Wang, J. Zhao, and N. Li, "An improved 2P-CFAR method for ship detection in SAR images," in *Proc. CIE Int. Conf. Radar (Radar)*, 2021, pp. 1274–1277.

[20] B. Cheng, F. Hu, and R. Yang, "Study on target detection of SAR image using improved fractal feature," *J. Elect. Syst. Inf. Technol.*, vol. 31, no. 1, pp. 164–168, 2009.

[21] D. Charalampidis and T. Kasparis, "Wavelet-based rotational invariant roughness features for texture classification and segmentation," *IEEE Trans. Image Process.*, vol. 11, no. 8, pp. 825–837, Aug. 2002, doi: 10.1109/TIP.2002.801117.

[22] L. M. Kaplan, "Improved SAR target detection via extended fractal features," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 37, no. 2, pp. 436–451, Apr. 2001, doi: 10.1109/7.937460.

[23] G. Stein, E. Zelnio, F. Garber, and D. Charalampidis, "Target detection using an improved fractal scheme," in *Proc. Int. Soc. Opt. Eng.*, 2006, vol. 6237, Art. no. 62370L, doi: 10.1117/12.666598/.

[24] Y. Zhang and Y. Hao, "A survey of SAR image target detection based on convolutional neural networks," *Remote. Sens.*, vol. 14, no. 24, Dec. 2022, Art. no. 6240, doi: 10.3390/rs14246240.

[25] W. Zhu, Y. Zhang, L. Qiu, and X. Fan, "Research on target detection of SAR images based on deep learning," in *Proc. Image Sig. Process. Remote Sens.*, 2018, vol. 10789, Art. no. 1078921, doi: 10.1117/12.2500089.

[26] T. Zheng, J. Wang, and P. Lei, "Deep learning based target detection method with Multi-features in SAR imagery," in *Proc. Asia-Pac. Conf. Synthetic Aperture Radar*, 2019, pp. 1–4.

[27] Z. Geng, Y. Xu, B. Wang, X. Yu, D. Zhu, and G. Zhang, "Target recognition in SAR images by deep learning with training data augmentation," *Sensors*, vol. 23, no. 2, Jan. 2023, Art. no. 941, doi: 10.3390/s23020941.

[28] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Comput. Soc.*, 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[29] P. Chen, Y. Li, H. Zhou, B. Liu, and P. Liu, "Detection of small ship objects using anchor boxes cluster and feature pyramid network model for SAR imagery," *J. Mar. Sci. Eng.*, vol. 8, no. 2, 2020, Art. no. 112, doi: 10.3390/jmse8020112.

[30] L. Chen, P. Zhang, J. Xing, Z. Li, X. Xing, and Z. Yuan, "A Multi-Scale deep neural network for water detection from SAR images in the mountainous areas," *Remote Sens.*, vol. 12, no. 19, 2020, Art. no. 3205, doi: 10.3390/rs12193205.

[31] R. Luo et al., "A fast aircraft detection method for SAR images based on efficient bidirectional path aggregated attention network," *Remote. Sens.*, vol. 13, no. 15, 2021, Art. no. 2940, doi: 10.3390/rs13152940.

[32] T. Zhang, X. Zhang, J. Shi, and S. Wei, "High-speed ship detection in SAR images by improved Yolov3," in *Proc. Int. Comput. Conf. Wav. Act. Media Tech. Infor. Process.*, 2019, pp. 149–152, doi: 10.1109/ICCWAMTIP47768.2019.9067695.

[33] L. Chen, R. Luo, J. Xing, Z. Li, Z. Yuan, and X. Cai, "Geospatial transformer is what you need for aircraft detection in SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5225715, doi: 10.1109/TGRS.2022.3162235.

[34] 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, doi: 10.1109/TGRS.2019.2923988.

[35] L. Chen et al., "Employing deep learning for automatic river bridge detection from SAR images based on adaptively effective feature fusion," *Int. J. Appl. Earth Observation Geoinf.*, vol. 102, 2021, Art. no. 102425.
[36] Q. Xiao et al., "Improved region convolutional neural network for ship detection in multiresolution synthetic aperture radar images," *Concurrency Computation, Pract. Experience*, vol. 32, 2020, Art. no. e5820.
[37] Y. Li, J. Chen, M. Ke, L. Li, Z. Ding, and Y. Wang, "Small targets recognition in SAR ship image based on improved SSD," in *Proc. IEEE Int. Conf. Sig. Infor. Data Process.*, 2019, pp. 1–6, doi: 10.1109/ICSIDP47821.2019.9173411.
[38] X. Fu and Z. Wang, "SAR ship target rapid detection method combined with scene classification in the inshore region," *J. Sig. Process.*, vol. 36, no. 12, pp. 2123–2130, 2020.
[39] Y. Chang, A. Anagaw, L. Chang, Y. Wang, C. Hsiao, and W. Lee, "Ship detection based on YOLOv2 for SAR imagery," *Remote Sens*, vol. 11, no. 7, 2019, Art. no. 786.
[40] T. Zhang, X. Zhang, J. Shi, and S. Wei, "High-speed ship detection in SAR images by improved Yolov3. Conference," in *Proc. 16th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process.*, 2019, pp. 149–152.
[41] S. Liu et al., "Multi-scale ship detection algorithm based on a lightweight neural network for spaceborne SAR images," *Remote Sens*, vol. 14, no. 5, 2022, Art. no. 1149.
[42] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13708–13717.
[43] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Comput. Soc.*, 2017, pp. 6450–6458.
[44] X. Gu, L. Akoglu, and A. Rinaldo, "Statistical analysis of nearest neighbor methods for anomaly detection," Jul. 2019, doi: 10.48550/arXiv.1907.03813.
[45] 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, doi: 10.1109/TPAMI.2019.2913372.
[46] D. Li, H. Yang, C. Yuan, and X. Qin, "Leaf classification utilizing densely connected convolutional networks with a self-gated activation function," in *Proc. Intell. Comput. Methodol.*, 2018, vol. 10956, pp. 370–375.
[47] G. Ding et al., "Anchor-free feature aggregation network for instrument detection in endoscopic surgery," *IEEE Access*, vol. 11, pp. 29464–29473, 2023, doi: 10.1109/ACCESS.2023.3250400.
[48] P. Goyal et al., "Accurate, large minibatch SGD: Training ImageNet in 1 hour," Jun. 2017, doi: 10.48550/arXiv.1706.02677.
[49] Z. Ge, S. Liu, Z. Liu, O. Yoshie, and J. Sun, "OTA: Optimal transport assignment for object detection," in *Proc. IEEE Comput. Soc.*, 2021, pp. 303–312.
[50] 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, 2019, Art. no. 765.

Qianqian Mao received the bachelor's degrees in electronic information engineering from the Tianjin University of Technology, Tianjin, China, in 2020, and the master's degree in communication engineering from the University of Shanghai for Science and Technology, Shanghai, China, in 2024.

Her main research interests include intelligent interpretation of images in remote sensing field, including the detection of stationary targets in radar images and the tracking of dynamic targets.

Yinwei Li (Member, IEEE) received the bachelor's degrees in electronic information engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, and the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2014.

He was with the Shanghai Radio Equipment Research Institute, Shanghai, China, from 2014 to 2019. He is with the Terahertz Technology Innovation Research Institute, Terahertz Spectrum and Imaging Technology Cooperative Innovation Center, University of Shanghai for Science and Technology, Shanghai, China. He is also with the School of Intelligent Emergency Management, University of Shanghai for Science and Technology, and Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai. He is currently an Associate Professor with the University of Shanghai for Science and Technology. His research interests include (inverse) synthetic aperture radar (SAR/ISAR) imaging, interferometric SAR system design and signal processing, terahertz radar imaging, and so on.

Yiming Zhu received the Ph.D. degree in electronic engineering from the University of Tokyo, Tokyo, Japan, in 2008.

He is currently with the Shanghai Key Lab of Modern Optical System and the Terahertz Technology Innovation Research Institute, University of Shanghai for Science and Technology, Shanghai, China, and also with the Terahertz Spectrum and Imaging Technology Cooperative Innovation Center, Shanghai. He is the Director of the Terahertz Spectrum and Imaging Technology Cooperative Innovation Center and the Terahertz Precision Biomedical Technology Overseas Expertise Introduction Center for Discipline Innovation, University of Shanghai for Science and Technology. Up to now, he has published more than 100 articles, including more than 40 articles in the *Light: Science & Applications, the Advanced Optical Materials*, the *Applied Physics Letters*, the *Optics Letters*, and the *Optics Express* (Top 5%), including five articles are selected as ESI articles. His research interests include terahertz technologies and applications, including terahertz devices, terahertz spectroscopy, imaging systems, terahertz bio-applications, and so on.

Dr. Zhu is currently an Award Committee Member of the International Society of Infrared, Millimeter Wave and Terahertz (IRMMW-THz), and a Council Member of the China Instrument and Control Society, the China Optical Engineering Society, and the Young Scientist Club of China Electronics Society. He is also the Topic Editor of the "*Light: Advanced Manufacturing*," "*PhotoniX*," and so on.