An Improved Deep Neural Network for Small-Ship Detection in SAR Imagery

Boyi Hu and Hongxia Miao

***Abstract*—Ship detection by using remote-sensing images based on a synthetic aperture radar (SAR) plays an important role in managing water transportation and marine safety. However, complex background, a small-ship size, and low focus on small ships results in difficulties in feature extraction and low detection accuracy. This study proposes a new small SAR ship-detection network. First, a transformer-based dynamic sparse attention module is used to improve the focus and extraction of small-ship features. Second, the feature maps are fused with deep layers, and small target-friendly detection heads are used to improve the processing of global information in the network. Third, a more suitable fused loss function is used for small ships to ensure the multiscale detection capability. Experimental results on publicly available datasets, LS-SSDD_v1.0 and AIR-SARShip-1.0, show that the proposed method effectively improves the detection accuracy of small ships on SAR images without computational burden boost. Compared with other methods based on the convolutional neural network, the proposed method demonstrates the better multiscale detection performance.**

***Index Terms*—Convolutional neural network (CNN), ship detection, synthetic aperture radar (SAR), transformer.**

I. Introduction

SYNTHETIC aperture radar (SAR) plays an important role in real-time object detection because of its unique high-resolution imaging technology and its near-independence from the factors of weather and time. SAR-image-based ship detection provides timely warning and facilitates in rescue operations during emergencies at sea. This is critical for the maintenance of marine security and the reduction of accidents.

The development of the SAR allowed for the obtaining of many excellent high-resolution images for research. Studies have also proposed several SAR-based methods for ship detection. For example, the constant false alarm rate (CFAR) is a widely used conventional detection method. Liu et al. [1] proposed the optimization of CFAR for ship detection in polarimetric SAR (PolSAR) by using a quadratic programming approach with good accuracy and robustness. Then, scholars [2] proposed to simultaneously diagonalize two symmetric matrices by using a combination of eigenvalue decomposition and matrix iteration techniques to further improve the performance of CFAR for detecting PolSAR ships. However, most of the conventional detection methods are aimed at first eliminating regions that are unlikely to contain targets, selecting regions-of-interest (ROIs) that may contain targets, and then performing subsequent steps of identification and classification. The selected ROI usually contains a large amount of cluttered background, causing missed detections and a high false-alarm rate [3]. These algorithms are dependent on manual feature annotation and have poor robustness.

In recent years, with advancements in faster processors, several deep-learning-based detection methods have been developed. Deep-learning methods have made breakthroughs in many fields owing to their excellent model structure and rich parameters. A convolutional neural network (CNN) is a classical deep neural network. Guan et al. [4] proposed a new deep learning model that effectively extracts image features and enhances the representation of key features by utilizing residual networks and attention mechanisms after image preprocessing. In addition, studies have proposed and applied many excellent CNN-based detectors, capable of maintaining robustness in more complex scenarios [5], [6], [7], [8], [9] with mostly satisfactory results. These detectors integrate feature extraction and classification into a framework that uses existing data for feature extraction and learning, and finally new data can be detected and identified based on the learned features. Although the training process is time-consuming, the trained model can detect new data in a short time and adapt to large-scale datasets. Furthermore, it meets the requirements of real-time detection and demonstrates stronger robustness than conventional detection methods [10]. The detectors can be broadly classified into the following two types: single- and two-stage detectors. The single-stage detectors comprise the single-shot multibox detector (SSD) [11] and you only look once (YOLO) series detectors [12], which display satisfactory performance and a rapid detection speed. Chen et al. [13] proposed an attention-combined single-stage detection network for target localization in complex scenarios. Representative two-stage detectors include faster Region-CNN (R-CNN) [14] and Cascade R-CNN [15], which have the advantage of a more accurate prediction of the location of a bounding box. Zhao et al. [16] proposed an attention perception pyramid network to enhance the relationship between nonlocal features and thus improve the multiscale detection capability. Gao et al. [17] proposed cross-modal domain transfer learning to improve the generalization of SAR target-recognition models on real data by performing feature alignment and knowledge

Manuscript received 11 October 2023; revised 6 December 2023; accepted 22 December 2023. Date of publication 27 December 2023; date of current version 10 January 2024. This work was supposed in part by the National Natural Science Foundation of China under Grant 62201078. *(Corresponding author: Hongxia Miao.)*

The authors are with the State Key of Networking and Switching Technology, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: huboyi@bupt.edu.cn; hongxia_miao@bupt.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3347660

migration between synthetic and real data. In [18], a dual prototype structure was proposed and a lightweight adaptive task attention module was introduced to adaptively enhance the ability to focus on key features. However, convolutional attention tends to filter out small target features, making the detection of small SAR ships less effective.

Other deep learning networks have been proposed. In 2017, Vaswani et al. [19] proposed a revolutionary network architecture that uses attention mechanism to connect encoders and decoders, namely a transformer. The excellent performance of the transformer has attracted considerable research attention, and it has been recently introduced into object detection tasks. Dosovitskiy et al. [20] used image patch sequences directly as input to the transformer, terming it a vision transformer (ViT). The self-attention mechanism enables the transformer module to better capture global information. Liu et al. [21] proposed a new ViT, called swin transformer (or SwTR in this article) that can be used as a backbone network in vision tasks. The SwTR is a breakthrough that uses a form of shifted window to calculate the attention between pixels and preserves the interaction between neighboring windows of ViT, allowing the transformer to better learn global information and improve multiscale feature detection. Compared to ViT, the nonpower exponent in SwTR makes it significantly less complex to compute [21]. The corresponding sampling process uses different sampling multipliers, significantly improving its detection performance in small target datasets than that in Vit. Some scholars have introduced SwTR into deep learning networks and achieved good experimental results, further verifying the effectiveness of its structure [22], [23].

In an image, small targets usually occupy fewer pixels. In addition, distinguishing the size of a target solely based on its pixel size is not entirely reasonable. In detection tasks, the size of targets is relative and should not be distinguished simply by the size of the pixel value but also in relation to the resolution of the image in which it is located. In the detection network, "small" targets should refer to those that occupy a small proportion of image pixels, resulting in insufficient target information and difficultly in object detection owing to larger amount of interference information than target information. The term "small target" does not simply refer to a target with "few pixels." In addition, in a multiscale environment, a "small target" can be a target with relatively fewer pixels. Therefore, in this article, the term "small ship" refers to a ship that is "relatively small." Owing to the special SAR imaging mechanism, the smaller size of a ship implies weaker scattering intensity, and the imaged target not only carries less characteristic information but is also easily disturbed by background clutter. Therefore, small-ship detection remains a major issue in SAR-based image detection. Jiao et al. [24] proposed a multiscale dense connection based on a Faster R-CNN, which can solve the problem of the detection of small ships. Yang et al. [25] proposed a detection network with an increased receptive field and the introduction of a coordinate attention module to enhance the multiscale detection of large-scale images. In general, these algorithms have two main problems in detecting small SAR ships. First, the detection layer with a large receptive field tends to ignore the features of small ships when extracting feature information. Second, when the background is more complex, the clutter further interferes with the detection of small ships, resulting in a high rate of missed or false alarms. Several improved methods have been proposed to alleviate these problems. For example, Li et al. [26] proposed a YOLOSR-IST network by adding SwTR blocks as detection heads to YOLO, thus reducing the false-detection rate and false alarms. Cui et al. [27] proposed a spatial shuffle-group enhanced attention module based on CenterNet, which reduces the missed detection of small ships. Gong et al. [28] proposed an improved feature pyramid and sample enhancement to effectively extract image features and increased the number of training samples to improve the detection performance of small ships. Sun et al. [29] combined feature fusion and cross-layer connection techniques to improve the response speed and localization accuracy of small targets by effectively integrating image features and establishing cross-layer connections. Zhou et al. [30] proposed the sidelobe-aware mechanism to reduce the effects of strong scattering points, while improving the neck structure and loss function to enhance the accuracy of small-ship detection.

To further reduce the false alarms and missing targets, this study proposes an improved network, which includes a dynamic sparse attention module, an improved feature extraction neck structure, and a fused loss function. Specifically, we introduce a transformer-based attention module at the end of the backbone, and it only contains GPU-friendly matrix multiplication. In addition, the module fully exploits the feature information of small ships without a large amount of computational load. Then, the shallow feature map is fused with the deeper features, thus enhancing the global feature information and reducing feature loss from convolution and downsampling. The detection head after multiple downsamplings easily ignores the small target features. To alleviate this problem, we introduce the SwTR into the detection head. Not only does SwTR not add significantly to network complexity but it also improves the network's ability to process background images. Third, the normalized Gaussian Wasserstein distance (NWD) [31] loss is fused with the intersection over union (IoU), thus reducing the sensitivity of IoU and improving the network's ability to regress on small ships.

In brief, the main contributions of this article are summarized as follows.

1) To allow the model to better focus on the target ROI, a transformer-based dynamic sparse attention module is added between the backbone and neck of the model.
2) To enhance the global information of the target, improve the detection capability of small ships, and adapt the model to multiscale detection tasks, characteristic maps of the superficial layers of the backbone are introduced into the deeper layers of the neck, and SwTR is introduced in the third detection head.
3) To better regress small ships and improve the multiscale detection ability of the model, NWD is introduced into the loss function, and a new calculation formula of the loss function is obtained.

The rest of this article is organized as follows. Section II specifies the proposed methodology. Section III validates and analyzes the proposed method by proving a comparison of the experimental results. Finally, Section IV concludes this article.

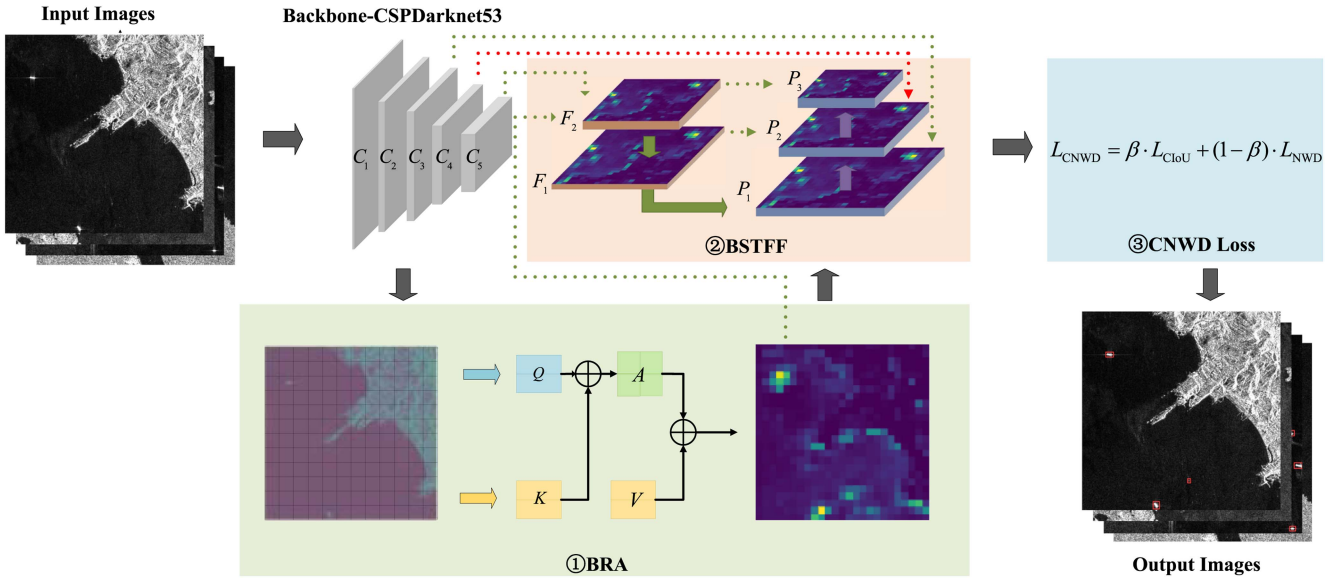

Fig. 1. Overall structure of the proposed network. BRA represents the BRA module, BSTFF represents the improved neck, and represents the fused loss function.

II. Methodology

The network structure of the proposed model is illustrated in Fig. 1. First, a dynamic sparse attention module, bilevel routing attention (BRA), is added at the end of the backbone. Then, based on the feature pyramid network (FPN) and path aggregation network (PAN) structures, the C4 feature layer is introduced into the neck and the third detection head is replaced with the SwTR. In addition, a new fused loss function is introduced in the final regression stage.

A. Dynamic Sparse Attention Mechanism

1) Attention Mechanism in a Transformer: The attention mechanism can effectively capture long-range dependencies, and significantly improve model performance. Convolution is a local operator, whereas attention focuses on a global feel [19]. However, conventional attention tends to ignore small target features, while the self-attention mechanism makes transformer attention friendlier to small targets. The attention function transforms each query into a weighted sum of values, and the calculation of the weights is termed as a normalized dot product between the query and corresponding key.

Taking queries $\mathbf{Q} \in \mathbb{R}^{N_q \times C}$, keys $\mathbf{K} \in \mathbb{R}^{N_k \times C}$, and values $\mathbf{V} \in \mathbb{R}^{N_v \times C}$ as input, the output of the attention function is calculated as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}}\right)\mathbf{V} \tag{1}$$

where C is the number of channels, $\sqrt{C}$ is a scalar factor that avoids gradient vanishing [19], softmax is a normalized exponential function, the height and width of the feature map are H and W, respectively, and N_k and N_v represent N of keys and values, respectively, and $N = H \times W$.

Moreover, multiheaded self-attention (MHSA) was used in the transformer. Self-attention implies that the values of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ demonstrate a linear projection from $\mathbf{X} \in \mathbb{R}^{N \times C}$. Here, $\mathbf{X}$ is a spatially flattened feature map in ViT. The multiple head infers to the fact that the output will be divided into h heads, and the projection weights in different blocks are independent. Formally

$$\text{MHSA}(\mathbf{X}) = \text{concat}(\mathbf{head}_0, \mathbf{head}_1, \ldots, \mathbf{head}_h)\mathbf{W}^\circ \tag{2}$$

$$\mathbf{head}_i = \text{Attention}(\mathbf{X}\mathbf{W}_i^q, \mathbf{X}\mathbf{W}_i^k, \mathbf{X}\mathbf{W}_i^v) \tag{3}$$

where $i = 0, 1, 2, \ldots, h$, $\mathbf{head}_i \in \mathbb{R}^{N \times \frac{C}{h}}$ is the output of the ith attention head, concat is a function that concatenates feature maps, and $\mathbf{W}_i^q, \mathbf{W}_i^k, \mathbf{W}_i^v \in \mathbb{R}^{C \times \frac{C}{h}}$ are the corresponding input projection weights. All the heads are combined by a linear transformation with a weight matrix $\mathbf{W}^\circ \in \mathbb{R}^{C \times C}$.

However, the MHSA requires a large computational burden. The global view of perception inevitably comes with a corresponding cost, i.e., the calculation of the affinity of pairs of tokens at all spatial locations brings unavoidable complex computations and burden computational resources. Thus, to alleviate this problem, many researchers have tried to reduce the attentional operations by restricting them within local windows, axial stripes, or dilated windows [32], [33], [34]. However, different semantic regions focus on significantly different key-value pairs, and forcing all queries to focus on the same set of tokens is suboptimal. In addition, as bilevel routing can be used to achieve more flexible computational resource allocation, BRA has been proposed earlier [35].

2) Bilevel Routing Attention: As shown from (2), in the case of h queries, each query focuses on h key-value pairs. Such a structure inevitably increases computational complexity and introduces serious scalability issues in terms of the spatial resolution of inputs.

In this study, we introduce a new sparse attention module, namely BRA, the structure of which is shown in Fig. 2. This

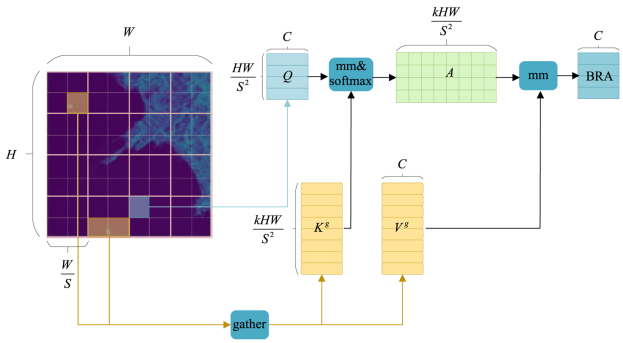

Fig. 2. Overall structure of BRA.

module causes each query to focus on a small set of semantically most relevant key-value pairs. In this way, attention becomes a dynamic and query-aware sparsity mechanism. To efficiently locate valuable key-value pairs, an area-to-area routing method is used. The core of this approach is to filter out irrelevant key-value pairs at the coarse-grained level to facilitate the application of attention mechanisms in subsequent regions. For convenience, the single input and single header cases are discussed. Given a two-dimensional (2-D) input feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, it is first divided into $S \times S$ nonoverlapping regions, each containing HW/S^2 feature vectors, and $\mathbf{X}$ is reshaped into $\mathbf{X}^r \in \mathbb{R}^{S^2\times \frac{HW}{S^2}\times C}$. Then, a linear projection is used to derive $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{S^2\times \frac{HW}{S^2}\times C}$

$$\mathbf{Q} = \mathbf{X}^r\mathbf{W}^q, \mathbf{K} = \mathbf{X}^r\mathbf{W}^k, \mathbf{V} = \mathbf{X}^r\mathbf{W}^v. \tag{4}$$

In addition, a region-level affinity graph is constructed and pruned, with each node retaining only the first k connections, so that attending regions can be formed. Specifically, the required $\mathbf{Q}^r, \mathbf{K}^r \in \mathbb{R}^{S^2\times C}$ must be obtained by applying the average of each region. Then, $\mathbf{Q}^r$ and $\mathbf{K}^r$ matrices are multiplied to obtain the adjacency matrix of the region-to-region affinity graph, $\mathbf{A}^r \in \mathbb{R}^{S^2\times S^2}$

$$\mathbf{A}^r = \mathbf{Q}^r(\mathbf{K}^r)^T. \tag{5}$$

Then, retain only the top-k links of each region for pruning the affinity graph. Specifically, a routing index matrix, $\mathbf{I}^r \in \mathbb{N}^{S^2\times k}$ is derived as

$$\mathbf{I}^r = \text{topkIndex}(\mathbf{A}^r) \tag{6}$$

where topkIndex represents the top-k links.

Then, to apply token-to-token attention. With a region-to-region routing index, $\mathbf{I}^r$, a fine-grained level of attention can be performed. Modern GPUs can rapidly load dozens of contiguous bytes simultaneously. The process of implementing a focus on these routing regions, which are expected to be scattered throughout the feature graph, is not very complicated. As such, the tensor of key values is collected as follows:

$$\mathbf{K}^g = \text{gather}(\mathbf{K}, \mathbf{I}^r), \mathbf{V}^g = \text{gather}(\mathbf{V}, \mathbf{I}^r) \tag{7}$$

where $\mathbf{K}^g, \mathbf{V}^g \in \mathbb{R}^{S^2\times \frac{kHW}{S^2}\times C}$ denotes the key-value pair tensor and gather is a function that collects values along the axis specified by dim. Attention functions can be applied to these key-value pairs as follows:

$$\mathbf{O}_{\text{BRA}} = \text{Attention}(\mathbf{Q}, \mathbf{K}^g, \mathbf{V}^g) + \text{LCE}(\mathbf{V}) \tag{8}$$

where LCE denotes a local context enhancement term [36].

BRA involves only hardware-friendly dense matrix multiplication when collecting key-value tokens, enabling a good performance, computational tradeoff by focusing on important key-value pairs in each query in a content-aware manner.

B. Bidirectional and SwTR Feature Fusion Neck

The ship size differs in SAR images. Especially, in high-resolution SAR images, networks cannot fully extract the ship features of different sizes, and thus, the features of small ships are easily lost in this process. Insufficient feature extraction of small targets results in information loss, which adversely affects the subsequent detection.

Based on the Bi-FPN structure [37], fully utilizing a feature map for multiscale feature fusion can alleviate the information loss in the backbone. As such, the feature map obtained by the detection head has more useful features and stronger semantic information. In addition, to avoid the heavy computational burden caused by the increase in network complexity, we simply performed a deep fusion for the C4 layer information, as shown in Fig. 4. The neck is based on the backbone as the input features and FPN-PAN as the structure. To enhance the target feature information and fully retain the small-ship features, we concat the C4 layer with the intermediate detection head to fuse the detection with the multiscale semantic information.

The end of the network comprised the detection head developed by gathering all the features of the network. Next, regression prediction is performed using the detection heads based on the predefined size of the bounding box. However, when detecting small targets, the detection heads can easily ignore some feature relationships between small targets and the background after multiple downsampling; this, is not conducive to the network's perception of all images. In some scenarios, where the background is highly variable, this situation decreases the robustness and increases the false detection rate of the model. To improve the network's understanding of the whole image and allow it to better process images, we introduce SwTR. A SwTR block is a basic computational unit composed of a shift-window-based multihead self-attention (MSA) module, as shown in Fig. 3. As shown, the window MSA (W-MSA) and shifted W-MSA denote self-attention, based on division configurations of regular and shift windows, respectively. In addition, the entire structure contains multilayer perceptron (MLP) modules and is preceded by a layer-norm layer for both MSA and MLP modules. Each module is followed by a residual connection. Two consecutive submodules form an SwTR block.

The blind introduction of SwTR could bring about a large computational overhead, and the final detection results may not be satisfactory. Because of the characteristics of the overall network structure, the third detection head loses the most amount of feature information after multiple downsampling. Therefore, as shown in Fig. 4, we use the SwTR module instead of the third *C*3 head with the *C*3-SwTR module. In summary, the new

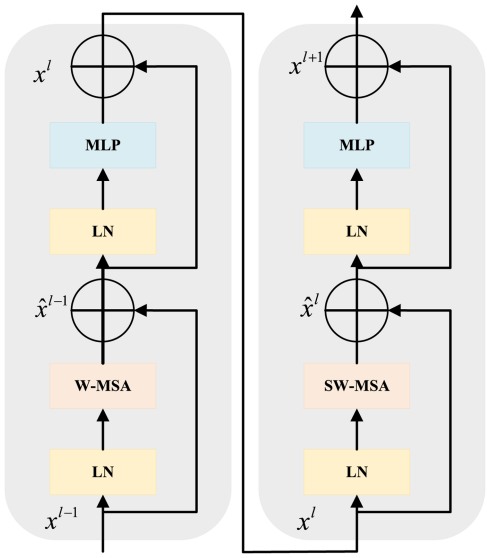

Fig. 3. Structure diagram of the SwTR block.

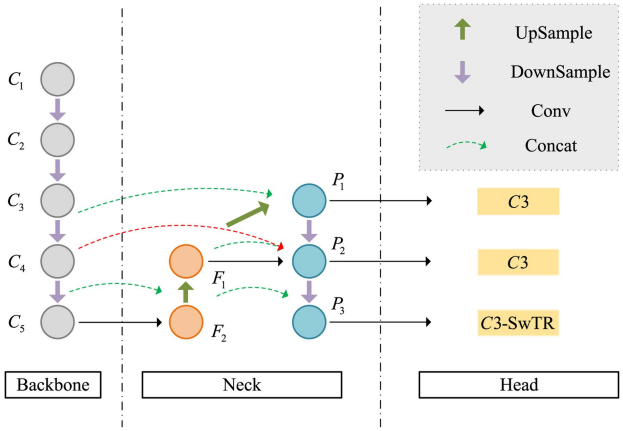

Fig. 4. Structure diagram of the proposed BSTFF.

neck structure first considers the rich semantic information in the backbone to enable the detection head to fully perform feature fusion. Then, the third detection head is replaced to improve the ability to sense global features for multiscale information detection.

C. Complete NWD Loss

1) IoU: The target detection process terminates with the regression of the prediction frame, and the metrics used by many advanced detectors are based on IoU, which is an important evaluation metric for existing loss functions. In simple terms, IoU is used to measure the degree of overlap between the detection and target boxes

$$\mathrm{IoU}(B_a, B_b) = \frac{|B_a \cap B_b|}{|B_a \cup B_b|} \tag{9}$$

where B_a and B_b represent the prediction and real boxes, respectively. The loss function is defined as

$$L_{\mathrm{IoU}} = 1 - \mathrm{IoU}(B_a, B_b). \tag{10}$$

However, the IoU demonstrates a satisfactory performance only when the two boxes overlap. For some extreme cases, distance IoU (DIoU) [38] was proposed; it adds a penalty term that represents the distance between the centroids of the prediction and true boxes. The loss function is defined as

$$L_{\mathrm{DIoU}} = 1 - \mathrm{IoU}(B_a, B_b) + \frac{\rho^2(d, d^{\mathrm{gt}})}{c^2} \tag{11}$$

where d and d^{gt} denote the centroids of the prediction and true boxes, respectively, $\rho(\cdot)$ is the Euclidean distance, and c is the diagonal length of the smallest box that contains both boxes.

Ships of different aspect ratios (width-to-height ratio) could significantly impact the loss function. The complete IoU (CIoU) [38] was proposed to add the factor of the aspect ratio to the DIoU, and the loss function is denoted as

$$L_{\mathrm{CIoU}} = 1 - \mathrm{IoU}(B_a, B_b) + \frac{\rho^2(d, d^{\mathrm{gt}})}{c^2} + \alpha\upsilon \tag{12}$$

where α is the loss tradeoff parameter and υ measures the similarity of the aspect ratio

$$\alpha = \frac{\upsilon}{(1 - \mathrm{IoU}(B_a, B_b)) + \upsilon} \tag{13}$$

$$\upsilon = \frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}} - \arctan\frac{w}{h}\right)^2 \tag{14}$$

where w and h are the width and height of the bounding box, respectively, and w^{gt} and h^{gt} are the width and height of the ground-truth (GT) box, respectively.

The IoU is sensitive to the position deviation of the target [39]. Especially, for small targets, a deviation in pixel position, even a small one, can result in a dramatic deterioration in the detection performance for anchor-based detectors. Fig. 5 shows that the sensitivity of IoU to targets of different scales varies greatly. This difference could be attributed to the fact that the location of the bounding box can only vary discretely. For small targets (8×8 pixels), a small deviation in position causes a sharp drop in the IoU metric (from 0.62 to 0.08), which further results in an inaccurate label assignment strategy. However, for normal-sized targets (40×40 pixels), changes in the IoU are not particularly noticeable at the same deviation and will not fluctuate much for the final detection results. The sensitivity of IoU on small targets causes positional deviations that flip the anchor labels, complicating network convergence.

2) NWD: Dynamic assignment strategies, such as adaptive training sample selection, can form IoU thresholds to assign pos/neg labels based on the statistical characteristics of the target. However, the sensitivity of IoU is not conducive to this process, making it difficult for the detector to possess high-quality pos/neg samples for feature learning. To alleviate this problem, a new metric, NWD [31] is proposed to replace the IoU with the Wasserstein distance, and the similarity of the bounding box is represented by this distance. The NWD does not affect the measurement of the distribution similarity for a small or no overlap. In addition, NWD is independent of the scale of the target and has an advantage over IoU in terms of the measurement of small targets. Owing to its computational independence, NWD can be used not only for anchor-based

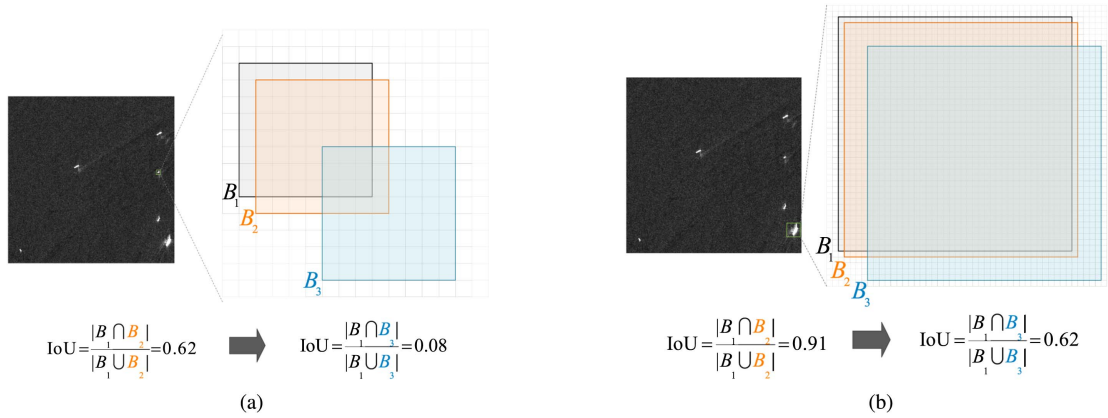

Fig. 5. IoU sensitivity for small and normal-sized ships. Each grid represents one pixel, box B_1 represents the real box, and B_2 and B_3 represent the predicted boxes deviated by 1 and 4 pixels, respectively. (a) Small ships. (b) Normal ships.

single-stage detectors, but also as a good alternative to IoU in multistage detectors.

The Wasserstein distance is calculated using the optimal transport theory, in which for two 2-D Gaussian distributions, $\mu_1 = \mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ and $\mu_2 = \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$, and the second-order Wasserstein distance between μ_1 and μ_2 is defined as

$$W_2^2(\mu_1, \mu_2) = \|\mathbf{m}_1 - \mathbf{m}_2\|_2^2 + \mathbf{Tr}\left(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 - 2\left(\mathbf{\Sigma}_2^{\frac{1}{2}}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{\frac{1}{2}}\right)^{\frac{1}{2}}\right). \quad (15)$$

This can be simplified as

$$W_2^2(\mu_1, \mu_2) = \|\mathbf{m}_1 - \mathbf{m}_2\|_2^2 + \left\|\mathbf{\Sigma}_1^{\frac{1}{2}} - \mathbf{\Sigma}_2^{\frac{1}{2}}\right\|_F^2 \quad (16)$$

where $\|\cdot\|_F$ is the Frobenius norm.

For Gaussian distributions $\mathcal{N}_a$ and $\mathcal{N}_b$, the models are divided from bounding boxes, $B_a = (cx_a, cy_a, w_a, h_a)$ and $B_b = (cx_b, cy_b, w_b, h_b)$, respectively. Equation (16) can be further simplified as

$$W_2^2(\mathcal{N}_a, \mathcal{N}_b) = \left\|\left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2}\right]^{\mathrm{T}}, \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2}\right]^{\mathrm{T}}\right)\right\|_2^2. \quad (17)$$

However, $W_2^2(\mathcal{N}_a, \mathcal{N}_b)$ is a distance metric that must be normalized using an exponential form to obtain a new metric, namely NWD

$$\mathrm{NWD}(\mathcal{N}_a, \mathcal{N}_b) = \exp\left(-\frac{\sqrt{W_2^2(\mathcal{N}_a, \mathcal{N}_b)}}{D}\right) \quad (18)$$

where D is a constant that is closely related to the dataset. Based on the results obtained from several experiments, D can be set to the average absolute size of the dataset to obtain the best performance. In addition, D is robust within a certain range.

NWD performs well in detecting small targets, such as scale-invariant insensitivity to position deviation, and can measure the similarity between bounding boxes when they do not overlap. In addition, NWD can be integrated into any anchor-based detector to replace the IoU.

IoU-Loss is not suitable for small targets detectors, because it cannot provide gradients for the optimized network when there is no overlap between the predicted bounding box, B_p, and the GT box, B_{gt}, i.e., ($|B_p \cap B_{gt}| = 0$) or when box B_p completely contains box B_{gt} or vice versa, i.e., ($|B_p \cap B_{gt}| = B_p$ or B_{gt}) [31]. Both of these are very common scenarios in small target detection. Although CIoU and DIoU can be used to manage these two cases, their essence is still based on IoU. CIoU and DIoU still encounter the problem of positional deviation sensitivity. To deal with this problem, a new NWD loss function is proposed

$$L_{\mathrm{NWD}} = 1 - \mathrm{NWD}(\mathcal{N}_p, \mathcal{N}_{gt}) \quad (19)$$

where $\mathcal{N}_p$ is the Gaussian distribution model of box B_p and $\mathcal{N}_{gt}$ is the model of box B_{gt}. The NWD-based loss function can still provide gradients normally, while avoiding gradient vanishing even in the cases of $|B_p \cap B_{gt}| = 0$ and $|B_p \cap B_{gt}| = B_p$ or B_{gt}.

To improve the performance of small-ship detection while retaining the robustness of normal-scale ship detection, the NWD and IoU are combined in some ratio to form a new loss function [40], which fully preserves the multiscale detection capability. The experimental results show an improvement in the detector performance after the addition of the NWD.

3) CIoU and NWD Fused Loss: For the CIoU loss function in (12), the aspect ratio is ambiguous when the detection target is relatively small; this may limit the loss function and affect the regression. To fully improve the detection of small ships, we introduce a fusion-improved loss function (CNWD) that fuses NWD and CIoU in some ratio, and it is calculated as follows:

$$L_{\mathrm{CNWD}} = \beta \cdot L_{\mathrm{CIoU}} + (1 - \beta) \cdot L_{\mathrm{NWD}} \quad (20)$$

where $\beta \in (0, 1)$ is a scaling factor that can be flexibly adjusted to fully utilize the fusion loss function based on the specifics of the target size of ships in the dataset. This loss function is applicable to various detection networks.

The loss function theory is applied to the LS-SSDD_v1.0 and AIR-SARShip-1.0 datasets. This loss fully integrates the

TABLE I
DETAILS OF DATASET

Information	LS-SSDD_v1.0	AIR-SARShip-1.0
Satellite	Sentinel-1	Gaofen-3
Image Size(pixel)	24000×16000	3000×3000
Resolution	5m×20m	1m,3m
Polarization	VV,VH	Single
Image number	15	31
Ship number	6015	461

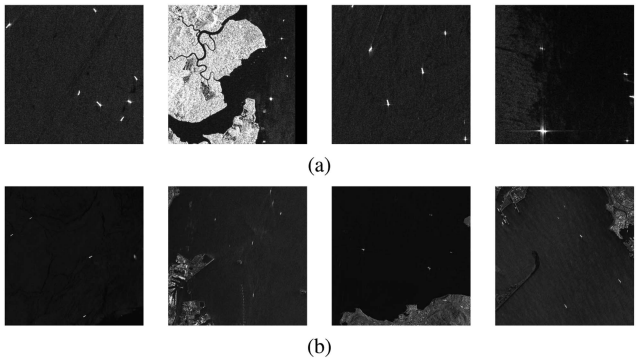

Fig. 6. Sample images from different datasets. (a) LS-SSDD_v1.0. (b) AIR-SARShip-1.0.

advantages of NWD, while retaining the IoU loss function for better detection of ships of different scales.

III. EXPERIMENTS

The proposed method is validated using the LS-SSDD_v1.0 and AIR-SARShip-1.0 datasets. This section first provides a brief description of the datasets used for the experiments, followed by a description of the experimental environment. Then, the evaluation metrics of the experiments are introduced, followed by ablation and comparison experiments to fully analyze and validate the proposed method.

A. Datasets and Environment Settings

1) Datasets: Information on the two datasets used is shown in Table I. As shown, the LS-SSDD_v1.0 [41] dataset is a large-scene small-ship dataset taken by Sentinel-1. It contains 15 large, 24 000 × 16 000 pixel images of scenes with a resolution of 5 m × 20 m. These images are subdivided into 9000 subimages of size 800 × 800 pixels, containing only one category, ships. The 9000 images are further divided into a training set and a validation set, containing 6000 and 3000 images, respectively. The small size of the ship targets complicates the detection tasks.

The AIR-SARShip-1.0 [42] dataset is obtained from the GaoFen-3 satellite, and it comprises 31 high-pixel SAR ship images with a resolution from 1 to 3 m. Spotlight and streak maps are used as the imaging modes. The training and test sets comprise 21 and 10 images, respectively. The total number of ship targets is 461. To train the model better, we augmented the data with random flipping, panning, and cropping. Finally, we obtain 1281 and 310 training and test images, respectively. The example images of these two datasets are shown in Fig. 6.

TABLE II
ENVIRONMENT CONFIGURATION

Project	Configuration/Parameter
CPU	Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz
RAM	64GB
GPU	NVIDIA Quadro RTX4000
Python	Python 3.8
Framework	torch 1.10.1/torchvision 0.11.2
Environment	Quadro RTX 536.25/CUDA 11.3/CUDNN 8.2.1

2) Settings: For a fair comparison, we conducted experiments on the same machine, and all models did not use pretrained weights. Pytorch is used the framework for the whole experiment, and the server configuration is Intel(R) Xeon(R) Silver 4216 CPU and NVIDIA RTX 4000 GPU, running on torch-1.10.1, CUDA-11.3, and CUDNN-8.2.1. The hardware information is given in Table II. The model uses stochastic gradient descent for training, with a learning rate of 0.001, momentum of 0.937, and decay rate of 5e−4. The anchors are automatically generated using K-means clustering. We set the batch-size to 16 and epochs to 200. The rest of the parameters are kept to the default values.

B. Evaluation Metrics

The performance evaluation metrics of the model are precision, recall, and average precision. Precision refers to the percentage of ships detected correctly to all detected ships, and it is defined as

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{21}$$

where true positives (TP) represent the number of correct detections by the model and false positives (FP) represent the number of incorrect detections (also called false alarms).

Recall is the percentage of correctly detected ships to GT and is denoted as

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{22}$$

where false negatives (FN) represent the number of missed detections.

Precision–recall (PR) is the curve formed by precision and recall, and average precision (AP) is the area formed by this

TABLE III
RESULTS FOR LS-SSDD_V1.0 DATASET

BRA	BSTFF	CNWD	mAP$^{\%}$	AP$^{\%}_{50}$	AP$^{\%}_{50:95}$	FPS	Params$^{\times 10^6}$
-	-	-	75.6	75.0	28.5	45.66	7.01
✓	-	-	76.3	75.7	29.1	44.44	8.07
✓	✓	-	77.5	76.9	**29.1**	41.32	8.27
✓	✓	✓	**78.3**	**77.5**	28.9	**47.85**	8.27

The bold value means the best performance.

TABLE IV
RESULTS FOR AIR-SARSHIP-1.0 DATASET

BRA	BSTFF	CNWD	mAP$^{\%}$	AP$^{\%}_{50}$	AP$^{\%}_{50:95}$	FPS	Params$^{\times 10^6}$
-	-	-	63.6	62.0	25.9	**16.03**	7.01
✓	-	-	64.3	62.1	26.1	15.37	8.07
✓	✓	-	67.0	63.7	26.5	14.22	8.27
✓	✓	✓	**67.3**	**64.1**	**27.1**	14.51	8.27

The bold value means the best performance.

curve, which is defined as

$$\mathrm{AP} = \int_0^1 \mathrm{P(R)}d\mathrm{R}. \tag{23}$$

In addition, IoU represents the ratio of the intersection and concatenation of the predicted and real boxes, and it is calculated as

$$\mathrm{IoU} = \frac{S_\cap}{S_\cup} \tag{24}$$

where $S_\cap$ represents the area where the two boxes overlap and $S_\cup$ represents the area of union.

In this study, the evaluation metrics used are **mAP** based on not only PASCAL VOC but also on COCO evaluation metrics to better assess the detection performance of ships at different scales. In both datasets, areas smaller than 32^2 pixels are small ships, areas between 32^2 and 96^2 pixels are medium-pixel ships, and areas larger than 96^2 pixels are large ships. Their corresponding evaluation metrics are represented as AP_s, AP_m, and AP_l. The value of AP differs for different IoU scores. In addition, AP_{50} in the COCO metrics represents the AP value when the IoU = 0.5, while $\mathrm{AP}_{50:95}$ represents the average value of AP for IoU = 0.5–0.95. The value of $\mathrm{AP}_{50:95}$ has more stringent evaluation criteria. In addition, AR_{100} is the recall when the maximum target detection frame is 100. Furthermore, FPS and parameters are used to evaluate the additional detection speed and model computational complexity resulting from the proposed method.

C. Ablation Experiment

1) Efficacy Analysis of BRA: The impact of the BRA module is first analyzed, and YOLOV5s is used for the baseline model. The results of the ablation experiments on the LS-SSDD_v1.0 dataset are presented in Table III. As shown, the inclusion of the BRA module improves the detection performance. The mAP, AP_{50}, and $\mathrm{AP}_{50:95}$ are improved by 0.7%, 0.7%, and 0.6%, respectively. Moreover, the number of model parameters is only improved by 1 M and the FPS is only decreased by 1.22, indicating that the BRA improves the accuracy without imposing too much computational burden on the model. As mentioned earlier, the blind introduction of transformer-based self-attention can significantly increase the computational burden. However, the BRA module considers both model detection and inference, and it is a lightweight module. This is further conformed by the results in Table IV. On the AIR-SARShip-1.0 dataset, the BRA module improves the model performance. With a 0.7% increase in mAP and no significant decrease in FPS, the results demonstrate the low impact on inference speed. To fully validate the effectiveness of the BRA module, the detection results are visualized in Figs. 8 and 9. The visualized detection results show that the attention of the baseline is not focused and the features are not obvious. Fig. 8 shows the heat maps of the detection process. The heat-map results show that the BRA module makes the scattered attention more focused on ship targets. Fig. 9 illustrates feature maps of the detection process. The feature-map results show that the BRA module improves the key features of ships, and this is conducive to further feature extraction and detection of small ships by the network. In addition, Fig. 7 visually illustrates the effect of the BRA module compared with that of the baseline. Fig. 13 shows the results of the ablation experiment, where row 1 is the GT and rows 2–5 represent the results after adding the modules sequentially. The green, red, blue, and yellow boxes represent the real ship target, detection results of the different scenarios, missed targets, and false alarms, respectively. Fig. 7 and rows 2 and 3 of Fig. 13 show that the baseline has more omissions and false alarms, and this is especially obvious in small targets; the BRA module effectively

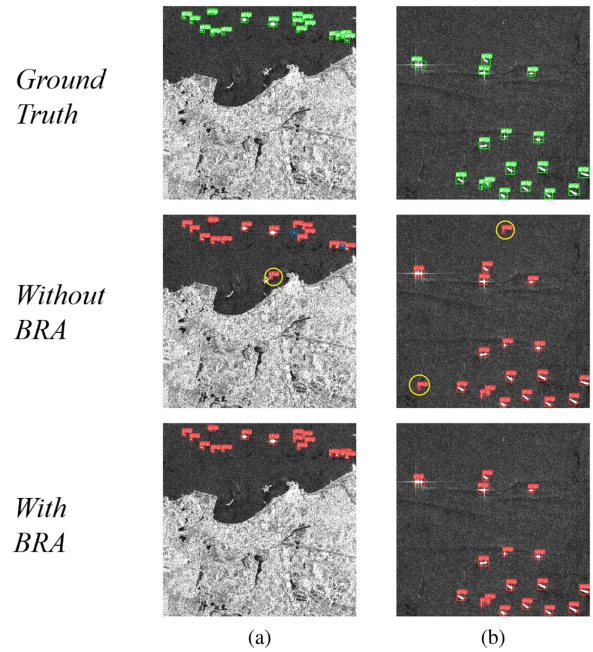

Fig. 7. Visual detection results of BRA. The yellow circles represent the false alarms, while the red and blue boxes represent the detection targets and missing targets, respectively. (a) Inshore detection results. (b) Offshore detection results.

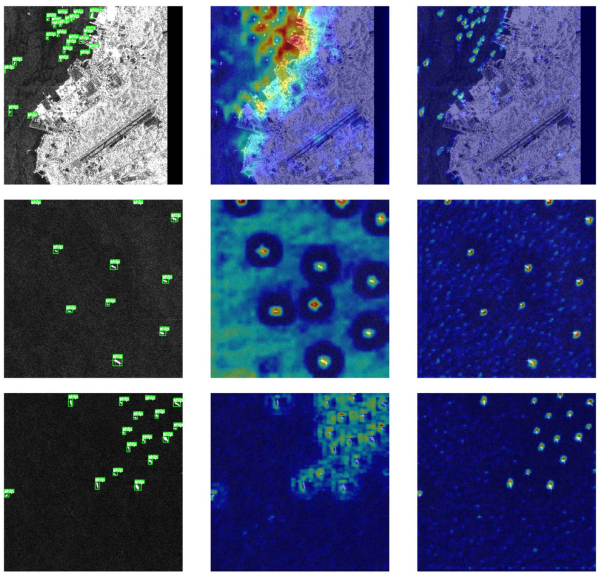

Fig. 8. Some visual heat maps of BRA. (a) GT. (b) Heat maps without BRA (Baseline). (c) Heat maps with BRA.

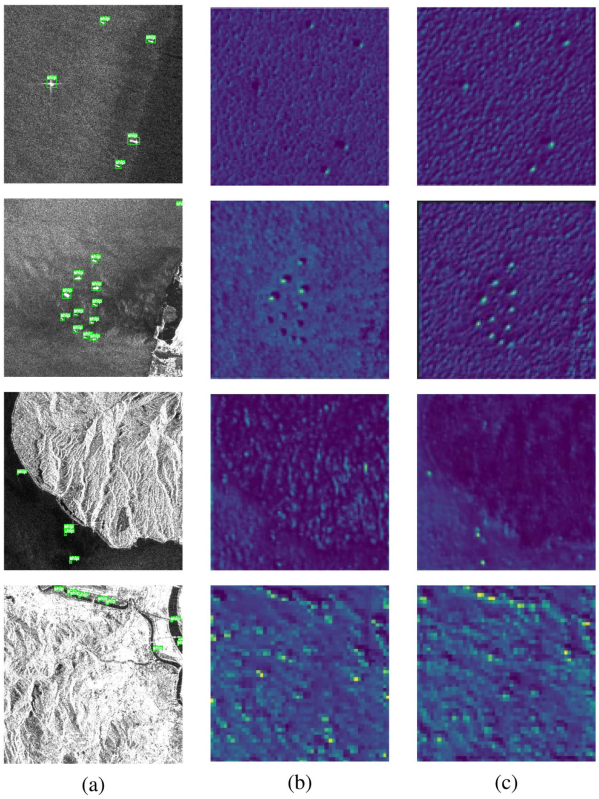

Fig. 9. Some visual feature maps of BRA. (a) GT. (b) Feature maps without BRA (Baseline). (c) Feature maps with BRA.

reduces these problems. The final detection of the ships is more accurate, further proving the effectiveness of this module.

2) Efficacy Analysis of BSTFF: Further experiments are conducted to analyze efficacy of BSTFF. As shown in Table III, the adjustments to the neck structure resulted in a steady

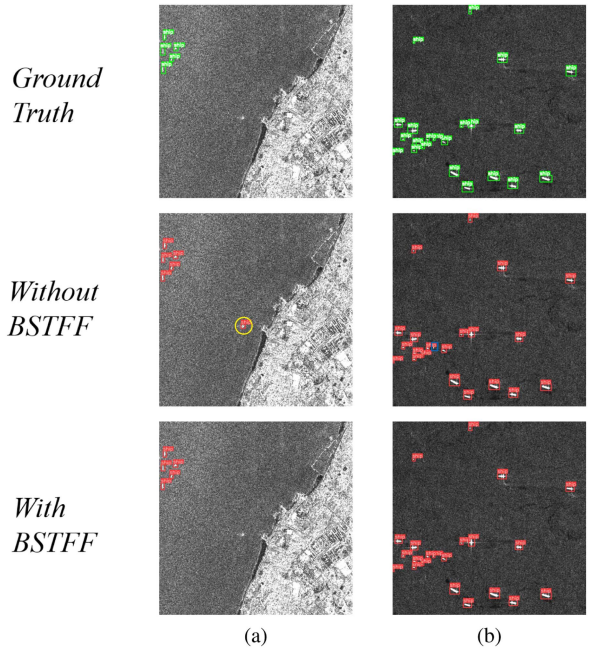

Fig. 10. Visual detection results of BSTFF. The yellow circles represent the false alarms, while the red and blue boxes represent the detection targets and missing targets, respectively. (a) Inshore detection results. (b) Offshore detection results.

improvement in detection performance. Compared to BRA, the improved neck structure resulted in a 1.2% increase in the mAP, a 1.2% increase in the AP_{50}, a 0.2 M increase in the parameters, and a 3.12 decrease in the FPS. These results show that the fusion of the shallow and deep feature maps significantly improve the performance at a relatively small computational cost, while the addition of the SwTR module does not significantly affect the model inference. That is, the proposed neck balances the accuracy with the detection speed. Fig. 10 visually illustrates the effect of the BSTFF compared with that of the FPN-PAN. The detection results in rows 3 and 4 of Fig. 13 also show that the deep feature fusion as well as the SwTR module can effectively reduce the problem of high false-alarm rate, especially reducing the false detection of small ships. This has also been illustrated by the results on the AIR-SARShip-1.0 dataset, where the model's mAP is improved by 2.7% and AP_{50} by 1.6%, as shown in Table IV, with only a slight decrease in FPS. Fig. 11 shows the PR curves of the proposed network on the two datasets. As shown, the area enclosed by the curves and axes increases with each additional scheme, and the envelope area reaches its maximum after the superposition of the three schemes. This also represents the best detection performance of the improved network.

3) Efficacy Analysis of CNWD: In addition, the CNWD loss function further improves the network performance without additional parameters. As shown in Table III, mAP improves by 0.8% and AP_{50} improves by 0.6%, with FPS displaying a significant improvement. Although $AP_{50:95}$ slightly decreases, it does not affect the overall detection. These results show that the IoU has larger parameters when regressing and is not friendly to small targets, while the CNWD performed lesser calculations and enhances the regression on small ships. In short, the CNWD

TABLE V
COMPARISON WITH OTHER METHODS ON LS-SSDD_V1.0 DATASET

Methods	mAP$^{\%}$	AP$^{\%}_{50}$	AP$^{\%}_{50:95}$	AP$_s$	AP$_m$	AR$_{100}$	FPS	Param$^{\times 10^6}$
Cascade-RCNN [15]	66.51	65.9	24.2	0.238	0.300	0.360	8.98	66.51
FCOS [43]	65.91	59.2	19.8	0.189	0.308	0.309	10.18	32.12
Faster-RCNN [14]	67.20	64.2	23.5	0.229	0.300	0.340	6.70	42.67
YOLOX [44]	72.91	65.2	23.3	0.218	0.372	0.402	4.57	8.94
CenterNet [45]	69.88	58.9	19.9	0.191	0.345	0.340	4.48	83.62
SSD [11]	75.01	71.8	25.8	0.246	0.377	0.356	7.86	23.61
TOOD [46]	68.40	63.2	22.8	0.219	0.331	0.345	8.95	32.02
GFLv2 [47]	70.01	62.9	22.7	0.217	0.338	0.353	8.54	32.27
PPYOLOE [48]	73.89	68.4	26.4	0.253	**0.381**	0.417	20.68	**7.61**
Ours	**78.30**	**77.5**	**28.9**	**0.279**	0.373	0.417	**47.85**	8.27

The bold value means the best performance.

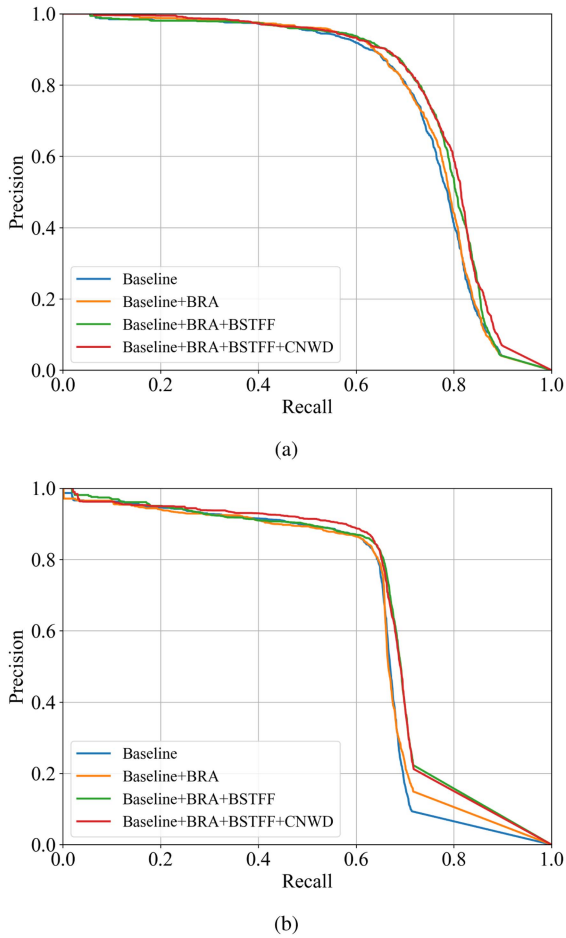

Fig. 11. PR curves of ablation experiments on different datasets. (a) LS-SSDD_v1.0. (b) AIR-SARShip-1.0.

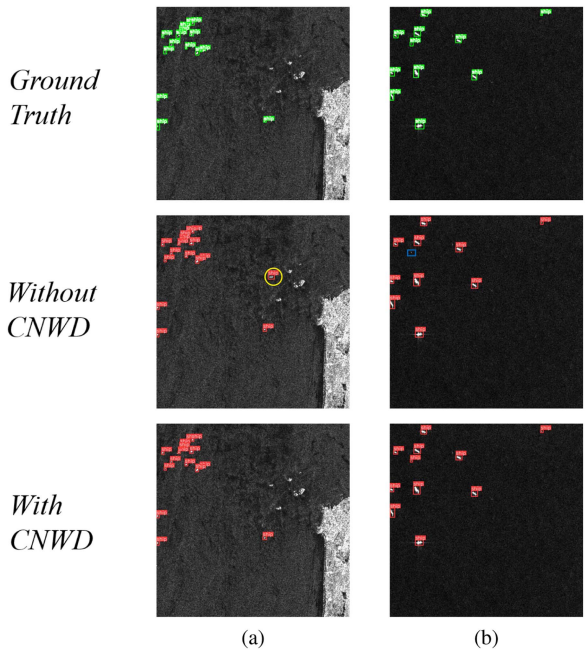

Fig. 12. Visual detection results of CNWD. The yellow circles represent the false alarms, while the red and blue boxes represent the detection targets and missing targets, respectively. (a) Inshore detection results. (b) Offshore detection results.

loss function enhances the detection accuracy of the model by reducing the inference time, while only slightly increasing the parameters. To further verify the effectiveness of CNWD, Fig. 12 shows the visual detection results of CNWD compared with those of CIoU. The detection results in rows 4 and 5 of Fig. 13 illustrate that the CNWD further reduces the missed detection and improves the regression of small ships. Table IV shows that the proposed CNWD also improves the mAP, AP$_{50}$, and AP$_{50:95}$. The FPS slightly decreases for the baseline because most of the share of IoU is retained in the CNWD loss function for detection on the AIR-SARShip-1.0 dataset, with a lesser share of NWD. This also illustrates that a different NWD ratio in the CNWD affects the model inference time, further validating that the IoU is more computationally intensive. The overall improvement of the network enhances the detection of small ships without affecting the original ship detection, displays better multiscale detection capability, and reduces missed and false detections. Finally, the proposed network with three improvements demonstrates 78.3% mAP for the LS-SSDD_v1.0 dataset and 67.3% for the AIR-SARShip-1.0 dataset. Fig. 11 shows the PR curves of the proposed network on the two datasets. As shown, the area enclosed by the curves and axes increases with each additional

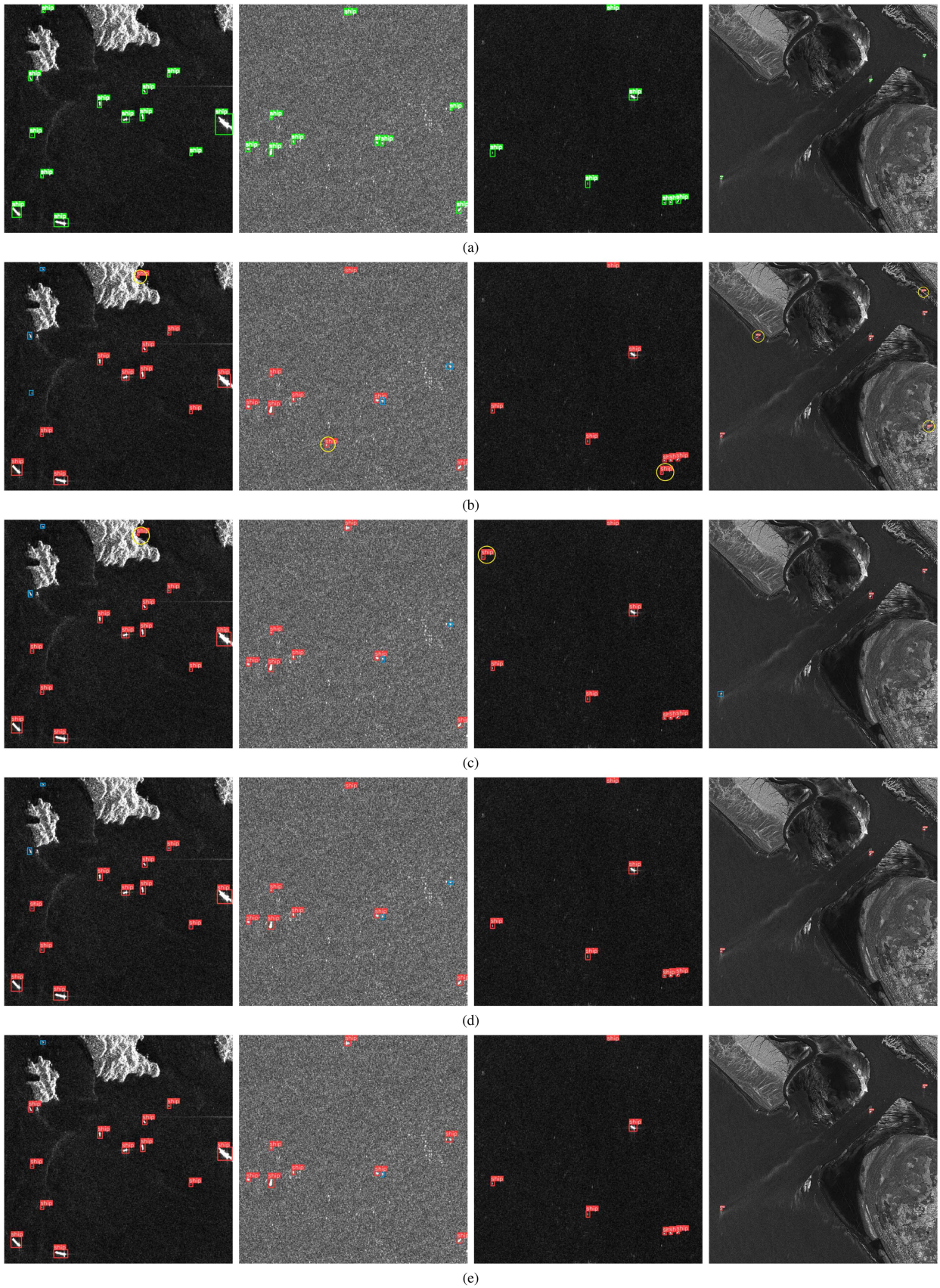

Fig. 13. Comparison of ablation experiment results. The yellow circles represent the false alarms, while the red and blue boxes represent the detection targets and missing targets, respectively. (a) GT. (b) Baseline. (c) Baseline + BRA. (d) Baseline + BRA + BSTFF. (e) Ours (Baseline + BRA + BSTFF + CNWD).

TABLE VI
COMPARISON WITH OTHER METHODS ON AIR-SARSHIP-1.0 DATASET

Methods	mAP$^{\%}$	AP$^{\%}_{50}$	AP$^{\%}_{50:95}$	AP$_m$	AP$_l$	AR$_{100}$	FPS	Param$^{\times 10^6}$
Faster-RCNN [14]	48.36	48.0	17.6	0.201	0.066	0.299	5.15	42.67
GFLv2 [47]	50.75	50.0	17.1	0.190	0.114	0.283	5.32	32.27
FCOS [43]	54.86	54.1	18.3	0.200	0.152	0.279	5.20	32.12
Cascade-RCNN [15]	52.56	53.9	18.9	0.210	0.102	0.314	5.38	69.15
TOOD [46]	50.0	49.6	17.7	0.201	0.087	0.308	5.24	32.02
YOLOX [44]	61.25	60.3	22.4	0.245	0.190	0.357	**15.93**	8.94
SSD [11]	59.30	58.4	21.6	0.237	0.184	0.346	8.91	23.61
PPYOLOE [48]	62.08	61.0	24.7	0.268	**0.219**	**0.390**	5.75	**7.61**
Ours	**67.3**	**64.1**	**27.1**	**0.302**	0.200	0.366	14.51	8.27

The bold value means the best performance.

scheme, and the envelope area reaches the maximum after the superposition of the three schemes. This also represents the best detection performance of the improved network.

D. Comparison With Other Methods

We compare the results obtained using the proposed scheme with some other state-of-the-art object-detection methods on both datasets, as shown in Tables V and VI. Table V shows that the proposed method achieves the highest mAP of 78.3%, which is 5.39%, 4.41%, and 3.29% higher than those of YOLOX, PPYOLOE, and SSD, respectively. AP$_{50}$, AP$_{50:95}$, AP$_m$, AR$_{100}$, and FPS reach 77.5%, 28.9%, 37.3%, 41.7%, and 47.85, respectively. The values of AP$_s$ of 27.9% indicate the highest accuracy in small-ship detection. Despite the smallest parameters, PPYOLOE does not perform well in accuracy and inference speed. The proposed method improves the accuracy of smaller ships in the detection task and only slightly influences the detection of normal-size ships. In addition, the parameters are smaller than those used in the other methods except for those in PPYOLOE, which is approximately 1/3 that of SSD, and 1/4 that of TOOD, GFLv2, and FCOS, and much smaller than those of methods, such as Cascade-RCNN, Faster-RCNN, and CenterNet. In terms of inference speed, the method far exceeds the two-stage detectors and reaches the highest values in many one-stage detectors, such as SSD, TOOD, and YOLOX. The results in Table VI show that the proposed method achieved a mAP of 67.3%, and it AP$_{50}$, AP$_{50:95}$, and AP$_m$ are the highest levels among all the compared methods, reaching 64.1%, 27.1%, and 30.2%, respectively. Although PPYOLOE has higher AP$_l$ and AR$_{100}$ values, all other AP values are lower than those of the proposed method, and the FPS is also lower. Despite YOLOX achieving the highest FPS, the fastest inference speed, and smaller parameters, the performances of the remaining metrics are not satisfactory. That is, YOLOX loses on model accuracy. The results show that the two-stage detectors perform relatively poorly in the multiscale detection task, in terms of not only the huge parameters and slow inference speed but also lower accuracy. Furthermore, the comparison of the PR curves of the different methods in Fig. 14 shows that the proposed method has the largest envelope area, verifying the robustness. Therefore, the proposed method can be concluded to consider the practical needs and improves the target detection accuracy of smaller ships, while fully balancing the detection speed and model parameters. The proposed network has good robustness, comprehensive detection performance, and better metrics than the other methods.

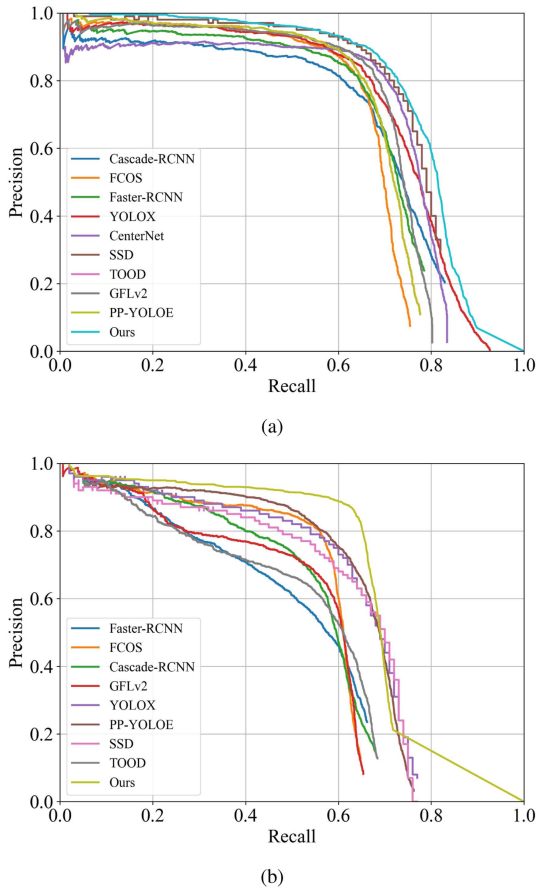

Fig. 14. PR curves for different methods on different datasets. (a) LS-SSDD_v1.0. (b) AIR-SARShip-1.0.

E. Experiment on Large Scene SAR Image

To verify the robustness of the proposed method on large scene SAR images, we select a large scene image from

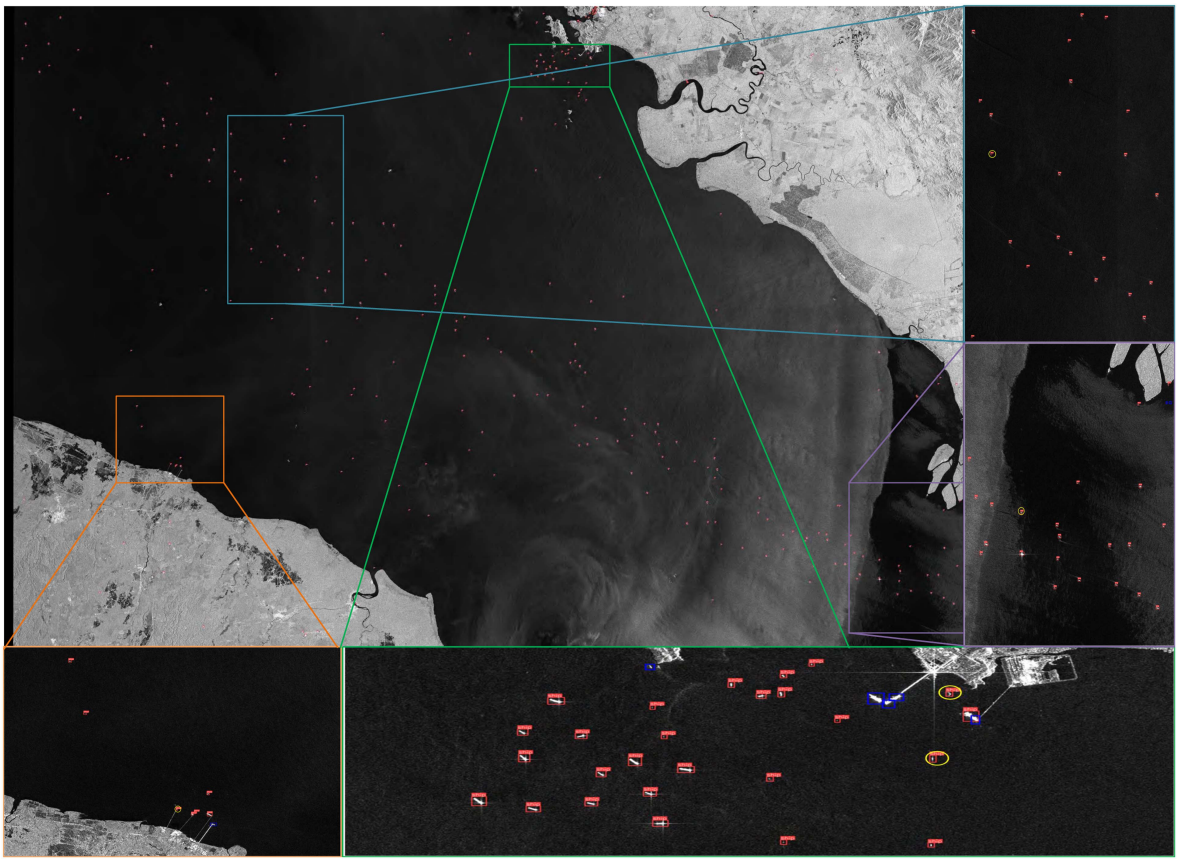

Fig. 15. Detection results of large SAR images on LS-SSDD_v1.0 (24 000 × 16 000 pixels). The yellow circles represent the false alarms, while the red and blue boxes represent the detection targets and missing targets, respectively.

the LS-SSDD_v1.0 dataset with an uncropped resolution of 24 000 × 16 000, and the inference results are shown in Fig. 15. As shown, the proposed method performs well, with the correct detection of most of the ship targets. Fig. 15 shows the detection results of the in-shore and off-shore scenarios. Few missing targets and false alarms prove the effectiveness and robustness of the proposed method.

IV. Discussion

The proposed method demonstrates three improvements with satisfactory detection results. To validate the effectiveness of the proposed method, several visualization results have been shown, strongly demonstrating the performance improvement brought by the different modules. On the one hand, the overall ablation experimental results in Fig. 13 show the reduced missing targets and false alarms with each additional improvement. The BRA module improves the small target attention, then the BSTFF enables the network to incorporate global information for feature extraction, and finally the CNWD reduces the small target sensitivity problem in regression computation. As shown in Tables V and VI, the proposed method displays a better performance than those of other methods. On the other hand, the in-shore background is more complex, with a considerable amount of interference information than that in the off-shore scenario. Especially, when the ship target intersects with the shore, the ship target cannot be detected correctly. The large scene detection results in Fig. 15 also show that the detection accuracy of the in-shore ship targets is lower than that of the off-shore scenario. Therefore, the missing targets and false alarms for small-ship detection in in-shore scenarios must be further reduced.

V. Conclusion

This study proposed a small-ship detection network for SAR images with dynamic sparse attention, improved neck structure, and fused loss function. The introduced BRA attention is based on the unique attention mechanism of the transformer, which deeply mines the feature information of small ships and effectively enhances the extraction of small-ship features. In addition, the improved neck structure can fuse contextual information. Moreover, the improvement of the detection head allows the network to be more approachable to detect small ships. Finally, the loss function that fuses CIoU and NWD demonstrates better robustness in small-ship regression and allows the network to better detect multiscale ships. The experimental results show that the proposed network improves the detection accuracy of small ships in SAR images and has better robustness of reducing missing and false detection. In future research, the detection accuracy of small ships in in-shore situations will be further improved. Moreover, further research should be to reduce detection speed and model size of the network for better performance and applicability to other platforms.

REFERENCES

[1] T. Liu, Z. Yang, G. Gao, A. Marino, S.-W. Chen, and J. Yang, "A general framework of polarimetric detectors based on quadratic optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5237418.

[2] T. Liu, Z. Yang, G. Gao, A. Marino, and S.-W. Chen, "Simultaneous diagonalization of hermitian matrices and its application in PolSAR ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–18, Nov. 2023.

[3] 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.

[4] Y. Guan et al., "Fishing vessel classification in SAR images using a novel deep learning model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–21, Sep. 2023.

[5] C. Chen, B. Liu, S. Wan, P. Qiao, and Q. Pei, "An edge traffic flow detection scheme based on deep learning in an intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1840–1852, Mar. 2021.

[6] T. Ye, X. Zhang, Y. Zhang, and J. Liu, "Railway traffic object detection using differential feature fusion convolution neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1375–1387, Mar. 2021.

[7] S. Zheng, J. Guo, X. Cui, R. N. J. Veldhuis, M. Oudkerk, and P. M. A. van Ooijen, "Automatic pulmonary nodule detection in CT scans using convolutional neural networks based on maximum intensity projection," *IEEE Trans. Med. Imag.*, vol. 39, no. 3, pp. 797–805, Mar. 2020.

[8] Z. Dong, M. Wang, Y. Wang, Y. Zhu, and Z. Zhang, "Object detection in high resolution remote sensing imagery based on convolutional neural networks with suitable object scale features," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2104–2114, Mar. 2020.

[9] Y. Wu, Y. Lin, Z. Zhou, D. C. Bolton, J. Liu, and P. Johnson, "DeepDetect: A cascaded region-based densely connected network for seismic event detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 62–75, Jan. 2019.

[10] Q. He, X. Sun, Z. Yan, and K. Fu, "DABNet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5601216.

[11] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2015, pp. 21–37.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

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

[14] 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.

[15] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.

[16] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, Jun. 2020.

[17] G. Gao, Y. Dai, X. Zhang, D. Duan, and F. Guo, "ADCG: A cross-modality domain transfer learning method for synthetic aperture radar in ship automatic target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Sep. 2023, Art. no. 5109114.

[18] G. Gao, P. Zhou, L. Yao, J. Liu, C. Zhang, and D. Duan, "A bi-prototype BDC metric network with lightweight adaptive task attention for few-shot fine-grained ship classification in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, Oct. 2023.

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

[20] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[21] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[22] N. I. Bountos, D. Michail, and I. Papoutsis, "Learning from synthetic InSAR with vision transformers: The case of volcanic unrest detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 4509712.

[23] G. Youk and M. Kim, "Transformer-based synthetic-to-measured SAR image translation via learning of representational features," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5206318.

[24] J. Jiao et al., "A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection," *IEEE Access*, vol. 6, pp. 20881–20892, 2018.

[25] 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, Feb. 2022, Art. no. 5217712.

[26] R. Li and Y. Shen, "YOLOSR-IST: A deep learning method for small target detection in infrared remote sensing images based on super-resolution and YOLO," *Signal Process.*, vol. 208, Jul. 2023, Art. no. 108962.

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

[28] Y. Gong, Z. Zhang, J. Wen, G. Lan, and S. Xiao, "Small ship detection of SAR images based on optimized feature pyramid and sample augmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 7385–7392, Aug. 2023.

[29] M. Sun, Y. Li, X. Chen, Y. Zhou, J. Niu, and J. Zhu, "A fast and accurate small target detection algorithm based on feature fusion and cross-layer connection network for the SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8969–8981, Oct. 2023.

[30] Y. Zhou, H. Liu, F. Ma, Z. Pan, and F. Zhang, "A sidelobe-aware small ship detection network for synthetic aperture radar imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5205516.

[31] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny object detection," 2021, *arXiv:2110.13389*. [Online]. Available: https://arxiv.org/abs/2110.13389

[32] J. Jiao et al., "DilateFormer: Multi-scale dilated transformer for visual recognition," *IEEE Trans. Multimedia*, vol. 25, pp. 8906–8919, Dec. 2023.

[33] J.-Y. Jiang, C. Xiong, C.-J. Lee, and W. Wang, "Long document ranking with query-directed sparse transformer," in *Proc. Findings Assoc. Comp. Linguist. Findings*, 2020, pp. 4594–4605.

[34] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," 2019, *arXiv:1912.12180*.

[35] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. H. Lau, "BiFormer: Vision transformer with bi-level routing attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10323–10333.

[36] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted self-attention via multi-scale token aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10843–10852.

[37] 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.

[38] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12993–13000.

[39] X. Lu, Y. Zhang, Y. Yuan, and Y. Feng, "Gated and axis-concentrated localization network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 179–192, Jun. 2020.

[40] J. Zhao and H. Zhu, "CBPH-Net: A small object detector for behavior recognition in classroom scenarios," *IEEE Trans. Instrum. Meas.*, vol. 72, Dec. 2023, Art. no. 2521112.

[41] T. Zhang et al., "LS-SSDD-v1.0: A deep learning dataset dedicated to small ship detection from large-scale Sentinel-1 SAR images," *Remote Sens.*, vol. 12, no. 18, Sep. 2020, Art. no. 2997.

[42] S. Xian et al., "AIR-SARShip-1.0: High-resolution SAR ship detection dataset (in English)," *J. Radars*, vol. 8, no. R19097, 2019, Art. no. 852.

[43] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9626–9635.

[44] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*. [Online]. Available: https://arxiv.org/abs/2107.08430

[45] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: https://arxiv.org/abs/1904.07850

[46] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3490–3499.

[47] X. Li, W. Wang, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss V2: Learning reliable localization quality estimation for dense object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11632–11641.

[48] S. Xu et al., "PP-YOLOE: An evolved version of YOLO," 2022, *arXiv:2203.16250*. [Online]. Available: https://arxiv.org/abs/2203.16250