A Novel Anchor-Free Model With Salient Feature Fusion Mechanism for Ship Detection in SAR Images

Yunlong Gao, Chuan Wu, and Ming Ren

***Abstract*—Ship detection in synthetic aperture radar (SAR) images has gained great attention in civil and military fields. Anchor-based detection algorithms usually rely on preset candidate boxes, and a large amount of anchor boxes with different sizes will result in a large amount of computing resources being consumed. Recently, anchor-free algorithms have found wide applications in ship detection from SAR images. However, there are still some problems which limit the ship detection performance to a certain extent, such as how to effectively fuse salient features and unbalanced distribution of positive samples. In order to tackle the above problems, we propose a novel anchor-free model named salient feature fusion (SFF)–YOLOX with SFF mechanism. First, we redesign the network of YOLOX to obtain the best balance between detection accuracy and running speed. Second, a saliency region extraction module is introduced to generate the corresponding salient guide map of the input image. Besides, the SFF mechanism is proposed by fusing deep features and salient features to better enhance the discrimination of the multiscale targets. Finally, we improve the SimOTA mechanism by combining the predicted intersection over union (IoUs) and the anchor IoUs to the ground truth bounding boxes to instruct label assignment. We evaluate the detection accuracy and running speed of SFF–YOLOX on the public dataset single shot detector and test the generalization ability on HRSID and two complex large-scale SAR images, and the experimental results prove the model's effectiveness for ship detection task in SAR images.**

***Index Terms*—Label assignment, salient feature fusion (SFF), ship detection, synthetic aperture radar (SAR), YOLOX.**

I. Introduction

SYNTHETIC aperture radar (SAR) [1] is an active microwave remote sensing imaging sensor, which can obtain massive high-resolution and wide-scale remote sensing images. With the continuous improvement of SAR imaging technology, it has been widely applied in all aspects of social and economic life, such as maritime monitoring, traffic control, natural disaster assessment, and environment management [2], [3], [4], [5], [6]. Among these applications, automatic ship detection in remote sensing images has attracted more and more interests because of its important practical value for both civil and military fields [7], [8], [9], [10]. Compared with optical sensors, SAR has the all-day and all-weather surveillance capabilities, making it possible to continuously monitor targets at sea. In recent years, ship detection in SAR images has attracted the attention of scholars, and many investigations that relate to this field have been carried out. Therefore, it is very significant to study the task of ship detection in SAR images.

In the field of object detection in SAR images, extensive studies have been proposed over the years [11], [12], [13], [14], which can be mainly divided into two categories: traditional algorithms and deep-learning algorithms. Traditional algorithms are usually based on statistical distribution analysis of image pixels, and most of that are threshold-based methods [15], [16], and these methods calculate the threshold which distinguishes the ship targets from the backgrounds. The constant false alarm rate (CFAR) [17] is the most classic threshold-based method and it is widely applied in ship detection system nowadays. There are kinds of variants of CFAR algorithm; however, CFAR-based methods require high computational complexity to statically model the ship targets and sea clutters, which is time-consuming in the real-time ship detection [18], [19]. Besides, different shapes, directions of targets, and complex scenarios also limit to instruct a unified statistical model; thus, the generalization ability of these methods is unstable, and the detection performance is barely satisfactory.

Compared with traditional algorithms, deep-learning algorithms are data-driven and do not require prior knowledge, such as the preset threshold and the distributions of sea clutters, which make them more convenient and feasible to be applied in ship detection systems. Nowadays, state-of-the-art deep-learning-based ship detection algorithms consist of one-stage and two-stage detectors. Generally speaking, the two-stage detector is a coarse-to-fine architecture, which mainly focuses on the improvement of detection accuracy, however, they may ignore the importance of running speed. To solve the problem of low detection accuracy, He et al. [20] studied a new approach which applies the Gabor filter to the principle of selective search in fast R-CNN, increasing the number of positive samples in region proposal. Wang et al. [21] utilized the maximum stability extremal region method as the threshold generating strategy to reassess the proposals with high scores in the second stage of faster R-CNN, which greatly reduce the detection errors. Ke et al. [22] boosted the performance of faster R-CNN by using the deformable convolution blocks to better model the geometric transformation of shape changeable ships, achieving a 2.02% accuracy improvement than the baseline network. Sun et al. [23]

Manuscript received 20 July 2023; revised 24 August 2023 and 21 September 2023; accepted 22 September 2023. Date of publication 27 September 2023; date of current version 13 October 2023. This work was supported in part by the National Nature Science Foundation of China under Grant 61401425, and in part by the Key R&D project of Jilin Provincial Science and Technology Development Plan under Grant 2022021146GX. *(Corresponding author: Yunlong Gao.)*

The authors are with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China (e-mail: gaoyl15@mails.jlu.edu.cn; wuchuan0458@163.com; renming5134@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3319831

proposed a two-step detection algorithm based on the coarse-to-fine architecture, which combines the gravitational field and improved mean dichotomy methods to complete precise detection. Kumar and Zhang [24] provided an anchor box optimization method and uses ResNet-50 as the backbone of faster R-CNN to obtain compatible experimental results. The two-stage detectors overcome the shortcomings of traditional algorithms, and realize the automation of ship detection at the same time.

The one-stage detectors pay more attention on how to effectively balance the detection accuracy and the running speed. Up to now, there have been some mainstream one-stage detectors which have been applied in real-time detection tasks. Single shot detector (SSD) [25] performs predictions on multiscale feature maps, which has been the most popular strategy to complete multiscale detection nowadays, and the improvements increase the accuracy of real-time detection. In the meantime, classical YOLO series [26], [27], [28] also achieve a high level in object detection, and the latest YOLOv7 [29] outperforms mainstream real-time detectors in both speed and accuracy on MS COCO [30] dataset. RetinaNet [31] designs the focal loss to tackle the one-stage target detection scenario in which there is an extreme imbalance of the foreground and the background classes during model training, and it is the first time to realize the comprehensive transcendence of one-stage detectors over two-stage detectors. As a matter of fact, these one-stage detectors would be taken as the first choice for real-time detection applications since increasing the running speed is as important as improving the detection accuracy and becomes an important metric of the detection model.

Furthermore, according to whether the anchors are used, deep-learning methods for object detection could be divided into the anchor-based algorithms and the anchor-free algorithms. Anchor-based algorithms first tile a large number of preset anchors on the input image, then predict the category and refine the coordinates of these anchors by one or several times, finally output these refined anchors as detection results. However, these algorithms have some shortcomings. First, all hyperparameters of anchors are preset as prior knowledge, if the detection task changes, the hyperparameters needs to be reset, so the generalization ability is usually low. Second, most candidate boxes are prone to contain backgrounds, only some candidate boxes involve ship targets, which brings about the extremely imbalance of positive and negative samples. Finally, dense candidate boxes are redundant, thereby consuming lots of computing resources. Recently, anchor-free algorithms have become popular because of the proposal of feature pyramid networks (FPNs) [32] and focal loss. Anchor-free algorithms consist of keypoint-based and center-based algorithms. The keypoint-based algorithms, such as CenterNet [33] and CornerNet [34], first detect the keypoints and then combine the keypoints for object detection, while the center-based algorithms, such as adaptive training sample selection [35] and fully convolutional one-stage object detection [36], directly detect the center point and predict the four distances to the target boundary. These anchor-free algorithms abandon or bypass the concept of anchor, and use a more streamlined way to determine positive and negative samples, which have achieved similar performance with anchor-based algorithms.

Despite the success of deep-learning-based algorithms in ship detection, there are still some problems which need to be coped with: 1) features still need to be effectively fused to better enhance the discrimination of the multiscale targets, and 2) imbalance positive and negative samples and how to define positive and negative training samples have a significant impact on the ship detection performance. In this article, we propose a novel anchor-free model named salient feature fusion (SFF)–YOLOX with SFF mechanism for accurate ship detection in SAR images. First, we redesign the network of anchor-free algorithm YOLOX [37] for the consideration of high detection accuracy and high running speed. Secon, a saliency region extraction (SRE) module is proposed to generate the corresponding salient guide map of the input image, and the backbone of SFF–YOLOX consists of two parallel pipelines. The upstream pipeline extracts multiscale deep features from the input images, while the downstream pipeline extracts multiscale salient features from the corresponding salient guide maps by SRE. Besides, we propose the SFF mechanism to perform multiscale feature fusion operations of the two parallel pipelines, and we utilize the state-of-the-art BiFPN [38] with the input of three-level multiscale feature maps to further fuse the features. Finally, we improve the SimOTA mechanism [37] by introducing the anchor IoUs to the ground truth bounding boxes to perform label assignment. The comparison experiments are conducted on public dataset, which prove that the proposed SFF–YOLOX outperforms the mainstream deep-learning-based algorithms.

The main contributions of this article are as follows:

1) For the consideration of both high detection accuracy and high running speed in ship detection, we redesign the network of the classic anchor-free algorithm YOLOX.
2) We introduce an SRE module to generate the salient guide map and two parallel pipelines to extract multiscale features.
3) We propose an SFF module based on attention mechanism to obtain the fused features by deep feature maps and salient feature maps, which highlight the salient regions of ship targets.
4) We improve the SimOTA mechanism by introducing the anchor IoUs, which will shield the adverse effect of the inaccurate predictions of SFF–YOLOX in the early stage of training for the tasks of object classification and bounding box regression.

II. Related Works

Due to the improvement brought by deep-learning-based algorithms, there have been more and more studies in which the deep networks are served as the solutions in the field of ship detection in SAR images. Miao et al. [39] proposed an improved lightweight RetinaNet for ship detection in SAR images by replacing the shallow convolutional layers of the backbone into ghost modules and reducing the number of the deep convolutional layers, which can significantly decrease the floating-point operations while maintaining the model's robustness and the ability to detect ship targets. Yang et al. [40] designed a one-stage ship detector with strong robustness against scale changes and various interferences and introduces a coordinate attention module to obtain more representative semantic features to accurately locate and distinguish ship object. Li et al. [41] proposed an improved YOLOv5 SAR image ship target detection network based on the lightweight ideas of GhostNet and DWConv, and the proposed model with only one-half of original YOLOv5's model size do not have much loss in mean average precision and recall metrics. Fu et al. [42] designed a detection method named feature balancing and refinement network to eliminate the effect of anchors by adopting a general anchor-free strategy that directly learns the encoded bounding boxes. Guo et al. [43] improved the CenterNet by introducing a feature pyramids fusion module and a head enhancement module to reach a high accuracy for small ship detection under complex background. Min and Liu [44] improved the performance of SSD algorithm by redesigning the shallow network structure and enlarging the receptive field of features, which raises the accuracy about 7% while reducing the model's size. For the cases of overcoming fewer training samples, Rai et al. [45] proposed a semisupervised segmentation algorithm for ship SAR images, which requires only a few labeled samples to outperform the current mainstream semisupervised and supervised models. Chen et al. [46] devised a semisupervised learning strategy, which makes full use of unlabeled ship data and iteratively outputs higher quality labeled samples, and the comprehensive results shows the superiority of the proposed model.

Detectors tend to detect large and medium ship objects, and the feature representation of small ships or weak ships still needs further improvement. A lot of tricks are applied to enhance the feature representation, and fusing salient feature has proved to be the most effective trick. Li et al. [47] exploited the channel attention and spatial attention mechanism to enable the FPN to learn semantic and multilevel features, and the results for multiscale ships are superior to the existing algorithms. Zhao et al. [48] proposed an orientation-aware feature fusion network, which fuses the global and local information in feature extraction stage. Zhang et al. [49] proposed a multilevel feature fusion module, which combines the location and semantic information of different level features, and the proposed model achieves a good detection performance in large-scale SAR images. Wang and Chen [50] introduced an optimal window selection mechanism by multiscale local contrast measure to distinguish the similarity between the ship object and surrounding anchors. Xie et al. [51] studied the fusion problem of two lightweight models and proposed a novel end-to-end object detection framework fused with a coordinate attention module and YOLOv5 detector, which show significant gains in both efficiency and performance. Gao et al. [52] introduced an anchor-free convolutional network with dense attention feature aggregation mechanism by combining the multiscale features through dense connections and iterative fusions, and the experimental results demonstrate the effectiveness for multiscale ship detection.

Label assignment plays an important role in modern target detection models and it samples positives and negatives while training. Anchor-based detectors, like RetinaNet, preset anchors of multiple scales and aspect ratios on each location and resort to the intersection over union (IoU) for defining positive and negative samples among spatial-level and scale-level feature maps. The positive samples are those anchors with greater IoUs than the predefined threshold, while the negative samples are those anchors with smaller IoUs than the threshold. There are also some detection models, which utilize two thresholds, one for positives and the other for negatives, those anchors whose IoUs are between the two thresholds are ignored during the training process. However, the strategies with fixed threshold for label assignment do not take into account the differences between objects due to their various shapes and sizes. Anchor-free detectors, like YOLOX, sample a fixed fraction of center area as positive candidates among spatial-level feature maps, and select certain positives from candidates for each object by scale constraints dynamically. SimOTA label assignment strategy applied in YOLOX first determines the parameter k for the number of positive samples of each object by counting the prediction IoUs between predicted bounding boxes and ground truth bounding boxes, and then calculates a cost matrix in which the smaller the value is, the more suitable for prediction of this anchor point, the k anchor points with the smallest cost values are finally selected as positives for model training. SimOTA completes the assignment for different scale ship targets with different number of positive samples, and the method succeeds in preventing the situation of assigning the same number of positive samples for different targets in a unified scenario in traditional label assignment strategies. In this article, we try to improve the SimOTA by adding anchor IoU to predicted IoU to calculate the threshold dynamically, which could help for selecting more high-quality positives.

III. Methodology

In this section, we present the overall pipeline of SFF–YOLOX at first, and then, the improvements which contribute to the performance of ship detection will be introduced, respectively.

A. *Overall Pipeline of SFF–YOLOX*

Fig. 1 gives the overall pipeline of the proposed SFF–YOLOX. The model takes the SAR images as input with resized scale of 640 × 640. Then, the proposed SRE module is utilized to perform salient region extraction operation by which we can obtain corresponding salient guide map of the original SAR image, we call the SAR image and its salient guide map by salient pair. Next, we redesign the backbone of YOLOX by adopting CSPDarknet [28] and Swin-Transformer (Swin-T) [53] as two parallel pipelines, and the network extracts multiscale deep feature maps and salient feature maps from salient pairs with feature sizes of 80 × 80, 40 × 40, and 20 × 20. Besides, we introduce an SFF module based on attention mechanism to complete SFF operation, the module replaces the global feature fusion, such as elementwise summation or concatenation with selective feature fusion. Specifically, the spatial attention is applied to multiscale salient feature maps to generate the weighted descriptor map, which will be then projected to the multiscale deep feature maps. In BiFPN module, we adjust the number of input and output into three-level which accelerates

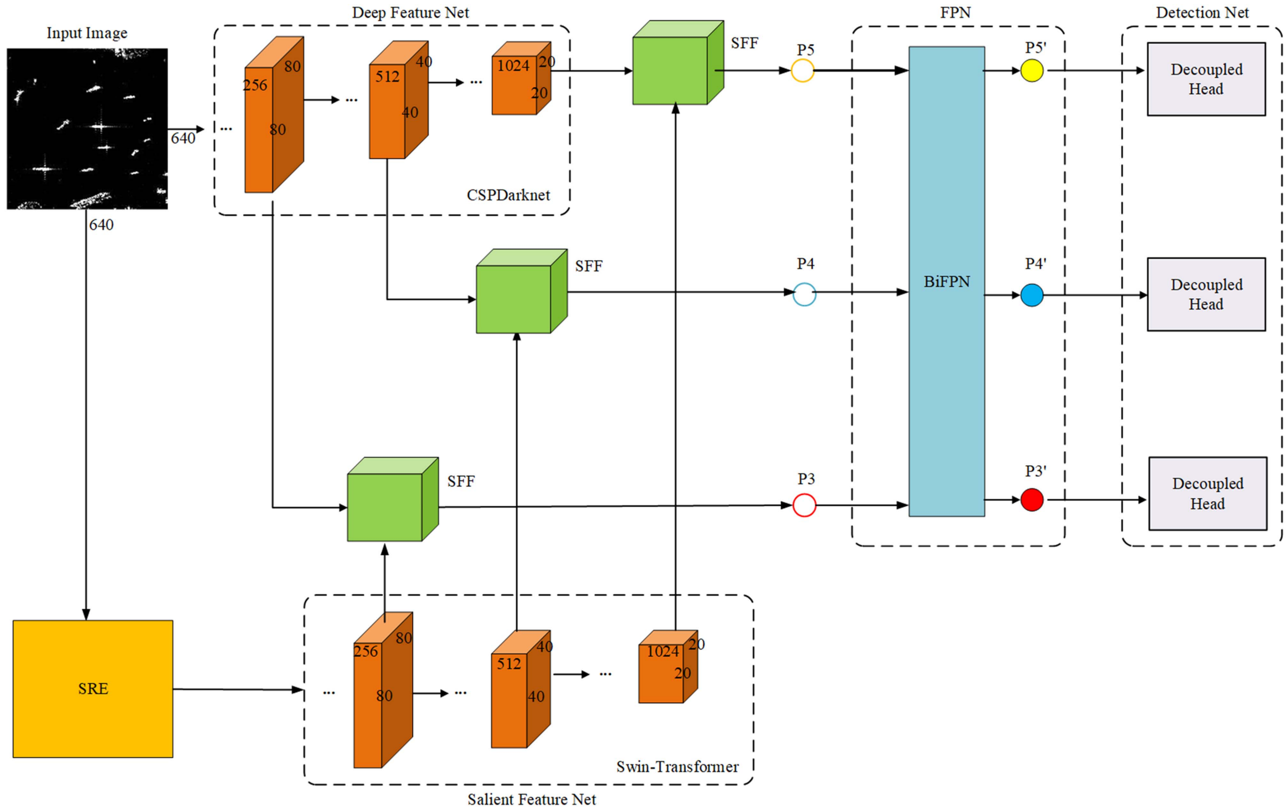


Fig. 1. End-to-end pipeline of SFF–YOLOX.

the running speed while maintaining the performance of the original BiFPN. Finally, the improved SimOTA is used as label assignment to select more high-quality positives.

B. Saliency Region Extraction

SRE is one of the research hotspots in the field of computer vision and image processing, whose goal is to quickly detect the salient region in an image, and it has been widely utilized in object recognition and object detection over the years. Therefore, we introduce the SRE into the detection model as one of the means to raise the detection accuracy.

The image consists of a low-frequency part and a high-frequency part in the frequency domain. The low-frequency part reflects the overall information of the image, such as the texture of the object and the basic composition area, while high-frequency part reflects the detailed information of the image, such as the outline of the objects. SRE uses more information of the low-frequency part. The proposed SRE algorithm analyzes the image from the angle of frequency and divides the process into five subtasks, including Gaussian smoothing, obtaining six-scale pyramid features, converting the color space, saliency calculation, and constructing salient guide map. The lightweight SRE algorithm highlights the salient regions of ship targets which will be used to optimize the feature representation.

Algorithm 1 presents the steps of the SRE algorithm in detail. In SRE algorithm, the upsampling and downsampling operations by the factor of 2 is used to obtain multiscale features, the I and L represent the six-scale pyramid feature set in RGB and LAB space, and the feature sizes are 640×640, 320×320, 160×160, 80×80, 40×40, and 20×20, respectively. The SF indicates the six-scale saliency feature set, which is used for constructing salient guide map in step 5. It is worth noting that the input SAR images follow a batch-normalization layer to conform to the same distribution, which accelerates the rapid convergence of the detection model.

C. Salient Feature Fusion

Feature fusion, the integration of features from different scales or branches, is often implemented by simple operations like elementwise summation or concatenation, but this might not be the best choice. Recently, attention mechanism has been widely introduced in multiscale feature fusion methods and FPN structure due to the ability of dynamically capturing the spatialwise and channelwise dependencies, providing new ideas for solving the fusion problems.

As depicted in Fig. 2, SFF module can be divided into three subnets, including feature encoding, feature refining, and feature decoding. Feature encoding based on spatial attention mechanism imposes average pooling and max pooling on the salient feature map to generate two intermediate tensors with size of $1 \times H \times W$, and then we complete the encoding operation via elementwise summation. Feature refining is a network which is made up by two fully connected layers (FC) and a sigmoid layer, the number of the activation units for the first FC layer reduces to $1 \times H \times W/r$, while for the second FC layer, it goes back to $1 \times H \times W$. The scaled weighted descriptor map s is then obtained by

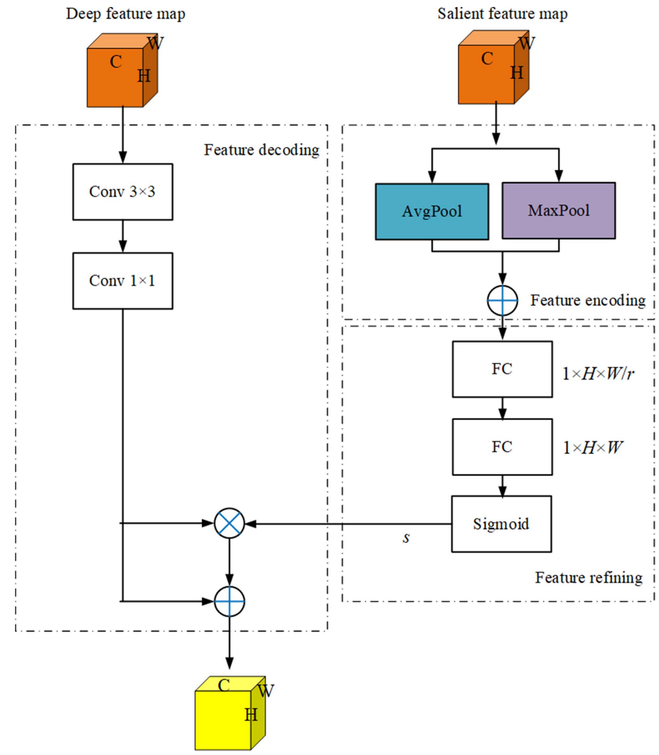


Fig. 2. Structure of SFF module.

Algorithm 1: SRE Algorithm.

```
Input: Resized SAR image $X$ with size of 640 × 640;
Output: The salient guide map $Y$ of the input;
STEP 1: Gaussian smoothing is applied to $X$ to filter
  high-frequency information: $X_f = G_{3\times3} \otimes X$, where
  $G_{3\times3}$ is a Gaussian operation with the kernel size of 3 ×
  3, and $\otimes$ is a convolution operator;
STEP 2: Down-sample $X_f$ to obtain the six-scale pyramid
  feature set $I$, $I = \{I_0, I_1, I_2, I_3, I_4, I_5\}$;
STEP 3: Convert the color space from RGB to LAB, and
  the feature set turn $L$, $L = \{L_0, L_1, L_2, L_3, L_4, L_5\}$;
STEP 4: for each $i \in [0,5]$ do
        Calculate the mean image of each channel in LAB
          space: $T_i = \mathrm{mean}(L_i)$;
        Calculate the sum of Euclidean distance in three
          channels, which represents the saliency feature of
          the image: $SF_i = \| T_i - L_i \|_2$, where $\| \bullet \|_2$ is the L2
          norm;
        $SF_i = \mathrm{ReLU}(SF_i)$, where ReLU is the leaky rectified
          linear operator;
    end for
STEP 5: $Y_5 = SF_5$;
    for $i = 5$; $i>0$; $i$– do
        $Y_{i\text{-}1} = \mathrm{ReLU}\ [SF_{i\text{-}1} +$ up-sampling $(Y_i)]$;
    end for
STEP 6: Salient guide map $Y = Y_0$.
```

the refining network via the sigmoid operation. The tensor s positions the salient area of the feature map and could be used to optimize the multiscale deep features in the feature decoding stage. The deep feature map is forward to the feature decoding network and is first processed by convolution operations with kernel size of 3 × 3 and 1 × 1, respectively, which is then optimized by the map s via elementwise multiplication and summation. These processes can be summarized as

$$K_{i1} = \mathrm{ReLU}(\mathrm{Conv}_{3\times3}(X_i)), i = 3, 4, 5 \tag{1}$$

$$K_{i2} = \mathrm{ReLU}(\mathrm{Conv}_{1\times1}(K_{i1})), i = 3, 4, 5 \tag{2}$$

$$P_i = K_{i2} \oplus K_{i2} \otimes s, i = 3, 4, 5 \tag{3}$$

where Xi represents the multiscale deep features and Pi denotes the fused feature maps of SFF module. Conv3 × 3 is the 3 × 3 convolutional layer and Conv1 × 1 represents the convolutional layer with kernel size of 1 × 1. $\oplus$ represents elementwise summation and $\otimes$ represents elementwise multiplication operation, respectively.

D. *Label Assignment*

SimOTA utilizes the predicted IoUs to dynamically allocate k positives for different targets. We improve the SimOTA mechanism by introducing the anchor IoUs as priors, and the process of the improved SimOTA is summarized as follows:

1) Consistent to SimOTA, the improved SimOTA identifies the positive sample candidate area based on each ground truth and the distance to the center of the target.
2) Compute the predicted IoUs and the anchor IoUs of each anchor in candidate area, and the final integrated IoUs are calculated via the summation operation.
3) Calculate the cost in candidate area.
4) Determine the number of positive samples k for each ground truth by ceiling the sum of maximum *n_candidate_k* (set to 10) integrated IoUs.
5) The k anchors with the lowest cost are served as positive samples for the ground truth, while the others are deemed negative.
6) Calculate the loss for training using the positive and negative samples.

IV. Dataset and Implementation Configurations

A. *Dataset and Evaluation Metrics*

In this section, we first introduce three public datasets that are used in this article to evaluate the proposed model SFF–YOLOX, then we briefly explain some popular evaluation metrics, respectively.

We use the SSD [54] which contains 43 819 ship images and a total of 59 535 ship objects as the dataset to train and test the models for the detection accuracy and the running speed, and the HRSID [55] dataset including 5604 images and 16 951 ship objects to evaluate the generalization ability of models. For the detection task in complex and large-scale SAR images, we crop image slices with 800 × 800 pixels under the overlapped ratio

TABLE I
DETAILED INFORMATION OF DATASETS

Datasets	Size (pixel)	Ship number	Imaging mode	Polarization	Image sensor	Resolution (m)
SSD	256 × 256	59 535	UFS/FS/QPS/ SM	VV/HH/VH/HV	Gaofen-3/Sentinel-1	3–10
HRSID	800 × 800	16 951	S3-SM/ST/ HS	VV/HH/VH/HV	Sentinel-1/TerraSAR-X/ TanDEM-X	1–15
Two complex large-scale SAR images	16746 × 24919 16374 × 21953	984	UFS/FS	VV/HH/VH/HV	Gaofen-3	3–10

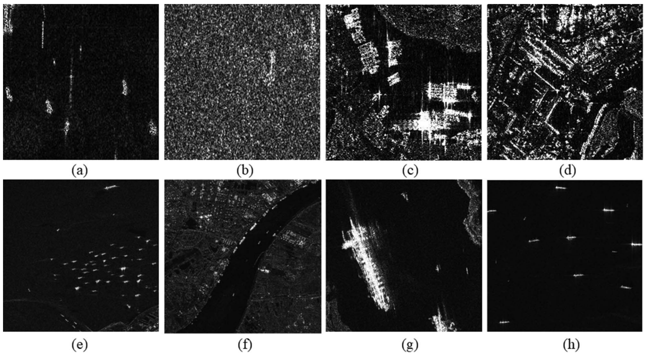


Fig. 3. Some image samples. The first row comes from SSD and the second row comes from HRSID.

of 30% from two complex large-scale SAR images to further evaluate the generalization ability of the proposed model. The detailed information of the datasets are presented in Table I. The images in the datasets are in gray-level image format which is same with the single-channel bitmap optical images, and the ship annotations are marked in a similar format to Pascal VOC, and according to the distribution information of ship widths and heights, the small ships belong to the targets whose scales do not exceed 32 × 32 pixels, the medium ships whose scales are between 32 × 32 and 96 × 96 pixels, and the large ships are the targets whose scales exceed 96 × 96 pixels. Besides, we also display some typical images of the datasets in Fig. 3, which vividly demonstrate the difficulty of multiscale ship detection and the importance of feature representation enhancement.

As for evaluation metrics, the precision, recall, F_1, AP, and FPS are selected to quantify the models, and the metrics are defined as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

where TP and FP express the amounts of true positives and false positives, respectively, and FN is the number of false negatives

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{6}$$

$$\text{AP} = \int_0^1 p(r)dr. \tag{7}$$

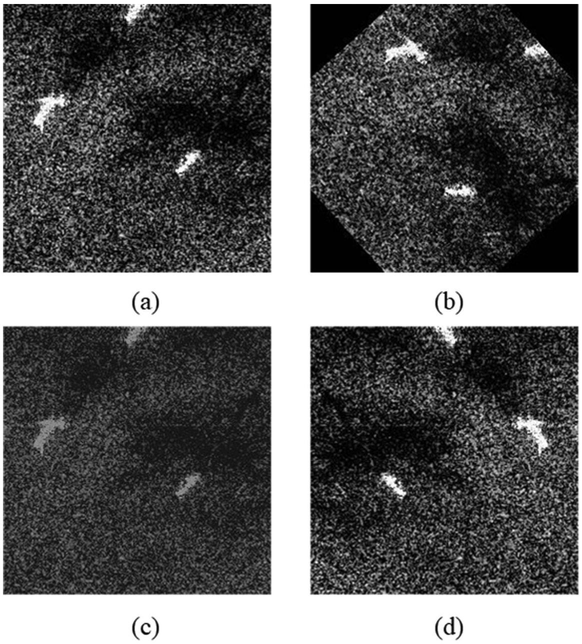


Fig. 4. Data augmentation effect on the input image. (a) is the original input image, and (b)–(d), respectively, present the RandomAffine, ColorJitter, and RandomFlip operations.

B. Implementation Configurations

We design and implement the models on the base of the framework of Pytorch and a computer with TITAN XP GPU. All models utilize the GPU platform to train in batch size of 16. We first use the data augmentation mechanisms to enhance the diversity of the SAR images, and then feed the transformed images to the trained model. Fig. 4 illustrates the data augmentation effect on the input image. The images of training datasets are randomly processed by the data augmentation mechanisms, where the ColorJitter operation changes the brightness of the input images, and the RandomAffine and RandomFlip operations change the geometry or position information of ship targets. Once the geometry or position information changed, we will update the corresponding annotation data. The parameter r in SFF module is set to 8 and the parameter n_candidate_k in improved SimOTA is set to 10. Label smoothing strategy is used for classification to prevent SFF–YOLOX from overfitting. We use the cosine annealing optimization method to adjust the learning rate during training, and the momentum and weight decay are, respectively, set to 0.9 and 0.0005. SFF–YOLOX is

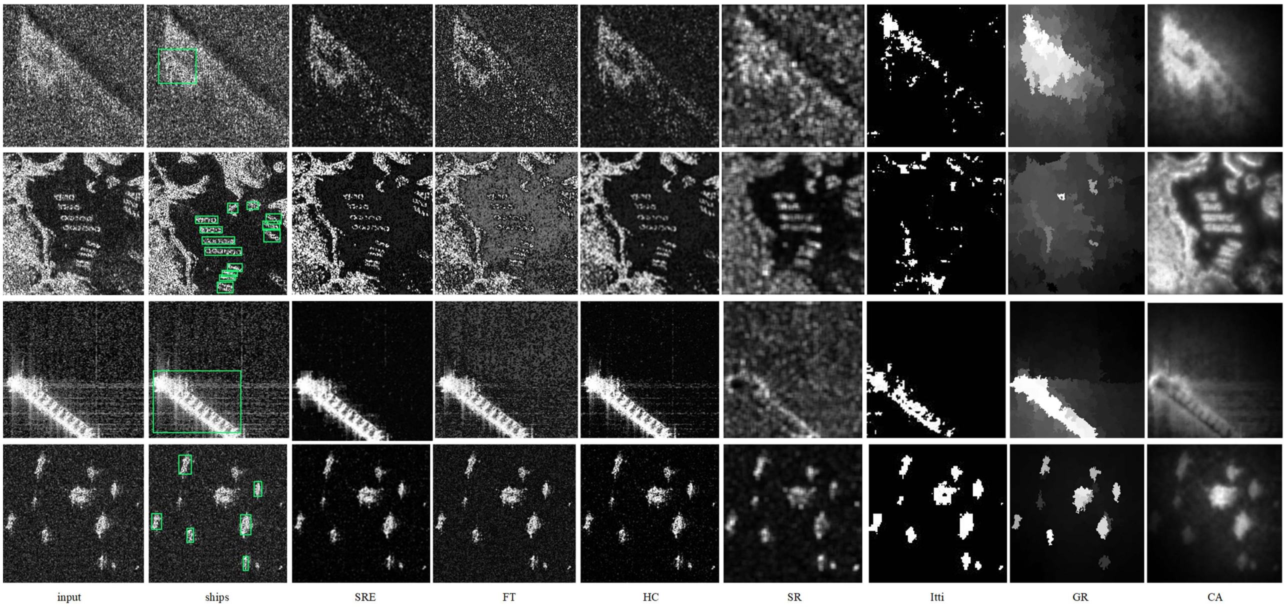


Fig. 5. Visual comparison between salient map extraction algorithms. The rectangle with green color denotes the ships.

trained for 2000 iterations, and the IoU threshold for NMS is 0.5 and 0.75, respectively. The SSD dataset is randomly divided into train, validation, and test subsets on a proportion of 7:2:1, while the whole HRSID and two complex large-scale SAR images are served as test sets to evaluate the generalization ability.

V. Experiments

A. Performance of SRE

In this section, we verify the performance of SRE by comparing with six salient map extraction algorithms, and the algorithms consist of frequency-tuned [56], histogram-based contrast (HC) [57], spectral residual [58], Itti [59], graph-regularized [60], and context-aware [61]. We first examine the visual comparison between the algorithms. We select four typical sample images from the dataset, which have the characteristics of complex noises, inshore backgrounds, strong backscatters, and multiple ship targets, then we forward the images into the salient map extraction modules and obtain the comparison results. As we can see from Fig. 5, SRE reduces the adverse impact of noises and backscatters to a certain extent, and distinguishes the ship objects well from the background at the same time compared with other algorithms, which will be beneficial for the following extraction of features. Besides, we transform the processing results of the algorithms in terms of heatmap and visualize the heatmaps in Fig. 6. The heatmaps will help us to estimate which part of the image has the most impact on the final results. As it is seen in Fig. 6, salient regions are marked with different degrees of red color according to the processing results. And SRE perfectly reserves the whole information of ship objects, in addition, more distinct outlines of the salient regions can also be captured by our proposed SRE algorithm over the others which will be validated effective for locating the ship targets.

To validate the effectiveness of SRE in objective metrics, the compared performance between YOLOX, SFF–YOLOX without SRE module (DeSRE–SFF–YOLOX), and SFF–YOLOX is presented in Table II. As we can conclude from Table II that DeSRE–SFF–YOLOX slightly outmatches the baseline YOLOX due to the other proposed improvements, with the results of 2.11%, 0.85%, 0.02, and 0.21% growth in precision, recall, F_1, and AP_{50}, and 0.58%, 2.28%, 0.01, and 0.09% growth in precision, recall, F_1, and AP_{75}. SRE significantly lifts the performance up by almost 2% in overall metrics versus DeSRE–SFF–YOLOX, by which the proposed SFF–YOLOX enriches the feature representation of ship objects.

Besides, we also test the performance of the six salient map extraction algorithms to further prove the effectiveness of SRE in Table II. The SRE is replaced by other algorithms while we keep other submodules in accordance with SFF–YOLOX. From Table II, introducing some salient map extraction algorithms to detection network could achieve better results than the DeSRE–SFF–YOLOX, which can help extract salient guide maps to optimize the position locating. Furthermore, according to the statistical results, our proposed SRE can be verified to perform better than the other algorithms since the active areas predicted by these algorithms are blurry or incomplete, resulting in errors of missing ships or false alarms. Overall, the experiment results demonstrate that the SRE is conductive to SAR ship detection task.

B. Performance of Salient Feature Net

The experiments are also conducted to delve into the performance of salient feature net with variant backbones. For selecting the backbone of salient feature net, we leverage some mainstream networks, including ResNet-50, ResNet-101, ResNet-152 [62], CSPDarknet, and Swin-T to test our method. Generally speaking, increasing the network depth reasonably can enrich multiscale features, which helps to accelerate the detector's performance, however, this could also easily trigger the problems of gradient explosion and gradient dispersion. ResNet-X is introduced to deal with the shortcomings by applying convolutional residual units which use cross-layer connections, and

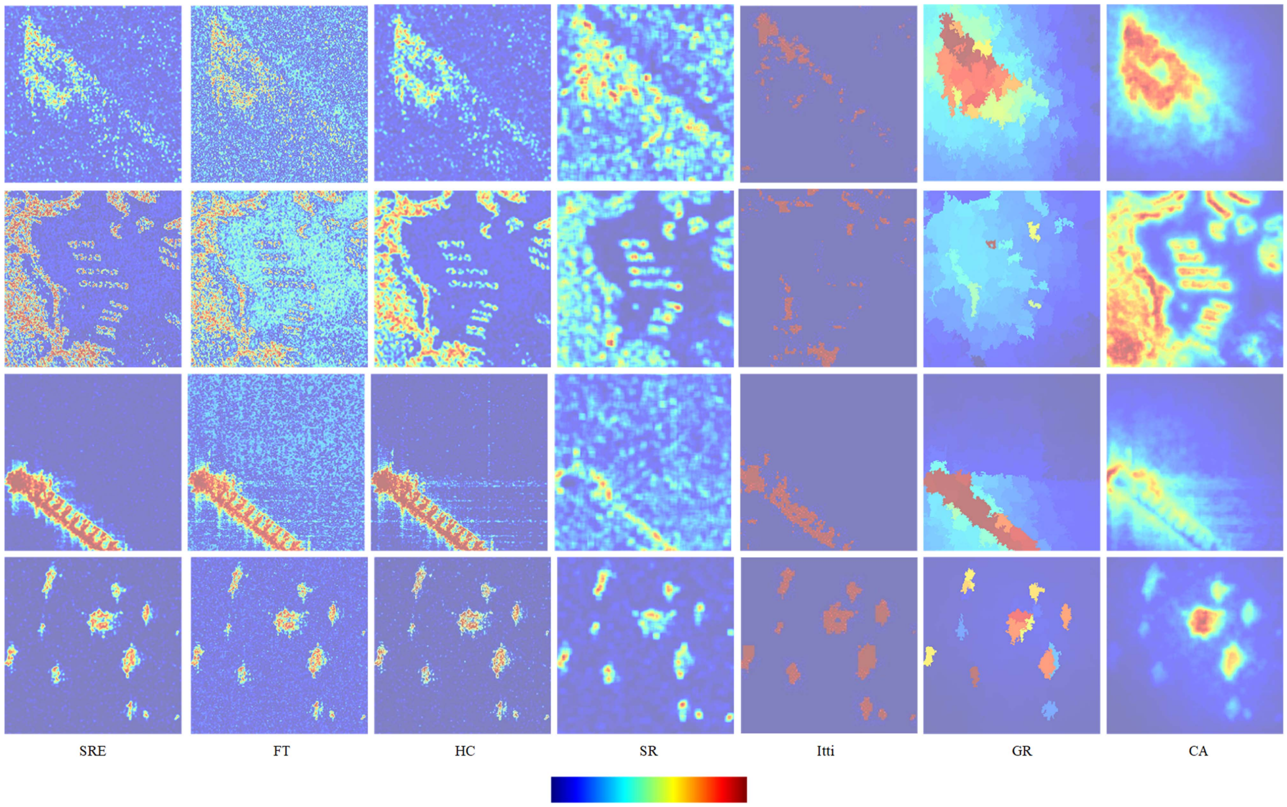


Fig. 6. Visualization of the heatmaps.

TABLE II
Comparison Results With Salient Map Extraction Algorithms

Method	IoU = 0.5				IoU = 0.75			
	AP_{50}	Precision	Recall	F_1	AP_{75}	Precision	Recall	F_1
YOLOX	91.56%	90.43%	85.26%	0.88	56.69%	65.44%	60.08%	0.63
DeSRE-SFF-YOLOX	91.77%	92.54%	88.11%	0.90	56.78%	66.02%	62.36%	0.64
SFF-YOLOX	**95.41%**	**94.04%**	**90.53%**	**0.92**	**60.85%**	**66.94%**	**63.11%**	**0.65**
DeSRE–SFF–YOLOX + FT	93.86%	92.68%	89.37%	0.91	59.28%	66.13%	62.54%	0.64
DeSRE–SFF–YOLOX + HC	94.24%	93.56%	90.06%	0.92	59.48%	66.50%	62.37%	0.64
DeSRE–SFF–YOLOX + SR	92.00%	90.62%	87.43%	0.89	57.24%	65.10%	61.49%	0.63
DeSRE–SFF–YOLOX + Itti	89.22%	88.42%	85.54%	0.87	52.15%	61.29%	56.12%	0.59
DeSRE–SFF–YOLOX + GR	90.35%	88.47%	86.13%	0.87	55.07%	63.86%	57.18%	0.60
DeSRE–SFF–YOLOX + CA	91.46%	91.72%	86.84%	0.89	56.93%	64.41%	60.58%	0.62

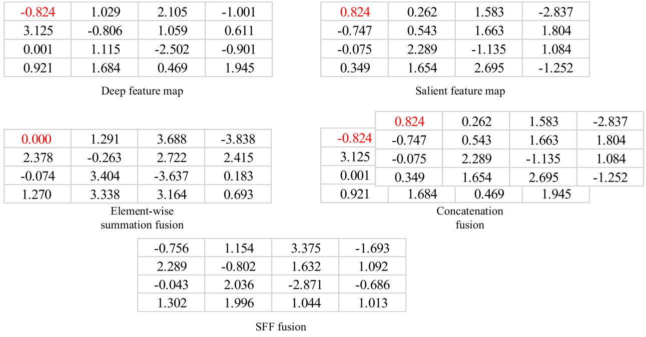
TABLE III
PERFORMANCE OF SALIENT FEATURE NET WITH DIFFERENT BACKBONES

Backbone of salient feature net	IoU = 0.5				IoU = 0.75			
	AP_{50}	Precision	Recall	F_1	AP_{75}	Precision	Recall	F_1
ResNet-50	92.12%	91.79%	88.04%	0.90	55.80%	62.67%	60.14%	0.61
ResNet-101	93.76%	92.87%	88.89%	0.91	58.57%	63.35%	60.81%	0.62
ResNet-152	93.57%	92.85%	88.70%	0.91	58.32%	63.11%	60.39%	0.62
CSPDarknet	94.19%	93.65%	90.44%	0.92	60.16%	65.48%	**63.14%**	0.64
Swin-T	**95.41%**	**94.04%**	**90.53%**	**0.92**	**60.85%**	**66.94%**	63.11%	**0.65**

since then the design has been widely used in many structures. CSPDarknet uses the split and merge strategy across stages to achieve a richer gradient combination while reducing the amount of computation. Swin-T does duty for a general-purpose pipeline in image processing which produces more efficiency by bounding self-attention operations to nonoverlapping local windows while taking into the cross-window connections. The comparison accuracies of backbones are listed in Table III. From Table III, Swin-T exhibits the best results for AP_{50}, AP_{75}, F_1, precision, and recall. ResNet-50 has relatively fewer parameters compared with others, which tends to miss the detections when forward the features lack semantic information. The location information of small-scale objects extracted by ResNet-101 or ResNet-152 is seriously lost and the accuracy will turn saturated or even decrease when increasing the network depth. CSPDarknet achieves better results than the ResNet series, indicating the enhancement of learning capability by CNNs. CNN network has great advantages in extracting the basic image elements and low-level features, while Swin-T pays more attention to how these elements are related together to form an object, and how the spatial relationship between objects forms a scene. The design of deep feature net and salient feature net aims to combine the advantage of CSPDarknet and Swin-T, and the experiment results prove that the combination can raise the detection performance.

C. Performance of SFF

In this part of experiments, we test the performance of SFF module and make comparison to several methods which are widely used to integrate the salient features to the deep CNN features. The concatenation method integrates the features by expanding dimensions which will augment the parameters and spatial complexity of the detectors. The elementwise summation method is easy to implement by adding tensors element by element; however, the fusion also easily triggers the feature disappearance in some cases which would ruin the distribution of features. Fig. 7 displays the fusion results of the concatenation method, the elementwise summation fusion method, and our proposed SFF method, deep feature map, and salient feature map are part of outputs of the two-stream network, respectively. The concatenation method keeps both features by expanding dimensions, and the fused feature may disappear by the elementwise summation fusion method, for example, the sum of two features "−0.824" and "+0.824" is "0." The purpose of SFF is to project the salient features to the multiscale deep feature, enhancing the diversity of deep CNN features.



-0.824	1.029	2.105	-1.001
3.125	-0.806	1.059	0.611
0.001	1.115	-2.502	-0.901
0.921	1.684	0.469	1.945

Deep feature map

0.824	0.262	1.583	-2.837
-0.747	0.543	1.663	1.804
-0.075	2.289	-1.135	1.084
0.349	1.654	2.695	-1.252

Salient feature map

0.000	1.291	3.688	-3.838
2.378	-0.263	2.722	2.415
-0.074	3.404	-3.637	0.183
1.270	3.338	3.164	0.693

Element-wise
summation fusion

	0.824	0.262	1.583	-2.837
-0.824	-0.747	0.543	1.663	1.804
3.125	-0.075	2.289	-1.135	1.084
0.001	0.349	1.654	2.695	-1.252
0.921	1.684	0.469	1.945	

Concatenation
fusion

-0.756	1.154	3.375	-1.693
2.289	-0.802	1.632	1.092
-0.043	2.036	-2.871	-0.686
1.302	1.996	1.044	1.013

SFF fusion

Fig. 7. Feature fusion results.

The comparison results of detection performance between the feature fusion methods are presented in Table IV. We can conclude that the SFF method achieves much better performance than the other two methods since that the attention mechanism is introduced to SFF to favor the assignment of available processing resources to the most salient parts of the deep feature maps, which can greatly help to highlight the objects of interest from backgrounds. Besides, the ablation experiments for the parameter r is also conducted and the results are listed in Table IV. As we can see from the results, adding the parameter brings the improvements in all metrics compared with condition when $r = 1$. And the condition when r is set to 8 obtains the best detection precision. In a word, the performance benefits from the better feature representation for ship targets.

D. Performance of BiFPN

In SFF–YOLOX, the applied BiFPN further integrates the three-level feature maps {P3, P4, P5}, producing balanced rich semantic and spatial location information for the fused feature maps. We design some relative experiments to test and compare with common FPNs and list the results in Table V. FPN [27]

TABLE IV
PERFORMANCE OF FEATURE FUSION METHODS

Feature fusion method	IoU = 0.5				IoU = 0.75			
	AP_{50}	Precision	Recall	F_1	AP_{75}	Precision	Recall	F_1
Concatenation	92.91%	92.32%	88.72%	0.90	57.60%	64.02%	59.63%	0.62
Element-wise summation	90.18%	88.39%	87.07%	0.88	55.23%	60.95%	56.48%	0.59
SFF ($r = 1$)	93.53%	92.54%	88.98%	0.90	58.85%	64.11%	58.30%	0.61
SFF ($r = 2$)	93.86%	93.25%	89.39%	0.91	58.96%	64.78%	59.43%	0.61
SFF ($r = 4$)	94.47%	93.56%	89.84%	0.91	59.53%	65.66%	62.38%	0.63
SFF ($r = 8$)	**95.41%**	**94.04%**	**90.53%**	**0.92**	**60.85%**	**66.94%**	**63.11%**	**0.65**
SFF ($r = 16$)	94.93%	93.70%	89.81%	0.91	60.06%	66.49%	62.17%	0.64

TABLE V
PERFORMANCE OF FPNS

Method	IoU = 0.5				IoU = 0.75				FPS
	AP_{50}	Precision	Recall	F_1	AP_{75}	Precision	Recall	F_1	
SFF–YOLOX + FPN	93.78%	92.42%	87.06%	0.90	55.93%	63.57%	59.49%	0.61	68
SFF–YOLOX + PAN	94.50%	93.71%	87.62%	0.91	57.44%	63.87%	60.00%	0.62	65
SFF–YOLOX + five-level BiFPN	95.05%	93.85%	89.96%	0.92	60.02%	66.56%	63.09%	0.65	57
SFF–YOLOX + three-level BiFPN	**95.41%**	**94.04%**	**90.53%**	**0.92**	**60.85%**	**66.94%**	**63.11%**	**0.65**	**62**

constructs a top-down structure with lateral connections for generating high-level semantic feature maps at all scales. PAN [28] increases the architecture hierarchy with accurate localization features in lower layers by bottom-up path augmentation, which enhances the fusion results between low-level and high-level features. However, these works treat the features of different resolutions with no distinction, ignoring the fact that they usually contribute to the fused features unequally. Five-level BiFPN is a simple but highly effective weighted bidirectional FPN, which introduces learnable weights for attaching the importance to different input features, repeatedly applying top-down and bottom-up multiscale feature fusion. Based on the five-level BiFPN, we adjust the number of input and output into three-level, which accelerates the running speed while maintaining the performance of the original BiFPN. As it is seen in Table V, the original FPN is inherently subjected to the one-directional feature information flow and therefore achieves the lowest accuracy, however, it has the fastest running speed. PAN has slightly better accuracy than the FPN owing to adding a bottom-up pathway on the top of the original FPN. Five-level BiFPN and our BiFPN achieve the best performance for multiscale feature fusion, but our BiFPN achieves better accuracy and efficiency tradeoffs.

E. Performance of Improved SimOTA

For proving the validity of the improved SimOTA, we test the performance difference between the original SimOTA and our improved SimOTA. We design an ablation experiment to study the ratios between the predicted IoUs and the anchor IoUs of each anchor, and the ratios preset are 0 : 1, 1 : 0, 0.5 : 1, 1 : 0.5, and 1 : 1, respectively. The results are listed in Table VI, and we can conclude from Table VI that simply adding them with the ratio between the two IoUs being 1:1 could achieve the best performance and the improvements happen in all metrics. It is worth noting that the improved SimOTA with the ratio of the anchor IoUs being 0 would turn into the original SimOTA. Besides, we give the training loss curve comparison of SFF–YOLOX with the original SimOTA and the improved SimOTA in Fig. 8, the loss does not have too much difference for both label assignments at the beginning of training due to the anchors dominating the integrated IoUs, gradually, the improved SimOTA presents a bigger decline in the training loss values verses the original

TABLE VI
ABLATION STUDY ON BALANCING THE IOUS

Predicted IoUs	Anchor IoUs	IoU = 0.5				IoU = 0.75			
		AP$_{50}$	Precision	Recall	F_1	AP$_{75}$	Precision	Recall	F_1
0	1	86.98%	83.17%	84.03%	0.84	54.31%	59.31%	55.66%	0.57
1	0	93.77%	92.89%	88.34%	0.91	59.36%	62.21%	60.09%	0.61
0.5	1	92.98%	92.07%	87.26%	0.90	58.20%	60.92%	58.12%	0.59
1	0.5	94.20%	93.34%	88.75%	0.91	59.76%	64.32%	60.69%	0.62
1	1	**95.41%**	**94.04%**	**90.53%**	**0.92**	**60.85%**	**66.94%**	**63.11%**	**0.65**

TABLE VII
DETECTION RESULTS OF DETECTORS ON SSD

Method	IoU = 0.5				IoU = 0.75				FPS	Parameters
	AP$_{50}$	AP$_{L}$	AP$_{M}$	AP$_{S}$	AP$_{75}$	AP$_{L}$	AP$_{M}$	AP$_{S}$		
RetinaNet	85.70%	81.27%	96.20%	85.58%	41.52%	39.59%	64.18%	40.25%	39	60.0M
CenterNet	84.19%	15.68%	89.46%	79.74%	32.91%	4.23%	44.77%	26.14%	78	**32.6M**
Faster-RCNN	83.80%	63.53%	94.57%	69.23%	21.83%	40.01%	42.06%	5.59%	16	61.8M
YOLOv3	90.98%	61.79%	95.96%	90.72%	48.15%	21.18%	62.65%	39.25%	61	62.0M
YOLOv4	93.69%	74.80%	96.42%	91.28%	50.42%	25.64%	64.67%	40.00%	50	64.0M
YOLOX	91.56%	63.95%	94.03%	88.78%	56.69%	38.39%	65.49%	48.78%	**95**	54.2M
YOLOv7	89.07%	60.90%	92.19%	70.13%	55.67%	35.64%	51.89%	35.94%	43	37.6M
SFF-YOLOX	**95.41%**	**85.25%**	**96.57%**	**94.62%**	**60.85%**	**45.36%**	**74.97%**	**51.29%**	62	86.4M

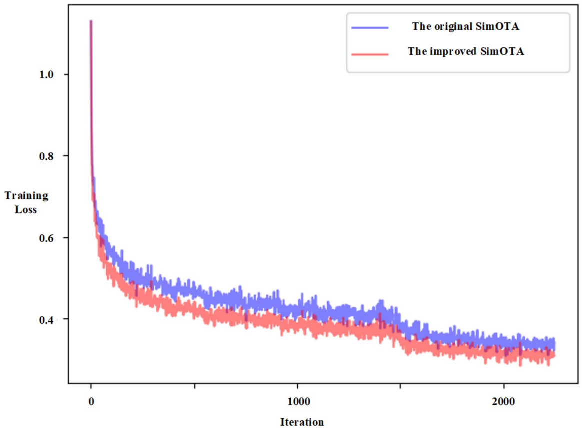


Fig. 8. Training loss comparison.

SimOTA, which further proves the effectiveness of the improved SimOTA for ship detection task.

F. Comparison With State-of-the-Art Methods

In this part of experiments, we implement comparisons between the proposed SFF–YOLOX and other seven deep-learning-based detectors and list the results in Table VII, from which we can clearly draw a conclusion that SFF–YOLOX obtains the uppermost overall performance among the detectors. Apart from AP$_{50}$ and AP$_{75}$, the metrics AP$_{L}$, AP$_{M}$, and AP$_{S}$ are also included to specifically present the detection ability of large-scale, medium-scale, and small-scale ships, respectively. And the metrics are significantly improved by SFF–YOLOX, especially for small-scale ship targets, this may benefit from the SFF and BiFPN for fusing the features, which guarantees the diversity and richness of the features extracted. Besides, we measure and count the FPS and parameters of the detectors and our proposed method achieves 62 FPS which is lower than CenterNet and YOLOX, but faster than the others even though the amount of parameters is largest among the detectors, that is, SFF–YOLOX obtains the best balance between detection accuracy and running speed.

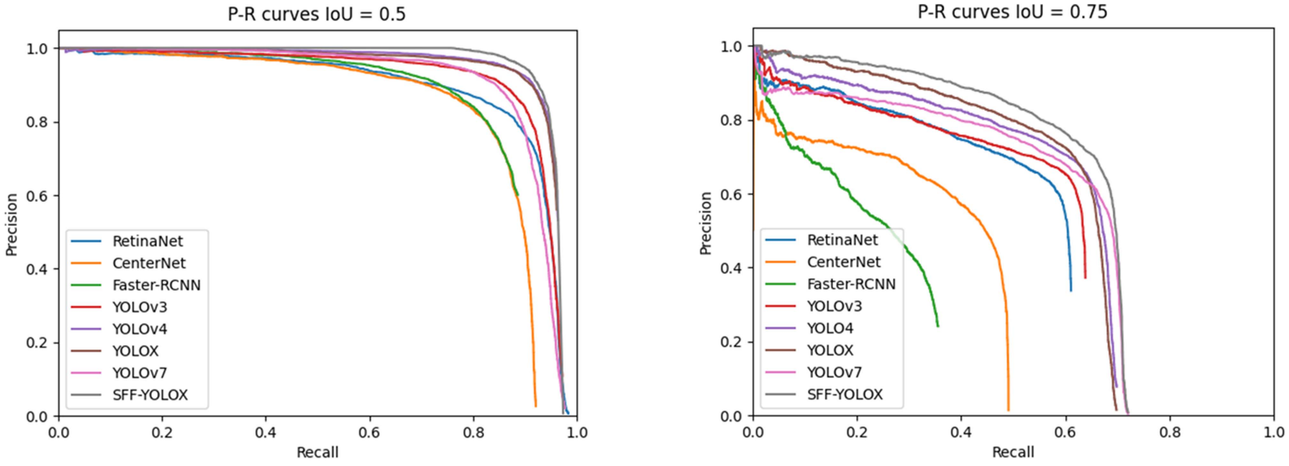


Fig. 9. P-R curves of detectors on SSD.

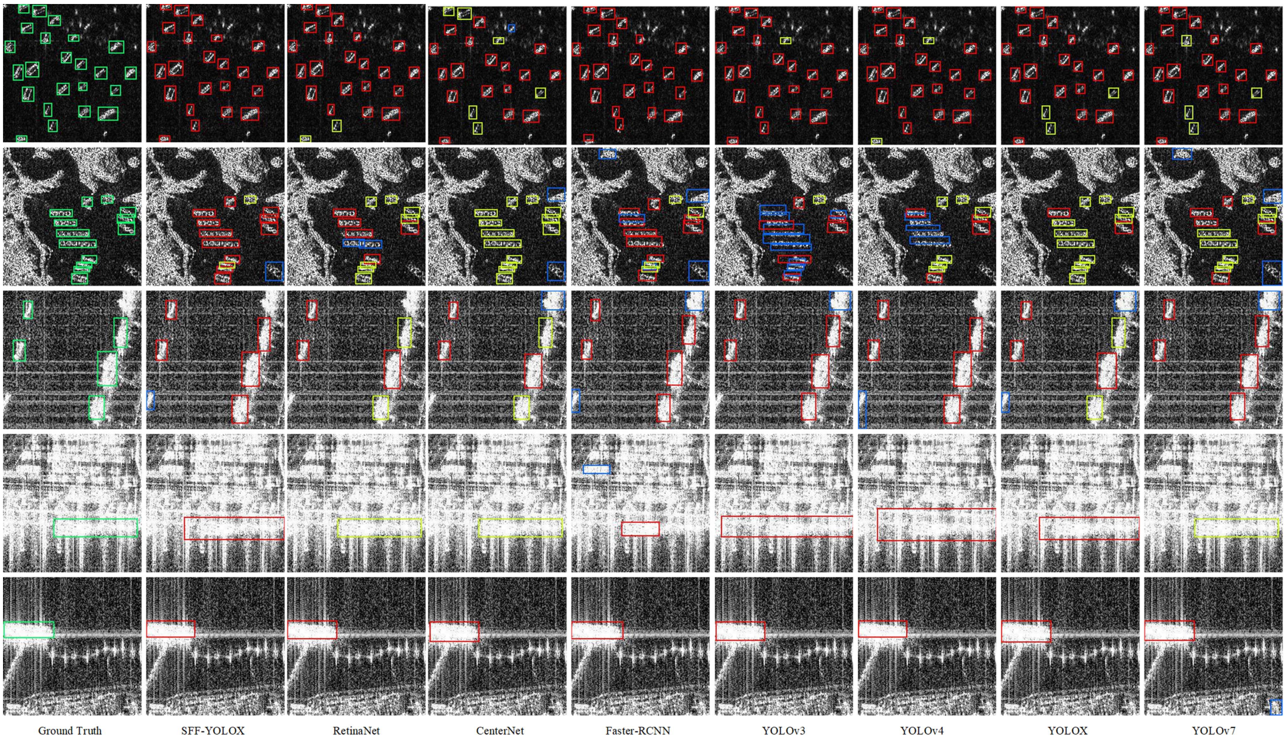


Fig. 10. Detection results on SSD.

Besides, we compare our SFF–YOLOX with CenterNet++ [43] and MGF [13] for ship detection on SSD. The best detection accuracies are cited and introduced to make fair comparisons. The precision, recall, F_1, and AP_{50} for CenterNet++ are, respectively, 83.50%, 97.60%, 0.90%, and 95.40%, while for MGF, the figures are, respectively, 81.98%, 92.35%, 0.87%, and 92.35%. By contrast, our model achieves relatively competitive detection performance.

Further, Fig. 9 displays the P-R curves of all the detectors and Fig. 10 illustrates the comparative detection results. The green, red, yellow, and blue rectangle boxes represent ground truths, detection outputs, missing targets, and false alarms, respectively. By comparing the detection results on five images, especially the second image, SFF–YOLOX produces one false alarm and two missing targets, while others misdetect the most of the ground truths due to the complex backgrounds. On the contrary, SFF–YOLOX is insusceptible to the influence of detection conditions, thus the method can cope with the detection task under different scenarios.

G. *Generalization Ability Testing*

Figs. 11 and 12, respectively, show the inshore and offshore detection results, and Table VIII presents the detection accuracies of these methods on HRSID. It can be observed from Figs. 11 and 12 that SFF–YOLOX misses a few prominent

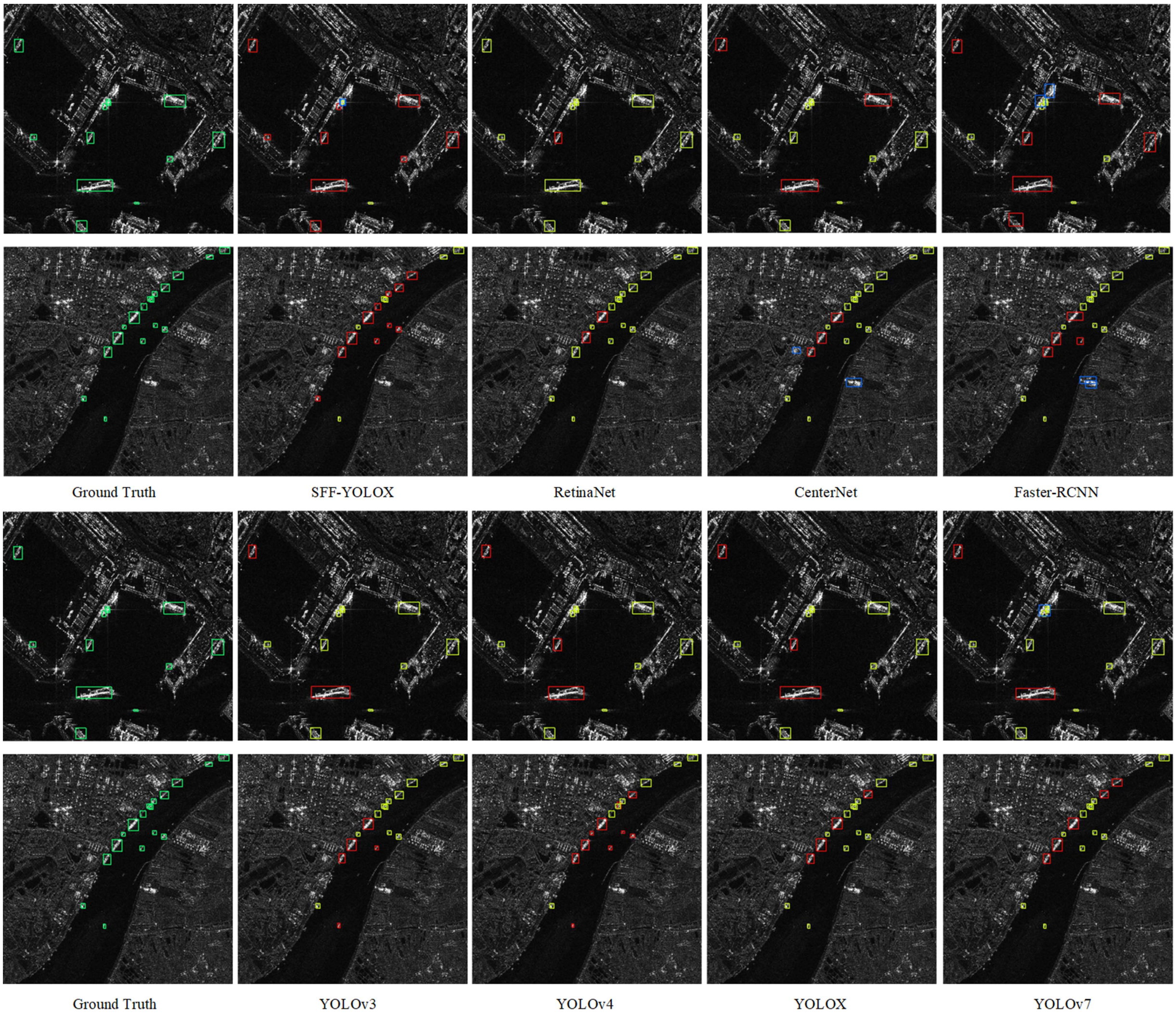


Fig. 11. Detection results for inshore scenes on HRSID.

ships with small sizes and produces false alarms by detecting two close ships into a big ship. However, RetinaNet, CenterNet, faster-RCNN, YOLOv3, YOLOv4, YOLOX, and YOLOv7 perform worse whether it is the inshore scene or not. As for multiscale ship detection in the last row of Fig. 12, SFF–YOLOX outperforms state-of-the-art methods and is least affected by the extreme aspect ratio of ships. Besides, when we compare the accuracy in Table VIII, we can figure out that anchor-free methods, such as SFF-YOLOX and YOLOX, obtain high figures even though the detection dataset changes, while anchor-based methods usually have relatively poor detection results owing to the preset hyperparameters of anchors. The comparison results indicate that the SFF–YOLOX possesses strong generalization ability and robustness.

H. Validation on Complex and Large-Scale SAR Images

In this section, we further validate the proposed model on two complex and large-scale SAR images, and the images are generated from Gaofen-3 satellite and contain multiscale ship targets under complex sea conditions. Figs. 13 and 14 are the detection results of two corresponding cropped slices from the large-scale images and the typical areas marked by bold green rectangles are enlarged and displayed on the right side of the detection results. From the results, we can conclude that SFF–YOLOX has excellent detection performance in both offshore and inshore scenes, there are only two false positives (marked by blue rectangle) and one false positive in Figs. 13 and 14, respectively, which means our proposed model is unlikely prone to the influence of the complex backgrounds. However, the detection results also reflect that SFF—YOLOX is not good for detecting small ships, especially for weak targets with small sizes, and there are two false negatives (marked by yellow rectangle) and six false negatives in Figs. 13 and 14, respectively. This may be due to the ships are too small so that the model lose the discernible features of ships after a certain number of convolution operations.

VI. Conclusion

An anchor-free ship detector named SFF–YOLOX is proposed for both accurate and fast-running ship detection task in

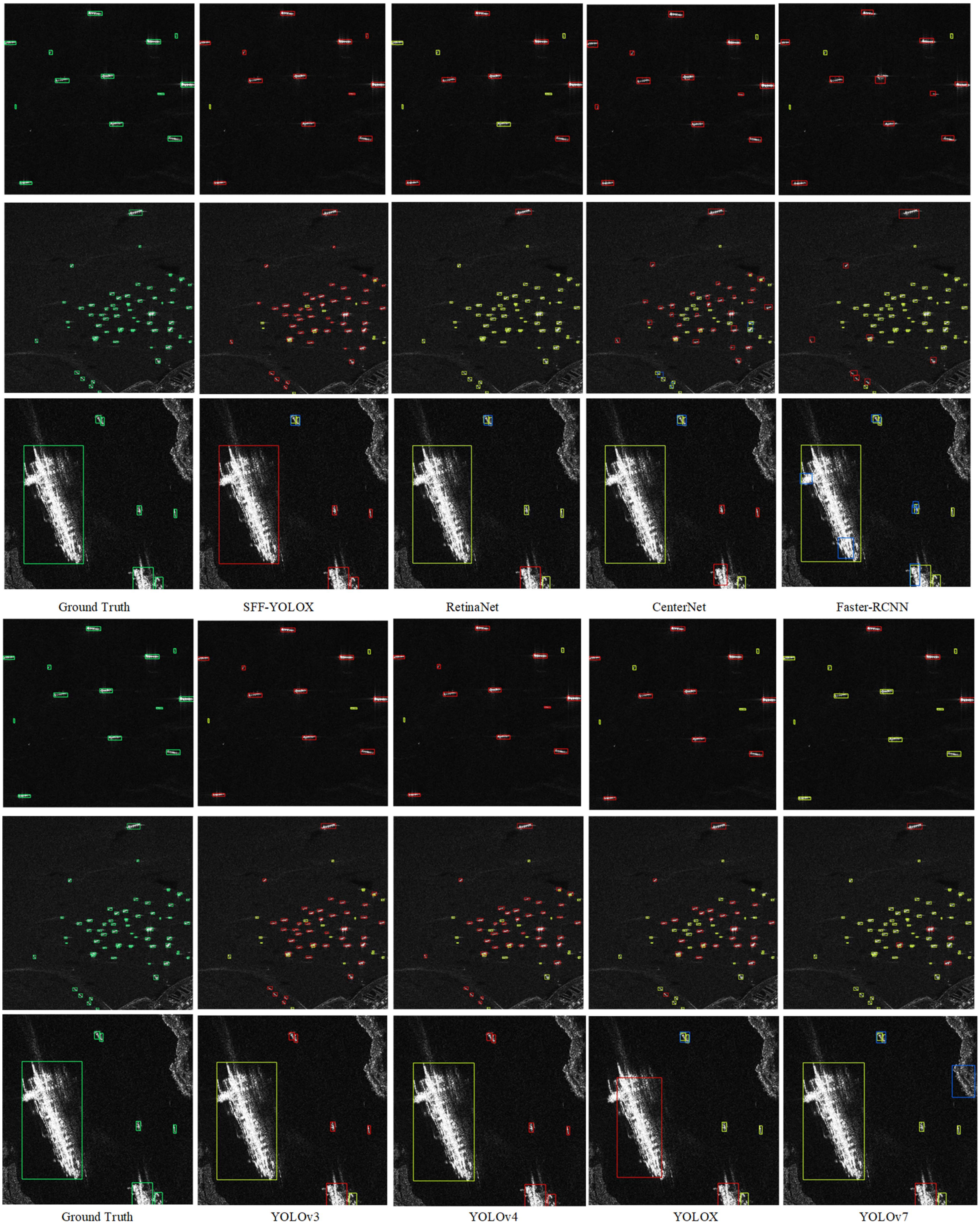


Fig. 12. Detection results for offshore scenes on HRSID.

SAR images. First, the SRE module is introduced to highlight the salient regions, which helps detector extract more discriminative features. Then, we redesign the one-stream feature extraction network of YOLOX into a two-stream network, which is applied to extract deep features and salient features, respectively. In addition, we propose the SFF module based on spatial attention mechanism which projects the salient feature maps to the deep feature maps. Finally, the improved SimOTA is served as label assignment to define positive and negative training samples dynamically. The comparison experiments are conducted on SSD and HRSID datasets to test the accuracy and generalization ability of detectors, and we also validate the proposed model on

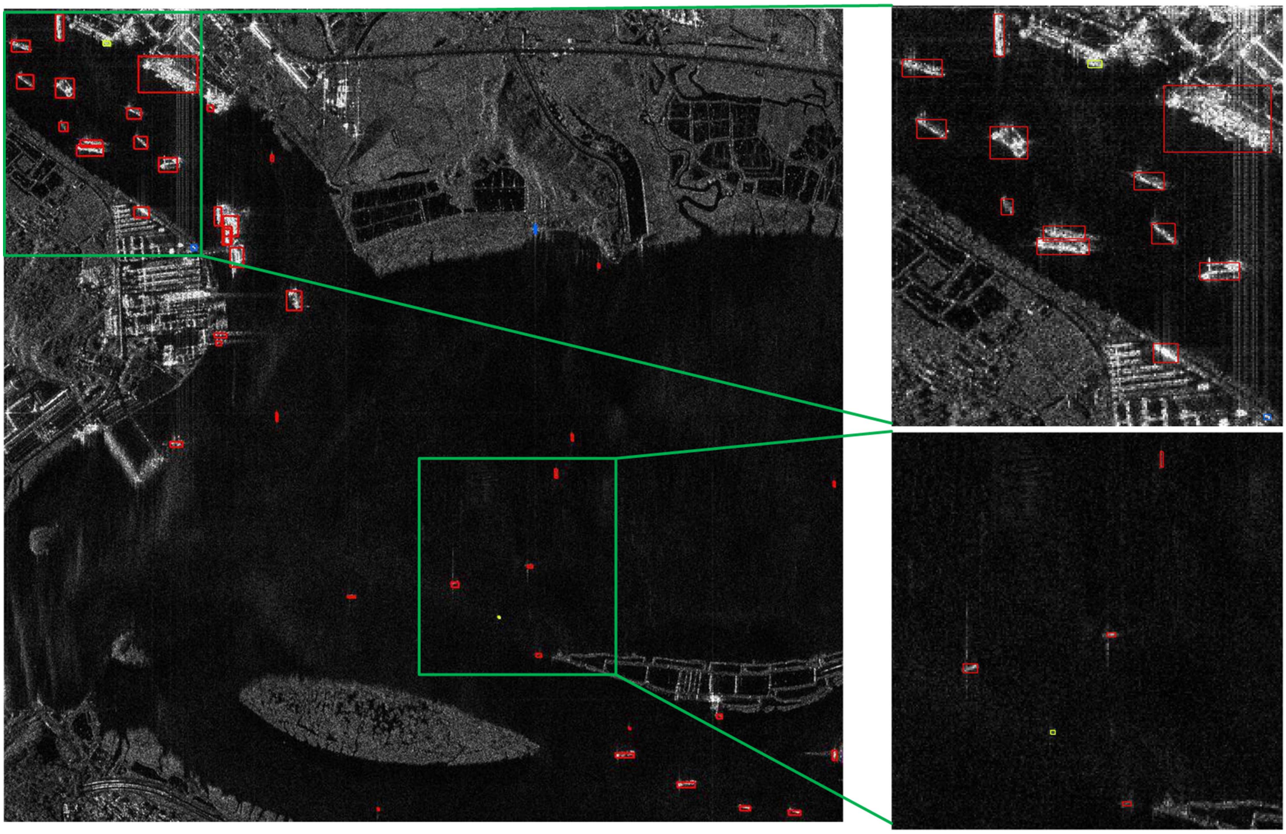


Fig. 13. Detection results on complex large-scale image (image 1).

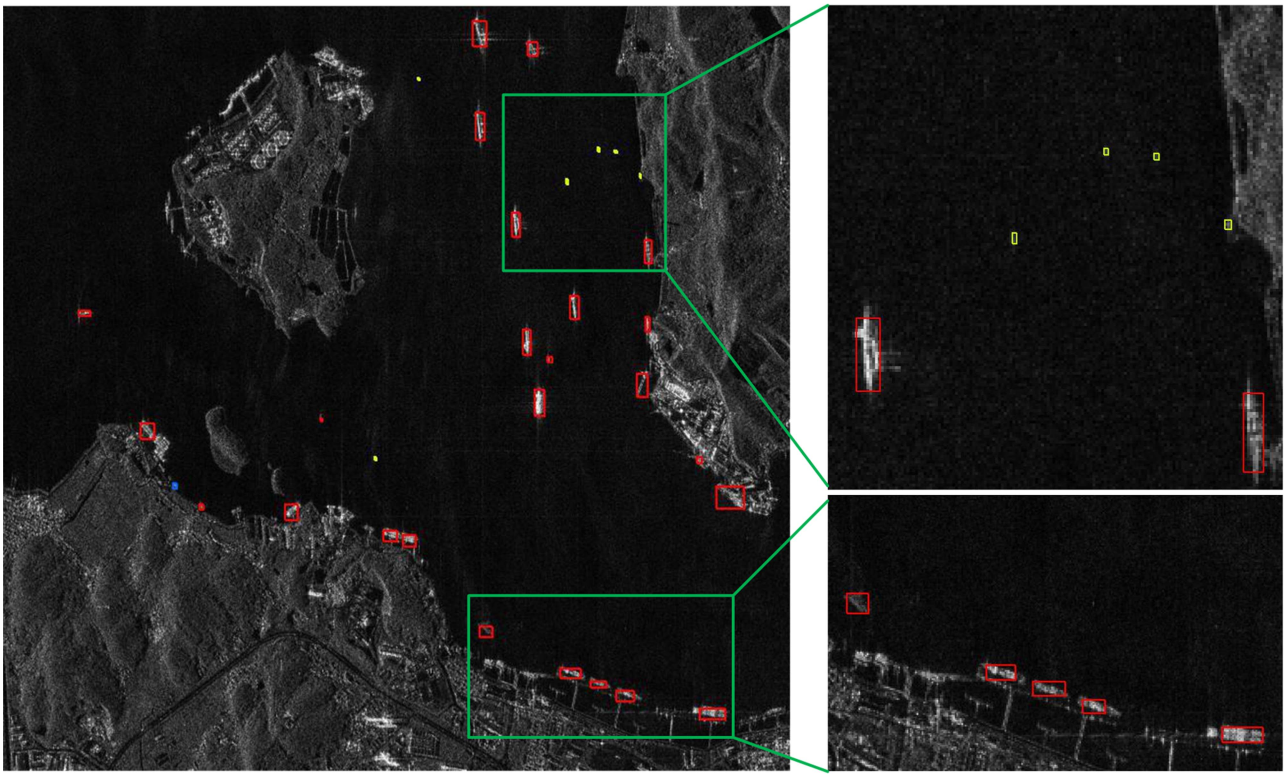


Fig. 14. Detection results on complex large-scale image (image 2).

TABLE VIII
Detection Results of Detectors on HRSID

Method	IoU = 0.5				IoU = 0.75				FPS
	AP_{50}	AP_{L}	AP_{M}	AP_{S}	AP_{75}	AP_{L}	AP_{M}	AP_{S}	
RetinaNet	54.68%	32.16%	43.70%	40.63%	30.54%	23.51%	26.74%	25.88%	39
CenterNet	54.59%	31.54%	43.38%	40.91%	11.51%	2.26%	6.92%	4.88%	78
Faster-RCNN	21.15%	13.15%	17.44%	15.05%	0.69%	0.05%	0.57%	0.51%	16
YOLOv3	67.35%	42.69%	60.05%	51.62%	23.46%	17.73%	22.43%	20.17%	61
YOLOv4	68.45%	44.71%	60.13%	52.22%	25.42%	19.06%	24.27%	20.48%	50
YOLOX	72.53%	50.59%	62.43%	57.66%	38.01%	30.50%	35.39%	31.81%	**95**
YOLOv7	66.88%	43.57%	59.93%	52.42%	30.99%	22.94%	26.83%	25.20%	43
SFF-YOLOX	**79.10%**	**57.15%**	**64.24%**	**62.29%**	**43.12%**	**33.69%**	**40.60%**	**36.21%**	62

two complex and large-scale SAR images. The results show that our SFF–YOLOX outperforms other mainstream deep-learning-based methods by a large margin, which can be applied in current ship detection and other vision tasks.

References

[1] C. Mao, L. Huang, Y. Xiao, F. He, and Y. Liu, “Target recognition of SAR image based on CN-GAN and CNN in complex environment,” *IEEE Access*, vol. 9, pp. 39608–39617, 2021.

[2] R. Chen, X. Li, and S. Li, “A lightweight CNN model for refining moving vehicle detection from satellite videos,” *IEEE Access*, vol. 8, pp. 221897–221917, 2020.

[3] A. M. Johansson, M. M. Espeseth, C. Brekke, and B. Holt, “Can mineral oil slicks be distinguished from newly formed sea ice using synthetic aperture radar?,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4996–5010, Aug. 2020.

[4] B. Brisco, M. Mahdianpari, and F. Mohammadimanesh, “Hybrid compact polarimetric SAR for environmental monitoring with the RADARSAT constellation mission,” *Remote Sens.*, vol. 12, no. 20, 2020, Art. no. 3283.

[5] T. Luti et al., “Land consumption monitoring with SAR data and multispectral indices,” *Remote Sens.*, vol. 13, no. 8, 2021, Art. no. 1586.

[6] 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.

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

[8] X. Leng, K. Ji, X. Xing, S. Zhou, and H. Zou, “Area ratio invariant feature group for ship detection in SAR imagery,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2376–2388, Jul. 2018.

[9] K. Sun, Y. Liang, X.-R. Ma, Y.-Y. Huai, and M.-D. Xing, “DSDet: A lightweight densely connected sparsely activated detector for ship target detection in high-resolution SAR images,” *Remote Sens.*, vol. 13, 2021, Art. no. 2743.

[10] R. Yang, Z. Pan, X. Jia, L. Zhang, and Y. Deng, “A novel CNN-based detector for ship detection based on rotatable bounding box in SAR images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1938–1958, Jan. 2021.

[11] M. Zhu et al., “ROS-Det: Arbitrary-oriented ship detection in high resolution optical remote sensing images via rotated one-stage detector,” *IEEE Access*, vol. 9, pp. 50209–50221, 2021.

[12] F. Ma, F. Gao, J. Wang, A. Hussain, and H. Zhou, “A novel biologically-inspired target detection method based on saliency analysis for synthetic aperture radar (SAR) imagery,” *Neurocomputing*, vol. 402, pp. 66–79, 2020.

[13] H. Qu, L. Shen, W. Guo, and J. Wang, “Ships detection in SAR images based on anchor-free model with mask guidance features,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 666–675, Dec. 2022.

[14] F. Gao, J. You, J. Wang, J. Sun, E. Yang, and H. Zhou, “A novel target detection method for SAR images based on shadow proposal and saliency analysis,” *Neurocomputing*, vol. 267, pp. 220–231, 2017.

[15] K. Eldhuset, “An automatic ship and ship wake detection system for spaceborne SAR images in coastal regions,” *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 4, pp. 1010–1019, Jul. 1996.

[16] C. Wackerman, K. Friedman, W. Pichel, P. Clemente-Colón, and X. Li, “Automatic detection of ships in RADARSAT-1 SAR imagery,” *Remote Sens.*, vol. 27, pp. 568–577, 2001.

[17] S. Gao and H. Liu, “Performance comparison of statistical models for characterizing sea clutter and ship CFAR detection in SAR images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7414–7430, Aug. 2022.

[18] G. Gao, Y. Luo, K. Ouyang, and S. Zhou, “Statistical modeling of PMA detector for ship detection in high-resolution dual-polarization SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4302–4313, Jul. 2016.

[19] G. Gao, K. Ouyang, Y. Luo, S. Liang, and S. Zhou, “Scheme of parameter estimation for generalized Gamma distribution and its application to ship detection in SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1812–1832, Mar. 2017.

[20] L. He, S. Yi, X. Mu, and L. Zhang, “Ship detection method based on Gabor filter and fast RCNN model in satellite images of sea,” in *Proc. 3rd Int. Conf. Comput. Sci. Appl. Eng.*, 2019, pp. 1–7.

[21] R. Wang et al., “An improved faster R-CNN based on MSER decision criterion for SAR image ship detection in harbor,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 1322–1325.

[22] X. Ke, X. Zhang, T. Zhang, J. Shi, and S. Wei, “SAR ship detection based on an improved faster R-CNN using deformable convolution,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 3565–3568.

[23] K. Sun, Y. Li, C. Li, Y. Liang, and M. Xing, “A two-step ship target detection method in high-resolution SAR image based on coarse-to-fine mechanism,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2811–2814.

[24] D. Kumar and X. Zhang, “Ship detection based on faster R-CNN in SAR imagery by anchor box optimization,” in *Proc. Int. Conf. Control, Automat. Inf. Sci.*, 2019, pp. 1–6.

[25] W. Liu, D. Anguelov, and D. Erhan, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[26] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.

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

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

[29] C. Wang, A. Bochkovskiy, and H. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7464–7475.

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

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

[32] T. 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.

[33] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6568–6577.

[34] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 642–656, 2019.

[35] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.

[36] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional onestage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

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

[38] M. Tan, R. Pang, and Q. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10778–10787.

[39] T. Miao et al., "An improved lightweight RetinaNet for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4667–4679, Jun. 2022.

[40] X. Yang, X. Zhang, N. Wang, and X. Gao, "A robust one-stage detector for multiscale ship detection with complex background in massive SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5217712.

[41] Y. Li, X. Lv, P. Huang, W. Xu, W. Tan, and Y. Dong, "SAR ship target detection based on improved YOLOv5s," in *Proc. Int. Conf. Control, Automat. Inf. Sci.*, 2021, pp. 354–358.

[42] 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.

[43] H.-Y. Guo, X. Yang, N.-N. Wang, and X.-B. Gao, "A CenterNet++ model for ship detection in SAR images," *Pattern Recognit.*, vol. 112, 2021, Art. no. 107787.

[44] F. Min and P. Liu, "Research on ship detection in the SAR image Algorithm based on improved SSD," in *Proc. 4th Int. Conf. Artif. Intell. Pattern Recognit.*, 2021, pp. 205–211.

[45] M. C. E. Rai, J.-H. Giraldo, M. Al-Saad, M. Darweech, and T. Bouwmans, "SemiSegSAR: A semi-supervised segmentation algorithm for ship SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jun. 2022, Art. no. 4510205.

[46] S. Chen, R. Zhan, W. Wang, and J. Zhang, "Domain adaptation for semi-supervised ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, May 2022, Art. no. 4507405.

[47] X.-Q. Li, D. Li, H.-Q. Liu, J. Wan, Z.-Y. Chen, and Q.-H. Liu, "A-BFPN: An attention-guided balanced feature pyramid network for SAR ship detection," *Remote Sens.*, vol. 14, 2022, Art. no. 3829.

[48] M. Zhao, J. Shi, and Y. Wang, "Orientation-aware feature fusion network for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 4504705.

[49] L.-M. Zhang, Y.-J. Liu, Q.-X. Guo, H.-Y. Yin, Y. Li, and P.-T. Du, "Ship detection in large-scale SAR images based on dense spatial attention and multi-level feature fusion," in *Proc. Assoc. Comput. Mach. Turing Award Celebration Conf.—China*, 2021, pp. 77–81.

[50] X. Wang and C. Chen, "Ship detection for complex background SAR images based on a multiscale variance weighted image entropy method," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 184–187, Feb. 2017.

[51] F. Xie, B. Lin, and Y. Liu, "Research on the coordinate attention mechanism fuse in a YOLOv5 deep learning detector for the SAR ship detection task," *Sensors*, vol. 22, no. 9, Apr. 2022, Art. no. 3370.

[52] F. Gao, Y. He, J. Wang, A. Hussain, and H. Zhou, "Anchor-free convolutional network with dense attention feature aggregation for ship detection in SAR images," *Remote Sens.*, vol. 12, no. 16, 2020, Art. no. 2619.

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

[54] Y.-Y. Wang, C. Wang, H. Zhang, Y.-B. Dong, and S.-S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 765.

[55] 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.

[56] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.

[57] M.-M. Cheng, N.-J. Mitra, X. Huang, and S.-M. Hu, "Salient shape: Group saliency in image collections," *Vis. Comput.*, vol. 30, pp. 443–453, 2014.

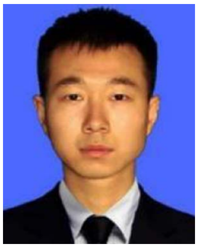
[58] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[59] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[60] C. Yang, L. Zhang, and H. Lu, "Graph-regularized saliency detection with convex-hull-based center prior," *IEEE Signal Process. Lett.*, vol. 20, no. 7, pp. 637–640, Jul. 2013.

[61] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[62] Q. Li, N. Miao, and X. Zhang, "Image recognition of maize disease based on asymmetric convolutional attention residual network and transfer learning," *Sci. Technol. Eng.*, vol. 21, pp. 6249–6256, 2021.



Yunlong Gao received the B.S. degree in computer science and technology and the M.S. degree in computer software and theory from the Jilin University, Changchun, China, in 2015 and 2018, respectively, and the Ph.D. degree in circuits and systems from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 2023.

His research interests include object detection and image processing technology.

Chuan Wu received the Ph.D. degree in mechatronic engineering from the Chinese Academy of Sciences, Beijing, China, in 2003.

He is currently a Research Fellow with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include target tracking and image processing technology.

Ming Ren received the B.S. and M.S. degrees in mechanical engineering from the Harbin Engineering University, Harbin, China, in 2017 and 2020, respectively.

His research interests include object detection and computational imaging technology.