A Small-Ship Object Detection Method for Satellite Remote Sensing Data

Xiyu Fan, Zhuhua Hu, *Senior Member, IEEE*, Yaochi Zhao, Junfei Chen, Tianjiao Wei, and Zixun Huang

***Abstract*—Satellite remote sensing technology can achieve real-time observation of ships at sea, and the remote sensing images obtained have the advantages of high contrast and low noise and have become one of the important means of marine monitoring. For the satellite remote sensing image data, there are two main problems: first, remote sensing data class-imbalance problem, and second the existing target detector in the presence of clouds, islands, farmed nets, and other interferences on the small-target ship, and there is a leakage of detection and wrong detection problem. To address the above problems, first, a new dataset containing 3881 images of remotely sensed ships in a variety of complex environments is constructed, which contains a total of 8418 ship instances. Second, we propose CSDP-YOLO for the small-target ship detection method with remote sensing data class imbalance. In order to enhance the performance of neural networks for small-target ship detection in remote sensing images, the innovative CSDP module is proposed, which uses deep large kernel convolution to enhance the sensory field of shallow features and mixes the channel positions using point convolution to obtain a more excellent feature extraction performance. Finally, the MPDIoU loss function is introduced to solve the class-imbalance problem between remote sensing small target ships and the background. We compare the performance with other state-of-the-art algorithms. The experimental results show that the proposed CSDP-YOLO algorithm can significantly improve the performance of small-target ship detection for private datasets. Its average precision, recall, and AP_{50} are improved to 90.1%, 86.6%, and 91.4%, respectively. For the SSDD public remote sensing dataset, its metrics can reach the highest 93.6%, 93.7%, and 96.8%, respectively.**

***Index Terms*—Class-imbalance, satellite remote sensing imagery, ship detection, small object detection, YOLOv7.**

I. Introduction

SHIPS are vital components of military and defense activities, and their movements must be closely observed since they are vital sea transit vehicles. The identification of ships is of utmost importance in the fight against illicit fishing, port trade, and marine traffic safety. At the moment, radar, optical and infrared reflectance, thermal infrared sensors, satellite remote sensing, and hyperspectral imaging are the main data sources used for the monitoring of small-sized ships in the ocean [1], [2]. The capacity of the satellite remote sensing technology to provide high-resolution, low-noise remote sensing photographs has attracted a lot of interest.

Small-target detection is an open problem in the field of remotely sensed imagery for a wide range of applications, including large-scale monitoring of marine ships, intelligent marine traffic, and ship position-based services. Traditional ship detection methods are mainly based on constant false alarm detection (CFAR) [3] to detect ships in remote sensing images. These methods first do a land–ocean segmentation, which allows land pixels to be suppressed and prevents interference with the CFAR step, thus limiting the speed required to acquire small-target ships. Finally, after CFAR prescreening, discriminators need to be designed to suppress noise. In addition, these methods usually rely on the statistical distribution of sea clutter, resulting in poor robustness of the new remotely sensed images [4].

With the development of deep learning-based target detection algorithms in computer vision (CV), researchers in the field of remote sensing also began to explore methods for detecting small-target ships from deep learning algorithms [5]. Due to the difficulty in acquiring remote sensing images and the fact that no researcher had yet developed a specialized remote sensing dataset at that time, deep learning-based detection methods could not be applied to remote sensing ship detection at the beginning. With the opening of the SSDD public dataset [6], a trend in the field of deep learning-based remote sensing ship detection was set off. SSDD provided researchers with a large amount of remote sensing data and evaluation criteria, which solved the problem of the lack of data for deep learning algorithms. By now, more and more researchers have adopted deep learning-based methods in this field. Sun et al. [53] focused on the complex environment and diverse ship scales of SAR images. An anchor-free detection method is proposed. It provides a better method for SAR ship detection. Wang et al. [54] proposed an MFFN multifeature fusion network, which can obtain ship seat texture information from the background of remote sensing images. Zhou et al. [55] proposed a network specifically for detecting small ships in remote sensing images to address the difficulty of detecting small ships. Gong et al. [56] proposed the enhancement strategy of SSPNet network and small ships, which contributed to the detection of SAR ships. Li et al. [7] used

Manuscript received 27 November 2023; revised 29 February 2024 and 11 June 2024; accepted 24 June 2024. Date of publication 27 June 2024; date of current version 12 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62161010 and Grant 61963012, in part by the Key Research and Development Project of Hainan Province under Grant ZDYF2022GXJS348 and Grant ZDYF2022SHFZ039, and in part by the Hainan Province Natural Science Foundation under Grant 623RC446. *(Corresponding author: Zhuhua Hu.)*

Xiyu Fan, Zhuhua Hu, Junfei Chen, Tianjiao Wei, and Zixun Huang are with the School of Information and Communication Engineering, Hainan University, Haikou 570228, China (e-mail: 22210810000042@hainanu.edu.cn; eagler_hu@hainanu.edu.cn; 22220854000129@hainanu.edu.cn; weitianjiao@hainanu.edu.cn; 23210810000004@hainanu.edu.cn).

Yaochi Zhao is with the School of Cyberspace Security (School of Cryptology), Hainan University, Haikou 570228, China (e-mail: zhyc@hainanu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3419786

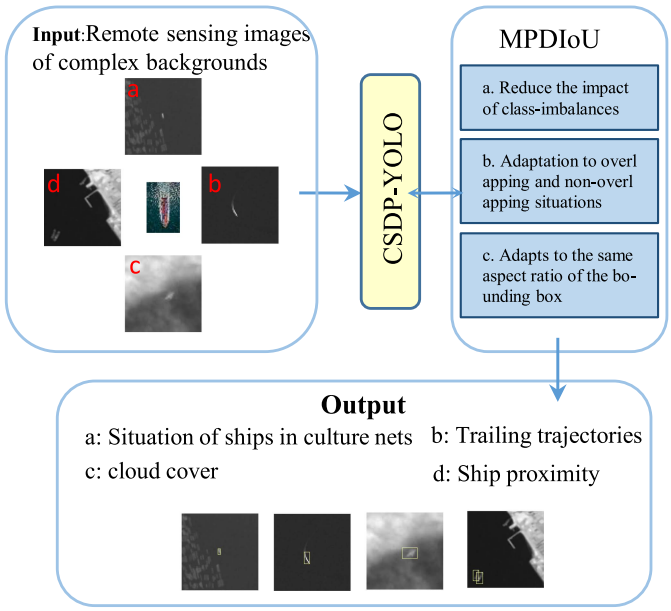

Fig. 1. Challenges of detecting small ships in complex backgrounds and proposed methods.

the improved Faster RCNN, a two-stage target detection method for ship detection, to improve the accuracy of small-target ship detection, but its detection time is long and it cannot detect ships in real time. Subsequently, the SSDD [8] detection algorithm and you only live once (YOLO) [9] single-stage detection algorithm have been introduced, which centralizes the target detection task and target category task into a single neural network model, achieving high accuracy and fast target detection performance. However, these generic target detectors have some problems when applied directly to small-target ship detection in remote sensing datasets.

1) First, there are fewer datasets used for the ship detection of small targets in the field of remote sensing imagery, and the datasets in recent years, such as SSDD, MSDS, and AIR-SAR Ship, the ship objects in these datasets suffer from the problems of ship multiscale and low-pixel value of the dataset, and the simple environment in which the ships are located, which cannot satisfy the ship detection task in the complex environment. Therefore, there is a need to establish a small-target ship dataset for remote sensing images in complex marine environments.
2) Second, current mainstream detection models often miss detecting ships with small or ambiguous targets. Our detection task faces challenges in the case of farmed nets, trailing trajectories, cloud cover, and ships in close proximity (as shown in Fig. 1). Since smaller ships usually occupy only a few pixels of the image [10], after a number of feature extractions, it will lead to the network's ability to lose the small-target features and the spatial layering information of the neural network.
3) Finally, remote sensing datasets are interfered by background noise, such as mariculture nets in near-shore harbors, lighthouses on the sea surface, and islands, which usually lead to false alarms for ship detection. Traditional algorithms use features to distinguish between small-target ships and other disturbances, but they usually lack accuracy and effectiveness.

Based on the above analysis, this constructs a dataset of remote sensing images of small-target ships. This dataset of remote sensing images is mainly from Hainan 1, after cutting, data enhancement, and other operations. In total, 3831 selected, high-quality remote sensing images were obtained, and we used horizontal bounding boxes and labeled 8418 instances of 1 category (ships). Second, YOLOv7, as a single-stage detector, still has the advantages of high detection accuracy and speed on small-target objects. In this article, we try to apply YOLOv7 to small-target detection of ships in remote sensing images. At the same time, we further explore optimize its accuracy and speed so that it can identify small targets with fuzzy ships more accurately and efficiently. To enhance the detection performance of small-target ships, we propose the CSDP-YOLO algorithm, which makes up for the shortcomings of YOLOv7 [11] in the performance of detecting small-target ships. The main contributions of the work are as follows.

1) A dataset of small-targeted ships from satellite remote sensing images in complex sea areas was constructed. The dataset contains 3831 remotely sensed images with 8418 labeled instances. This dataset does not require land and sea segmentation, which is helpful for the dynamic monitoring of ships in the sea and harbors.
2) Aiming at the problem that remote sensing small targets have few pixels, and when feature extraction is performed on the feature map, tiny pixel offsets will lead to a decrease in detection accuracy. In this study, we propose a CSDP structure based on deep convolution. The CSDP module consists of a full convolution block composed of large kernel deep convolution and point convolution, which replaces part of the extended high-efficiency layer aggregation network (ELAN) aggregation layer of the original YOLOv7 and enhances the model's ability of feature extraction for small targets. The large kernel deep convolution captures a wider range of contextual information and helps to identify complex patterns of ocean–land interaction in the image. Point convolution helps reduce computational cost and provides interchannel interaction while maintaining computational efficiency.
3) Aiming at the serious data class-imbalance problem between small-target ships and background, we introduce MPDIoU as the loss function of the penalty. So that the bounding box loss function can make full use of the geometric properties of the bounding box regression to speed up the convergence of the model and improve the detection accuracy by minimizing the distances of the top-left vertices and the bottom-right vertices between the predicted bounding box and the real bounding box.

The rest of this article is organized as follows. In Section II, the proposed dataset and the current methods used for ship remote sensing image detection are described in detail. Section III presents the detailed structure of the proposed CSDP-YOLO. Section IV describes the dataset and the experimental setup,

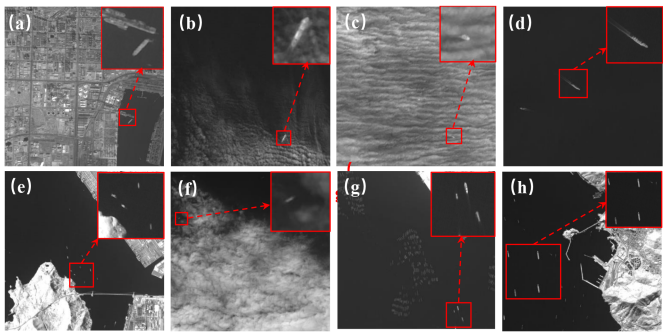

Fig. 2. Number of representative datasets of remotely sensed images from different backgrounds (The ship's target area has been magnified.)

verifies the improvement of the proposed algorithm through comparative experiments and analyses the experimental results, and verifies the generalization of the model on a publicly available dataset. Finally, Section V concludes this article.

II. Related Work

A. Remote Sensing Datasets

In the past few years, many target detection datasets, such as SSDD, OpenSARship [13], SAR-Ship-Dataset, AIR-SAR Ship, HRSID, LS-SSDDv1.0, FUSARv1.0, and MSDS, have been proposed in the field of aerial imagery to advance the research on ship detection in remote sensing imagery. However, the objects in these datasets have multiple scales, and therefore, these datasets are more suitable for evaluating detectors designed for multiscale object detection rather than small object detection. Although some work on small-target ship detection from remotely sensed imagery uses mainstream remote sensing data for training, the ocean is complex and most of the existing datasets do not have small-target ships in a variety of complex situations, such as cloudy, undersea nets, nearshore of harbors, islands, and so on, so this study constructs a small-target ship dataset in response to the complexity of the ocean environment in which small-target ships are located.

The construction strategy of the ship dataset in this study is as follows. We obtained 70 original remote sensing images from "Hainan No.1 Star 01," with a resolution of 28 000 × 28 000 pixels, and we set the resolution of the cropped images to 1024 × 1024 pixels, and the corners of the original remote sensing images are less than 1024 pixels. We set the resolution size of the cropped image to 1024 × 1024 pixels, and for the part of the original remote sensing image whose edges are less than 1024, we save it as the original image without filling. To facilitate network training, the cropped remote sensing images are manually filtered, and finally, 3831 high-definition remote sensing images are obtained. Our proposed dataset contains only one ship category object and 97% of the small target ship objects are less than 40 pixels, which pushes the difficulty of small-target detection to the extreme and meets the needs of practical scenarios for applications. Fig. 2 shows some representative samples of remote sensing images of ships, including cloud cover, harbor docking, and marine aquaculture nets, including harbors, cloudy climates, underwater nets, and other scenarios. Finally, the whole dataset is randomly split into training dataset (70%), validation dataset (20%), and test dataset (10%)

B. Deep Learning Object Detection

Since deep neural networks can automatically learn the threshold features and shape features of the target, they are of great research value in ship detection in remote sensing images. Object detection algorithms based on deep learning can be divided into two categories: two-level detectors and single-level detectors. The single-stage detector uses the full convolutional network to perform classification and regression tasks on the anchor frame only once to obtain detection results [14]. The two-stage detector uses a deep neural network to perform two classification and regression tasks on the anchor frame to obtain detection results.

So far, common two-level detectors include R-CNN, Faster R-CNN, feature pyramid network [15], Mask RCNN, etc. Most of the design ideas of the latter two-level detectors are based on the improvement of the previous network, and the starting point of improvement is mostly from the backbone network, regional suggestion network, etc. Although the two-level detectors are more accurate in detecting objects, the detection time has increased because of the method of extraction of boundary boundaries using enveloped neural networks. Common single-stage detectors include SSD [8], YOLO [9], and RetinaNet [17]. In addition, Zhang et al. [52] proposed a multiscale global scattering feature association network for remote sensing propagation target identification. It gave us important inspiration for ship identification technology in remote sensing images. Kang et al. [57] proposed a multilayer fusion convolutional neural network to solve the difficulty of detecting small-scale ships in SAR images. Sun et al. [58] aimed at the characteristics of multiscale and dense array of ships in high-resolution SAR images. A bidirectional fusion module is proposed for YOLO, which makes the model have better robustness and generalization.

The YOLO series is widely used in the field of ship detection of remote sensing images. Deng et al. [18] used YOLOv2 to detect ships in remote sensing images, and proposed YOLOv2-reduced, mainly by reducing the partial neural networks of YOLOv2, and it has achieved greater efficiency than the YOLOv2 detection (the AP of YOLOv2-Reduced is 89.76%) and low loss of accuracy (the AP of YOLOv2 is 90.05%). Zhang et al. [6], [52] have used the DarkNet-19 network to replace the original YOLOv3 backbone network, and the new network has greatly improved the detection efficiency of the sensing remote imaging ship. Subsequently, Wang et al. [20] proposed an SSS-YOLO network, cleverly designed the feature extraction layer of the neural network, and enhanced the semantic information of small target ships. Zhou et al. [27] proposed a multiscale ship detection network based on the YOLOv5 model, which achieved a good balance between the complexity of the model and the reasoning time. Tang et al. [28] proposed a convolved block attention mechanism with a multiscale receptive field based on the YOLOv7 and made full use of the information of

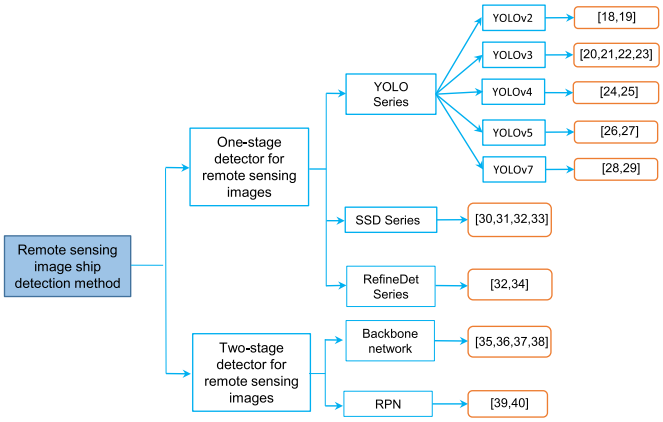

Fig. 3. Ship detection methods for remote sensing images with one-stage and two-stage detectors.

feature maps to accurately capture useful regions in the feature maps. The ship detection method of remote sensing images with single-order detector and double-order detector is shown in Fig. 3. Nowadays, the high accuracy and real-time performance of YOLOv7 is particularly prominent in the task of target detection, and YOLOv7 adopts a multilevel pyramid structure for target detection, which means that it can simultaneously predict the target location and category on the feature map of different resolutions. To better train the model, it allows auxiliary heads to be attached to the pyramid of the middle layer for training. The advantage of this training strategy is that it helps make up for information that might be lost in the next level of pyramid prediction. In other words, by making predictions on the pyramid of the middle layer, the model can better capture features at various resolution levels, thus improving the accuracy of target detection.

However, the performance of YOLOv7 on remote sensing image datasets is not very good. In order to improve its performance, CSDP is used as part of the backbone network of YOLOv7 based on the YOLOv7 model in this article. Deep packet convolution is used in the CSDP module to divide the input channels into multiple groups, and the channels in each group are convolved only with the convolution kernel in the corresponding group. The CSDP module is used for processing channel features, which can be regarded as a channel attention mechanism. This helps the model to better capture important information in the input features and improve the performance of small-target ship identification. At the same time, the MPDIoU loss function is also introduced to solve the problem of the imbalance between prospects and background classes, further improving the efficiency and accuracy of the model boundary frame regression.

III. Proposed Methods

YOLO is the most advanced single-stage target detection algorithm that has undergone multiple iterations [9]. In addition to the original version of YOLO, there are many derivative algorithms based on YOLO architecture, which are optimized and improved based on YOLO to meet the needs of different application scenarios. YOLOv7 is also an optimized version of the YOLO architecture, it adopts the extended high-efficiency layer aggregation network (ELAN) strategy [40]. By combining cardinality to combine different features, the neural network can learn and converge more effectively by controlling the shortest gradient path, and enhance the learning ability of the network without destroying the original gradient path. The excellent learning ability of YOLOv7 is more suitable for deployment in the detection of small-target objects. Based on YOLOv7, CSDP-YOLO, a method for small-target ships in remote sensing images, is proposed in this study.

A. Proposed CSDP-YOLO Framework

The CSDP-YOLO network architecture is shown in Fig. 4. First, two high-efficiency layer aggregation network (ELAN) modules of the backbone module of YOLOv7 are removed, and the CSDP layer is introduced into the backbone part to enhance the model's performance of extracting low-level feature maps. Low-level feature maps have higher resolution, containing information on the tail and shape of the small-target ship, as well as the location and details of the islands. It is helpful to improve the discrimination of the model and the accuracy of detection. Second, based on the imbalance between the background and foreground of small-target ships, MPDIoU is introduced as a loss function to solve the problem of small loss and slow gradient convergence in the training process. The proposed network architecture can be categorized into CBS, MPC, ELAN, and CSDP modules. CBS is a basic volume module, consisting of variable lengths of volume blocks. Cat acts as a multivoltage module that uses outputs from other volume layers to perform concat operations to improve the accuracy of the network. CSDP is the low-level feature map extraction layer proposed by the authors. Specifically, the DP module is a full convolution block composed of large core deep convolution and point convolution. We choose deep convolution to mix spatial locations, and point convolution to mix channel locations. Therefore, in the process of low-level feature extraction, there is a larger receptive field to pay attention to more detailed information, such as ship ends and shapes, so that the model can better take into account the global information, and at the same time, restrain the interference of islands, clouds, and other factors on small-target ships [48]. MP is a downsampling module, which helps to gradually reduce the size of the feature map, so that the network can detect the target at different resolutions, thus improving the detection ability of the model for the target of different sizes. SPPCSPC is an improved spatial pyramid pool structure that helps to handle objects at different scales, making the model more robust.

In summary, small-target recognition from remote sensing images is achieved using the trained CSDP-YOLO model. The training process is summarized in Algorithm 1.

B. CSDP Feature Extraction Architecture

Although the extended efficient layer aggregation network (ELAN) of YOLOv7 learns the features of the ships using different layer weights, it enhances the learning capability of the network by introducing the operations of expanding, shuffling, and

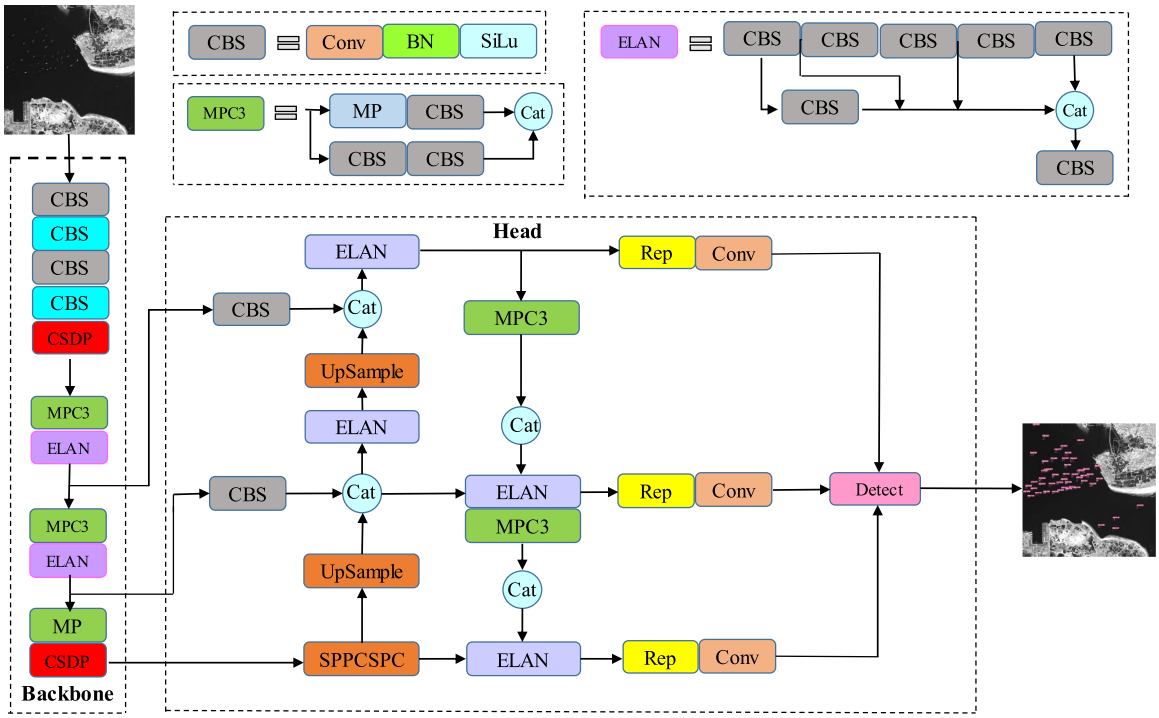

Fig. 4. Proposed small-target ship detection method: The CSDP-YOLO architecture.

Algorithm 1: Training Strategy of CSDP-YOLO.

Input: Given training samples $X = \{x_1, \ldots, x_k\}$ and labels $Y = \{y_1, \ldots, y_k\}, k \in N^+$

Output: An adeptly trained CSDP-YOLO model.

1: An adeptly trained CSDP-YOLO model is shown in Fig. 4.
2: Initialize the parameters $\theta = (w, b, r, \alpha)$
3: **repeat**
4: Randomly select a batch of instances X_b from X
5: Pass training samples forward through the CSDP-YOLO model
6: Compute the training loss $\mathcal{L}$ by $\mathcal{L} = box\_loss + object\_loss + class\_loss$
7: Propagate $\mathcal{L}$ back through CSDP-YOLO and update the parameters with SGD
8: Determine θ by minimizing the cost function $\mathcal{L}$ using X_b
9: Until the end of model convergence

merging bases while ensuring the continuity of the gradient paths to improve the performance and generalization [48]. However, these operations also increase the computational complexity of the network, while undergoing multiple convolutions can lead to a situation where the neural network loses information about small-target ships under remote sensing datasets.

To better integrate the small-target ships, islands, and other detailed information on the shallow feature map, for the remote sensing of small-target fuzzy, irregular shapes, and poor existing conditions [41], the CSDP feature extraction layer is constructed, and the module in the case of satisfying the computational parameters relative to the ELAN network is less, the grouping convolution of each input channel, so that the number of grouping is the same as the number of input channels. As a result, a point convolution is performed to mix the features of each output channel, to improve the ability of the neural network in the shallow feature map for small-target ships information acquisition so that better performance for the detection of small-target ships task. Dot convolution is performed to mix the features of each output channel, to improve the ability of the neural network to acquire information about small-target ships in shallow feature maps, and to achieve better performance for the task of small-target ship detection. As shown in Fig. 5, our proposed CSDP feature fusion structure consists of three feature-variable convolutional layers, a deep convolutional module, and a point convolutional module. The CSDP layer first accepts an input feature map, and then performs feature transformations through two 1×1 convolutional layers (CV1 and CV2) to adjust the dimensionality of the input feature map and passes it to the subsequent DP module and the CV3 convolutional layer, respectively. We set the input passed into the path of the DP module as X, the output is Z_1 after CV1 convolution, the output is Z_2 after CV2 convolution, and the size of the input feature map X is $P \times P \times C_{in}$. The kernel size of the convolution is 1×1 with a step size of 1, which is expressed in the following equation:

$$z_1 = \mathrm{BN}\left(\sigma\left\{\mathrm{Conv1}_{c_{\mathrm{in}} \to h}(X,\ \mathrm{s}\ = 1,\ \mathrm{k\_size}\ = 1)\right\}\right) \quad (1)$$

$$z_2 = \mathrm{BN}\left(\sigma\left\{\mathrm{Conv2}_{c_{\mathrm{in}} \to h}(X,\ \mathrm{s}\ = 1,\ \mathrm{k\_size} = 1)\right\}\right) \quad (2)$$

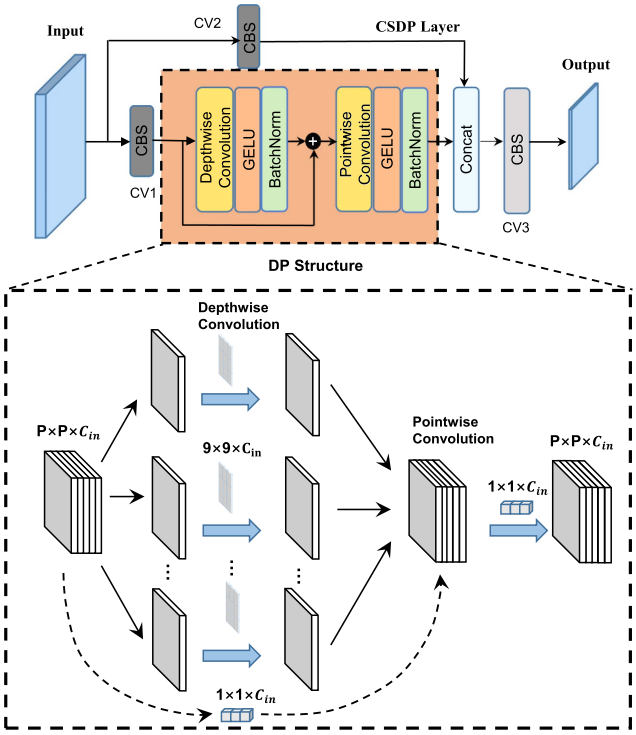

Fig. 5. Proposed CSDP network architecture, where CV1, CV2, and CV3 are 1×1 convolutional networks.

where z_l is the output of the residual structure, which can extract and fuse the features of remote sensing image after deep grouping convolution and shallow features, so that the deep network can better learn the small-target features. C_{in} is the number of channels into the CSDP structure. $\text{Kernel}_{\text{size}}$ is the size of the convolution kernel.

The DP block is composed of deep convolution (the number of grouped convolutions is equal to the number of channels h) and point convolution (the size of the convolution kernel is 1×1), and the DP module works well with a huge kernel (the convolution kernel is $9 \times 9 \times C_{\text{in}}$). We inserted the DP^1 and DP^2 modules into two parts of YOLOv7, and the number of channels of the inputs to the two modules is 128 and 1024, respectively. Each convolution is followed by a GELU activation function and a BatchNorm, which are expressed as follows:

$$z_l' = \text{BN}\left(\sigma\left\{\text{ConvDepthwise}\ (z_1,\ \text{s}\ =1,\ \text{k\_size}\ =9)\right\}\right) + z_1 \tag{3}$$

$$z_{l+1} = \text{BN}\left(\sigma\left\{\text{ConvPointwise}\ (z_l',\ \text{s}\ =1,\ \text{k\_size}\ =1)\right\}\right) \tag{4}$$

where Z_l' denotes the residual structure output after convolution with a large kernel ($k = 9$), the output of grouping convolution on each channel is mixed by pointwise convolution for channel mixing, and the 1×1 convolution is used to realize the exchange of information between the channels, which is specifically realized in the section is shown in Fig. 5. Finally, in order to enrich the expression of the shape features of the small-target ship, the output feature map of CV2 (the two output paths are independent

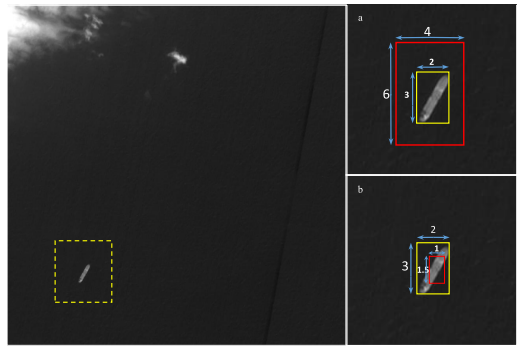

Fig. 6. CIoU failure when the predicted box (red) has the same aspect ratio as the real box (yellow).

of each other) is spliced with the output of the DP module after the channel mixing operation in terms of the channel dimensions (dim = 1), which fully integrates the feature information of the two paths to form a richer expression of the features of the small-target ship, as shown below:

$$\text{CSDP} = \text{BN}\left(\sigma\left\{\text{concat}\left(z_{l+1}, z_2\right)\right\}\right) \tag{5}$$

where z_{l+1} is the output of the DP module (4), and z_2 denotes the second independent path (the one passing through CV2) for the processing of the input feature map, which directly performs the convolution operation on the input feature map, and finally splice it again in the channel dimensions through the concat function to generate a richer representation of the features of the small ships.

Compared with the ELAN high aggregation network, our proposed CSDP module has the following advantages. First, it uses large kernel deep convolutional hybrid spatial locations as well as point convolutional hybrid channel locations, to fully utilize the feature information of small-target ships between convolutional groupings, and to improve the network's ability to extract the characterization of the small-target ships in complex environments. Second, all network layers of the module maintain the same resolution size of input and output, and there is no downsampling operation of the feature maps at the continuous network layer, which prevents the loss of small-target ship information. Finally, the number of parameters of the network is reduced due to the use of deep convolution versus point convolution for feature extraction, making CSDP-YOLO require less computational resources.

C. MPDIoU Loss Function

In the task of small-target detection in remote sensing images, for the problem of serious regional imbalance between the small-target ships and the background, the CIoU loss function used leads to the problem of small loss and slow convergence of the gradient of the network during the training process when the aspect ratio of the preframe and the real frame is the same. We introduce MPDIoU as the loss function of the penalty, which can minimize the class-imbalance problem. CIoU is used as the

Algorithm 2: Intersection Over Union With Minimum Points Distance.

Input: Two arbitrary convex shapes:A,B $\subseteq \mathbb{S} \in \mathbb{R}$
Output: MPDIoU.

1: For A and B,(x_1^A, y_1^A), (x_2^A, y_2^A) denote the coordinates of the upper-left and lower-right points of A, and (x_1^B, y_1^B), (x_2^B, y_2^B) denote the coordinates of the upper-left and lower-right points of B.
2: $d_1^2 = (x_1^B - x_1^A)^2 + (y_1^B - y_1^A)^2$, $d_2^2 = (x_2^B - x_2^A)^2 + (y_2^B - y_2^A)^2$
3: $\text{MPDIoU} = \frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2+h^2} - \frac{d_2^2}{w^2+h^2}$

loss function of the penalty in YOLOv7 as follows:

$$\text{CIoU} = \text{IoU} - \frac{\rho^2\left(\mathcal{B}_{gt}, \mathcal{B}_{\text{prd}}\right)}{C^2} - \alpha V \tag{6}$$

$$V = \frac{4}{\pi^2}\left(\arctan\frac{w^{\text{gt}}}{h^{\text{gt}}} - \arctan\frac{w^{\text{prd}}}{h^{\text{prd}}}\right)^2 \tag{7}$$

where $\rho^2(\mathcal{B}_{\text{gt}}, \mathcal{B}_{\text{prd}})$ is the Euclidean distance between the centroids of the predicted bounding box and the groundtruth bounding box, C is the area of the outer matrix covering the predicted box and the groundtruth bounding box, α is a weight parameter, and V is a similarity parameter measuring the aspect ratio [48]. However, when the predicted box and the groundtruth bounding box have the same aspect ratio (e.g., Fig. 6), the value of V of CIoU is 0 at this point, thus degenerating into DIoU, and the loss function of the bounding box regression fails, at which time the loss function cannot provide a good gradient for the model, thus limiting the speed of model convergence and detection accuracy. CIoU for the problem that the weight parameters of the network are difficult to update effectively when the neural network is updating the gradient. We make full use of the geometrical characteristics of the horizontal rectangle of the anchor box, and introduce the MPDIoU as the model's loss function for bounding box regression by minimizing the distances of the upper left and lower right points between the predicted bounding box and the groundtruth bounding box [12], which takes into full consideration the existing loss function of overlapping and nonoverlapping regions, height deviation, and the distance between the centroids of the bounding box, and simplifies the calculation process. The calculation of MPDIoU is summarized in Algorithm 2.

The MPDIoU metric simplifies the similarity comparison between predicted bounding boxes and groundtruth bounding boxes, allowing regression with and without overlapping bounding boxes. During training, the model's predicted bounding box $\mathcal{B}_{\text{prd}} = [x^{\text{prd}}, y^{\text{prd}}, w^{\text{prd}}, h^{\text{prd}}]^T$ approximates the groundtruth bounding box $\mathcal{B}_{\text{gt}} = [x^{\text{gt}}, y^{\text{gt}}, w^{\text{gt}}, h^{\text{gt}}]^T$ by minimizing the loss function, as in the following equation:

$$\mathcal{L} = \min_{\theta} \sum_{\mathcal{B}_{\text{gt}} \in \mathbb{B}_{\text{gt}}} \mathcal{L}\left(\mathcal{B}_{\text{gt}}, \mathcal{B}_{\text{prd}} \mid \theta\right) \tag{8}$$

where $\mathbb{B}_{\text{gt}}$ is the set of bounding boxes, $\mathcal{B}_{\text{prd}}$ is the set of predicted bounding boxes, θ is the regression parameter, and $\mathcal{L}$ is the ln paradigm. Based on the previously mentioned Algorithm 1, we

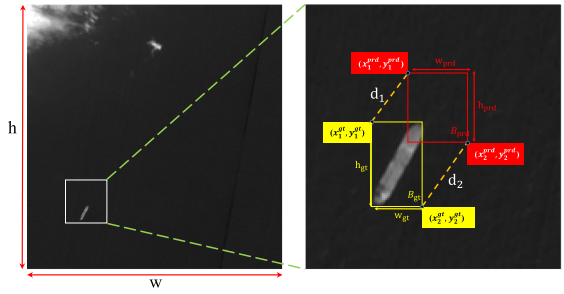

Fig. 7. We introduce the parameters of $\mathcal{L}_{\text{MPDIoU}}$ with the computational factors of the bounding box regression metrics.

define the loss function as follows:

$$\mathcal{L}_{\text{MPDIoU}} = 1 - \text{MPDIoU}. \tag{9}$$

Meanwhile, the parameters of bounding box regression can be determined by four coordinates, the regression factors are shown in Fig. 7, and the regression parameters are calculated in the following equation:

$$|C| = \left(\max\left(x_2^{\text{gt}}, x_2^{\text{prd}}\right) - \min\left(x_1^{\text{gt}}, x_1^{\text{prd}}\right)\right) \times \left(\max\left(y_2^{\text{gt}}, y_2^{\text{prd}}\right) - \min\left(y_1^{\text{gt}}, y_1^{\text{prd}}\right)\right) \tag{10}$$

$$x_c^{\text{gt}} = \frac{x_1^{\text{gt}} + x_2^{\text{gt}}}{2}, y_c^{\text{gt}} = \frac{y_1^{\text{gt}} + y_2^{\text{gt}}}{2}$$

$$x_c^{\text{prd}} = \frac{y_1^{\text{prd}} + y_2^{\text{prd}}}{2}, x_c^{\text{prd}} = \frac{x_1^{\text{prd}} + x_2^{\text{prd}}}{2} \tag{11}$$

$$w_{\text{gt}} = x_2^{\text{gt}} - x_1^{\text{gt}}, h_{\text{gt}} = y_2^{\text{gt}} - y_1^{\text{gt}}$$

$$w_{\text{prd}} = x_2^{\text{prd}} - x_1^{\text{prd}}, h_{\text{prd}} = y_2^{\text{prd}} - y_1^{\text{prd}} \tag{12}$$

where $|C|$ is the minimum outer join matrix between the bounding box and the real box, and $(x_c^{\text{gt}}, y_c^{\text{gt}})$ and $(x_c^{\text{prd}}, y_c^{\text{prd}})$ are the center coordinates of the real box and the bounding box. w_{gt} denotes the width of the real box, and h_{gt} denotes the height of the real box. w_{prd} denotes the width of the predicted bounding box, and h_{prd} denotes the height of the predicted bounding box. We used MPDIoU parameters, as shown in Fig. 7. When the predicted bounding box has the same aspect ratio as the groundtruth bounding box, $\mathcal{L}_{\text{MPDIoU}}$ has a lower value in the case where the predicted box is contained in the groundtruth bounding box. Hence, the introduction of MPDIoU ensures the accuracy of the bounding box regression. Algorithm 3 for the bounding loss box for both IoU and MPDIoU are as follows.

In Algorithm 3, B_{gt} denotes the matrix area of the bounding box, then $A^{\text{gt}} > 0$ as in Algorithm 3(d). We set the conditions as $\mathcal{A}_{\text{prd}} \geq 0$ $\mathcal{I} \geq 0$ (predicted area and intersection area are nonnegative), so for any state of the predicted bounding box $\mathcal{B}_{\text{prd}} = (x_1^{\text{prd}}, y_1^{\text{prd}}, x_2^{\text{prd}}, y_2^{\text{prd}})$, the intersection area of the predicted box with the groundtruth bounding box $\mu > 0$, and $\mu > \mathcal{I}$ (the area of the concatenation is greater than the area of the intersection). Then, $\mathcal{L}_{\text{MPDIoU}}$ is a bounded function. Thus, $\mathcal{L}_{\text{MPDIoU}}$ is a bounded function, and for any predicted box with IoU $= 0$ (the predicted box does not overlap the groundtruth bounding box),

Algorithm 3: IoU and MPDIoU as Bounding Box Losses.

Input: Predicted B_prd and ground truth B_gt bounding box coordinates,width and height of input image:w,h. where B_prd and B_gt are represented by $B\_prd = [x_1^{prd}, y_1^{prd}, x_2^{prd}, y_2^{prd}]$, $B\_gt = [x_1^{gt}, y_1^{gt}, x_2^{gt}, y_2^{gt}]$

Output: $\mathcal{L}_{IoU}, \mathcal{L}_{MPDIoU}$

1: Determine the predicted bounding box(B_prd) and ensure that $x_2^{prd} > x_1^{prd}, y_2^{prd} > y_1^{prd}$

2: $d_1^2 = (x_1^{prd} - x_1^{gt})^2 + (y_1^{prd} - y_1^{gt})^2$, $(x_2^{prd} - x_2^{gt})^2 + (y_2^{prd} - y_2^{gt})^2$

3: Calculate the area of the groundtruth bounding box and the predicted bounding box. $A_{gt} = (x_2^{gt} - x_1^{gt}) \times (y_2^{gt} - y_1^{gt})$, $A_{prd} = (x_2^{prd} - x_1^{prd}) \times (y_2^{prd} - y_1^{prd})$

4: Calculate the intersecting area of the groundtruth bounding box and the predicted bounding box $\mathcal{I} = (\min(x_2^{prd}, x_2^{gt}) - \max(x_1^{prd}, x_1^{gt})) \times (\min(y_2^{prd}, y_2^{gt}) \max(y_1^{prd}, y_1^{gt}))$

5: IoU $= \frac{\mathcal{I}}{\mu}$, $\mu = A_{gt} + A_{prd} - \mathcal{I}$

6: $\mathcal{L}$ = 1 - IoU, $\mathcal{L}_{MPDIoU}$ = 1 - $\left(\text{IoU} - \frac{d_1^2}{h^2+w^2} - \frac{d_2^2}{h^2+w^2}\right)$

for the loss of MPDIoU, we have $\mathcal{L}_{\text{MPDIoU}} = 1 - \text{MPDIoU} = 1 + \frac{d_1^2}{d^2} + \frac{d_2^2}{d^2}$, at which point the process of minimizing $\mathcal{L}_{\text{MPDIoU}}$ is minimizing $\frac{d_1^2}{d^2} + \frac{d_2^2}{d^2}$ (the distance between the top-left corner point of the bounding box and the bottom-right corner point).

IV. Experimental Results and Analysis

A. Datasets

The remote sensing dataset used in this study is derived from the Hainan-1 satellite, originating from Hainan province, China. This satellite has provided a wealth of remote sensing imagery for various maritime applications and ocean management, thereby enabling dynamic ship detection. We sliced the original remote sensing images into frames of 1024×1024 pixels. Due to the potential difficulty of accurately observing small-target ships with the naked eye, we carried out color depth changes on the original images. Subsequently, based on AIS information, we used labeling software to annotate the small-target ships, ultimately yielding 3831 images. We found that many small-target ships were hidden in complex backgrounds, exhibiting conditions, such as trailing ship trajectories, crowding, cloud cover, and proximity to the shore, among others (as shown in Fig. 2). Concurrently, approximately 97% of the targets do not exceed more than 0.15% of the image area. Because of this, the detection of smaller object targets necessitates augmented inference of both shallow and deep features. This dataset pushes the difficulty of small-target detection to the extreme, which fulfills the requirements of practical application scenarios. Fig. 2 illustrates several representative images. In the experiment, the training set, test set, and validation set were divided at a ratio of 7:2:1, respectively, encompassing 5795, 837, and 1786 instances. The image size input to the network is uniformly set to 640 × 640 pixels.

TABLE I
Experimental Setting

Item	Value
CPU	Intel Xeon Gold 6132 CPU @ 2.60 GHz
GPU	Integrated Matrox G200eW3 Graphics Controller
Cuda Vision	12.0
Data processing	Python3.8
Deep learning framework	PyTorch

B. Experimental Setup

We implemented CSDP-YOLO on PyTorch 2.0.1 and trained and tested using the integrated Matrox G200eW3 Graphics Controller. During training and testing, the operating system is Ubuntu. Specific details are given in Table I.

To ensure adequate neural network training, the batch size was uniformly set to 16 with an initial learning rate of 0.01, carried out over 300 rounds of training. The stochastic gradient descent (SGD) optimizer was selected to minimize the MPDIoU loss. During the training phase, partial pretrained models of YOLOv7 were not used. Although CSDP-YOLO and YOLOv7 share part of the network architecture, their use for the detection of small-target ships in remote sensing images may result in negative transfer, that is, a decline in performance. Initial training from scratch can avoid such circumstances as it prevents the model from being restrained by previous training experiences.

C. Experimental Metrics

To demonstrate the advantages of CSDP-YOLO, we use precision (P), recall (R), F1 score, average accuracy [mean average precision (mAP)], mAP at 0.5 intersection over union (mAP@0.5), and mAP@0.5:0.95 to evaluate the parameter equations as follows:

$$\text{precision}(P) = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{Recall}(R) = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{AP}_i = \int_0^1 P_i(R_i) d(R_i) \tag{15}$$

$$\text{mAP} = \frac{1}{n} \sum_{i=1}^{n} \text{AP}_i. \tag{16}$$

In the above equations, TP is defined as samples correctly identified in the positive class, FP is defined as samples incorrectly identified as the positive class in the negative class, FN is defined as samples incorrectly identified as the negative class in the positive class, and TN is defined as samples correctly identified as the negative class in the negative class. Typically, precision (P) refers to the proportion of true positives (TP) among all samples predicted as the positive class, whereas recall (R) refers to the proportion of true positives (TP) among all true

TABLE II
RESULTS OF DIFFERENT METHODS

Method	AP_{val}	R_{val}	mAP@0.5	mAP@0.5:0.95
YOLOv7 [11]	0.830	0.780	0.853	0.404
YOLOv7x [11]	0.848	0.782	0.857	0.398
Biformer-YOLOv7 [43]	0.844	0.801	0.863	0.410
CBAM-YOLOv7 [44]	0.841	0.811	0.862	0.414
CSDP-YOLOv7(ours)	**0.901**	**0.866**	**0.914**	**0.497**

The experimental data of our proposed method, which is marked with black marks and distinguished from other methods.

positives (actual targets) [51]. F1 is used to assess the model's precision and recall performance. mAP represents the average value of AP and is used to measure the overall detection accuracy of object detection algorithms. mAP@0.5 is used to quantify the model's average performance at different IoU thresholds [49].

D. Experimental Results and Analysis

We evaluated the CSDP-YOLO on an Integrated Matrox G200eW3 Graphics Controller. To thoroughly validate the effectiveness of our proposed methodology, we compared our network with the original YOLOv7, YOLOv7x, and a version of YOLOv7 enhanced with an attention mechanism.

The attention mechanism is a deep learning technique inspired by the human visual system, analogous to our inclination to focus on specific areas while observing the world. This mechanism allows computational models to focus on the most critical part of the task at hand. This is achieved by allocating weights to different inputs, thereby assigning higher weights to information relevant to the task and downscaling information that is irrelevant. Such a technique permits the model to better deal with complicated data, reaping substantial advancements across various applications. Wang et al. [42] proposed incorporating a CBAM module into YOLOv5, enhancing the detection performance of remote sensing image object detection. CBAM is a convolutional neural network-based attention mechanism module that introduces spatial attention and channel attention mechanisms, assisting CNN models in better understanding and utilizing the features of input data to enhance detection performance. Zhu et al. [43] introduced the BiFormer attention mechanism module, a dynamically sparse attention mechanism that filters redundant information, retaining only the integral parts of interest, greatly boosting the identification of small targets. Both of these methods are highly representative algorithms of the attention mechanism [49]. We introduced them into YOLOv7 to explore their performance in detecting small-target ships in remote sensing data [50], comparing them with the improvements proposed in this article. The entire network is illustrated in Fig. 4.

The detection results of various algorithms on the remote sensing small-target ship dataset are shown in Tables II and III, and Figs. 9 and 10 display comparison results of different methods. Fig. 8 illustrates the mAP and recall curves of YOLOv7 and CSDP-YOLO, respectively. Fig. 10(a) and (b) represents the detection results in near-shore ports and dense ship conditions, respectively, while Fig. 10(c) and (d) shows the detection results in the case of cloud cover and ocean waves, respectively. Since

TABLE III
NUMBER OF MISSES, TIME, AND MODEL PARAMETERS

Method	Omissions	Inference time(ms)	FPS(ms)	Params(M)
YOLOv7 [11]	184	4.3	5.1	141.8
YOLOv7x [11]	182	6.9	7.7	270.1
Biformer-YOLOv7 [44]	167	6.2	7.0	161.9
CBAM-YOLOv7 [45]	158	7.3	8.1	146.6
CSDP-YOLOv7(ours)	**112**	**4.8**	**0.7**	**140.4**

The experimental data of our proposed method, which is marked with black marks and distinguished from other methods.

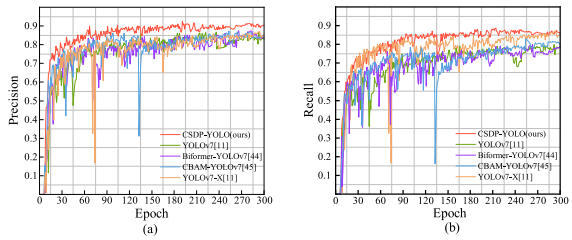

Fig. 8. Training process for five different methods. (a) Precision curve. (b) Recall curve.

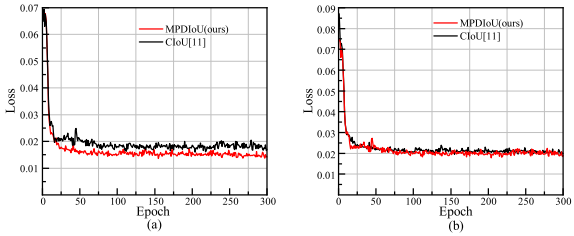

Fig. 9. MPDIoU and CIoU loss function training process. (a) Bounding box loss. (b) Overall loss.

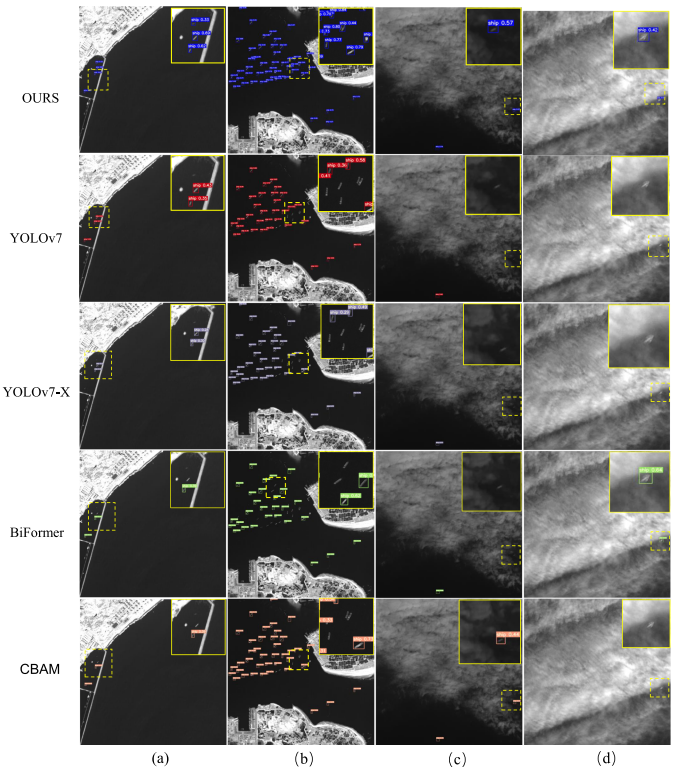

Fig. 10. Detection results of different methods for different scenarios in our early dataset. (a) Near-shore harbors. (b) Ship density and close proximity. (c) Cloud cover. (d) Ocean waves.

ships are the sole detection targets in the remote sensing dataset, precise localization of ship targets has more value than target classification. Compared with the original YOLOv7 network, the ELAN aggregation network layer of YOLOv7X is wider than the original ELAN layer, slightly enhancing detection accuracy and recall. However, this introduces more computations that significantly decrease FPS from 5.1 to 7.7 (as given in Table III), and the model's parameters increase from 141.8 to 270.1 M, which makes it harder to deploy on terminal devices. BiFormer [43], a dynamically sparse attention mechanism, filters most irrelevant key–value pairs in coarse area features. This retains only a small portion of the routing area, highlighting crucial information on small targets and substantially improving ship feature extraction. However, due to excessive focus on global information, many ships are omitted (up to 167 ships) and the FPS is notably affected [e.g., ships are not detected under circumstances of cloud and fog obstructions and proximity of ships, as shown in Fig. 10(c) and (b), respectively]. The CBAM model, which emphasizes important spatial information features, still does not improve false detection and omission in complex environments [as in Fig. 10(d), where ship omission occurs in a crowded port].

On the other hand, CSDP-YOLO significantly reduces ship omission (as given in Table III), and the parameter quantity of CSDP is less than that of the original YOLOv7 network, reaching 140.4 M, effectively reducing the required computational resources. Moreover, the class-imbalance problem between the background and foreground of small-target ships was solved through the CSDP layer in the backbone network and the MPDIoU loss function, maintaining stable detection performance under adverse weather or complex maritime conditions (as seen in Fig. 10). In comparison to YOLOv7, accuracy is improved by 8.6%, and recall rate by 11%. Among the detection tests on the test set (inclusive of 837 instances), CSDP-YOLO has the least omissions and no significant damage to the FPS, which is as high as 5.5.

Furthermore, Fig. 10 illustrates the mAP and recall curves of YOLOv7 and CSDP-YOLO. The comparison chart shows that the proposed CSDP-YOLO outperforms YOLOv7 in terms of average detection accuracy and recall rate. This indicates that the overall ability of CSDP-YOLO to recognize small-target ships in complex environments surpasses that of YOLOv7. Fig. 9 presents a comparison of boundary loss and overall loss between the proposed model and YOLOv7. Our proposed model uses MPDIoU, the boundary loss and overall loss of which are lower than YOLOv7's CIoU. The model that utilizes the MPDIoU loss function predicts frames more accurately.

E. Ablation Experiments

In this research, some high-aggregation network layers (ELAN) in YOLOv7 are replaced with the proposed CSDP network layer (as shown in Fig. 4). To assess the effectiveness of the proposed CSDP network and the introduced MPDIoU loss function, we independently examined each replacement of the CSDP module in the backbone network of YOLOv7, focusing on AP values as well as recalls. The results of the ablation experiments are given in Table IV. The ELAN network

TABLE IV
ABLATION STUDY OF SMALL TARGET DETECTION TASK IN REMOTE SENSING IMAGES

CSDP	ELAN	MPDIoU	AP^{val}	AP^{val}_{50}	AP^{val}_{75}	R^{val}
✓			0.886	0.906	0.430	0.865
	✓	✓	0.830	0.865	0.345	0.831
✓	✓	✓	0.874	0.900	0.389	0.865
✓		✓	**0.901**	**0.914**	**0.460**	**0.866**

The experimental data of our proposed method, which is marked with black marks and distinguished from other methods.

TABLE V
EXPERIMENTAL RESULTS OF DIFFERENT METHODS ON THE SSDD DATASET

Model	Precision	Recall	AP_{50}
Faster R-CNN [45]	81.65	85.31	89.63
SSD [46]	85.30	91.6	89.3
FCOS [45]	84.15	92.52	90.61
YOLOv3 [45]	89.11	85.03	91.54
YOLOv5s [46]	93.10	92.90	94.60
YOLOv7 [45]	91.05	84.92	93.68
CSDP-YOLO(ours)	**93.60**	**93.70**	**96.80**

The experimental data of our proposed method, which is marked with black marks and distinguished from other methods.

aggregation layer proposed in the original YOLOv7 did not deliver ideal detection results for small-target ships; however, our proposed CSDP network layer addresses the problem of information loss during the multiple feature extraction processes of small targets. This is achieved by using a mix of large kernel depth convolution and point convolution to blend channel and spatial position, which more accurately captures the details of small target ships. From the ablation experiment, we discovered that the model's performance was optimized when two of YOLOv7's ELAN aggregation network layers were replaced with the proposed CSDP layers in YOLOv7's backbone network. Consequently, the model's average detection accuracy reached 91.4%, and the recall rate amounted to 86.6%. This suggests that the introduction of the CSDP module has enhanced the feature extraction capability of neural networks and reduced instances of false detection and omission. Lastly, the application of the MPDIoU loss function significantly improved the recall rate of the neural network, reaching 83% without the addition of the CSDP module. As the MPDIoU adequately considers the geometric properties of bounding box regression, minimizing the distances between the top left corner vertices and bottom right corner vertices of the predicted and actual bounding boxes, it ultimately outperforms the original YOLOv7's CIoU loss function.

F. Generalizability Verification

To evaluate the generalizability of the proposed CSDP-YOLO model, we used the publicly available SSDD remote sensing dataset for ship detection. We compared the mainstream remote sensing ship detection models, namely, Faster RCNN, SSD, FCOS, YOLOv3, and YOLOv7, with the proposed CSDP-YOLO model under the same environment and conditions. The

experimental results are given in Table V. The experiment results demonstrate that the CSDP-YOLO detection accuracy and recall rate have, respectively, improved to 93.6% and 93.7%, securing the best detection performance. The CSDP network layer mixes the spatial information of small-target ships by applying convolution to each input channel in groups, setting the number of groups equal to the number of input channels. Finally, point convolution is used to mix the features of each output channel, allowing the network to pay more attention to the finer details of small-target ships and reducing interference from irrelevant information in remote sensing images, such as islands and lighthouses.

G. Discussions

The results in Tables II and III give that in the detection task of our proposed dataset, our proposed CSDP-YOLO approach to target detection outperforms other existing algorithms when compared with the detection performance using only the YOLOv7 algorithm, and as can be seen from the detection results in Fig. 10, the CSDP-YOLO algorithm achieves accurate detection for the case of ships nearby and cloud occlusion, which are better than the other algorithms. The results in Table V give that CSDP-YOLO still outperforms other existing algorithms in the SSDD public dataset. Tables II and III give that the target detection network proposed in this article can still miss or misdetect the detection of small targets. On the one hand, the dataset may have fewer remote sensing images for certain complex backgrounds (e.g., mariculture nets, ship trailing), which leads to the model learning less information from these complex backgrounds during the training process. On the other hand, the pixels of some small-target ships in the dataset presented in this article are too small, pushing the difficulty of small-target detection to the extreme, which means that the model inevitably loses feature information during the training process.

Considering the different traveling volumes of ships in the South China Sea region in different seasons, bad weather may cause the remote sensing images taken by the satellite to be not clear enough. The experimental images are selected in the harbor as well as the principle Hainan land marine conditions, and the harbor in and out of a large number of ships, for the effective detection of ships close to each other, can prevent the ship hitchhiking to carry out smuggling activities, see Fig. 10(b). And for far away from the land of Hainan Province marine territory, see Fig. 10(c) and (d), the sea situation is sudden and changeable, so the detection of ships in poor conditions can effectively ensure the safety of the ship, and the emergence of accidents can be realized accurately locate and quickly out of the police.

The use of deep learning technology to detect small-target ships can, on the one hand, greatly improve the accuracy of detecting ships under adverse conditions and reduce the false alarms of marine safety systems. On the other hand, introducing deep technology is conducive to constructing a new model of modern intelligent detection of ships. From the point of view of maintaining marine safety, in a situation where many ships are entering and leaving the port and the degree of density a large, efficient, and scientific analysis of the ship's situation can reduce the investment of marine traffic management resources. At the same time, it can reduce the human subjective judgment and the false alarm rate of marine safety brought afterward. This innovative style has an important role in promoting efficient ship detection, rational allocation of marine traffic management resources, and realizing intelligent management of marine traffic, and can provide favorable technical support for the policy of marine ship management.

V. Conclusion

In this article, we construct a new dataset of remotely sensed ship images containing 3881 remotely sensed images in complex situations, such as harbors, cloud occlusion, ships close by, farmed nets, and so on, with at least one instance of small target ships in each remotely sensed image, and this dataset kind of contains 8418 instances. In addition, we propose a CSDP network layer based on the ELAN high-aggregation network layer for small-target ships in the problem of unsatisfactory feature extraction, especially in the case of harbors, dense ships, proximity, and cloud cover, and use a large kernel deep convolution to mix the spatial and channel positions of the small-target ships to capture a wider range of contextual information without downsampling the continuous layer, even in the case of complex environments. The proposed model remains valid for small-target ships even in complex environments. In our proposed private dataset, studies comparing multiple advanced detection models (YOLOV7, YOLOv7-X, BiFormer-YOLOv7, and CBAM-YOLOv7) are compared. The experimental results show that the proposed CSDP-YOLO algorithm can effectively improve the average precision, recall, and AP_{50} of detecting small ship data to 90.1%, 86.6%, and 91.4%, respectively.

Small-target ships occupy few pixels in complex environments and for the problem of imbalance between foreground and background categories of small target ships. The MPDIoU loss function can still be optimized when the predicted box has the same aspect ratio as the groundtruth bounding box and achieves a more efficient and precise bounding box regression by minimizing the distance between the predicted upper left point and the lower right point. We compare the bounding box losses of CIoU and MPDIoU, and the experimental results show that the trained MPDIoU bounding box loss is lower than that of CIoU, indicating that MPDIoU motivates the model to predicted bounding boxes more accurately.

Finally, to verify the generalization of our proposed model, we do 300 rounds of training on the SSDD remote sensing dataset under the same conditions as before, while comparing a variety of models commonly used for remote sensing ship detection (Faster R-CNN, SSD, FCOS, YOLOv3, YOLOv5s, and YOLOv7), and the experimental results show that our proposed model detects the indicators are optimal, and the average precision, recall, and AP_{50} of detection are improved to 93.6%, 93.7%, and 96.8%, respectively. In future research, we will continue to explore the application of deep learning on our

proposed remote sensing dataset and further expand the remote sensing dataset, and we will investigate the lightweight of the model and its deployment on end devices.

ACKNOWLEDGMENT

The authors would like to thank the referees for their constructive suggestions. For datasets and codes related to this article, please contact the corresponding author.

REFERENCES

[1] X. X. Zhu et al., "Deep learning meets SAR: Concepts, models, pitfalls, and perspectives," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 143–172, Dec. 2021.

[2] Z. Wang, L. Du, J. Mao, B. Liu, and D. Yang, "SAR target detection based on SSD with data augmentation and transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 150–154, Jan. 2019.

[3] B. Lebona, W. Kleynhans, T. Celik, and L. Mdakane, "Ship detection using VIIRS sensor specific data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 1245–1247.

[4] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020, doi: 10.1109/JSTARS.2020.3005403.

[5] Z. Lin, K. Ji, X. Leng, and G. Kuang, "Squeeze and excitation rank faster R-CNN for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 751–755, May 2019.

[6] T. Zhang et al., "SAR ship detection dataset (SSDD): Official release and comprehensive data analysis," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3690.

[7] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. Sar Big Data Era: Models, Methods Appl.*, 2017, pp. 1–6.

[8] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis. ECCV*, 2016, pp. 21–37.

[9] J. Redmon et al., "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[10] L. Zhao and S. Ji, "CNN, RNN or ViT? An evaluation of different deep learning architectures for spatio-temporal representation of sentinel time series," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 44–56, 2023, doi: 10.1109/JSTARS.2022.3219816.

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

[12] M. Sharma et al., "YOLOrs: Object detection in multimodal remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1497–1508, 2021, doi: 10.1109/JSTARS.2020.3041316.

[13] M. Jiang et al., "Ship contour extraction from SAR images based on faster R-CNN and Chan–Vese model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5203414.

[14] W. Liu, K. Quijano, and M. M. Crawford, "YOLOv5-Tassel: Detecting tassels in RGB UAV imagery with improved YOLOv5 based on transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8085–8094, 2022, doi: 10.1109/JSTARS.2022.3206399.

[15] J. Hou et al., "Detecting diseases in apple tree leaves using FPN–ISResNet–Faster RCNN," *Eur. J. Remote Sens.*, vol. 56, no. 1, 2023, Art. no. 2186955.

[16] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[17] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[18] Z. Deng, H. Sun, S. Zhou, and J. Zhao, "Learning deep ship detector in SAR images from scratch," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4021–4039, Jun. 2019.

[19] Y. L. Chang, A. Anagaw, L. Chang, Y. C. Wang, C. Y. Hsiao, and W. H. Lee, "Ship detection based on YOLOv2 for SAR imagery," *Remote Sens.*, vol. 11, 2019, Art. no. 786.

[20] J. Wang, Y. Lin, J. Guo, and L. Zhuang, "SSS-YOLO: Towards more accurate detection for small ships in SAR image," *Remote Sens. Lett.*, vol. 12, pp. 93–102, 2021.

[21] Y. Chaudhary, M. Mehta, N. Goel, P. Bhardwaj, D. Gupta, and A. Khanna, "YOLOv3 remote sensing SAR Ship image DetectionM," in *Data Analytics and Management*. Singapore: Springer, 2021, pp. 519–531.

[22] Y. Chen, T. Duan, C. Wang, Y. Zhang, and M. Huang, "End-to-end ship detection in SAR images for complex scenes based on deep CNNs," *J. Sens.*, vol. 2021, 2021, Art. no. 8893182.

[23] Z. Hong et al., "Multi-scale ship detection from SAR and optical imagery via a more accurate YOLOv3," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 6083–6101, 2021, doi: 10.1109/JSTARS.2021.3087555.

[24] Z. Ma, "High-speed lightweight ship detection algorithm based on YOLO-V4 for three-channels RGB SAR image," *Remote Sens.*, vol. 13, 2021, Art. no. 1909.

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

[26] G. Tang, Y. Zhuge, C. Claramunt, and S. Men, "N-YOLO: A SAR ship detection using noise-classifying and complete-target extraction," *Remote Sens.*, vol. 13, 2021, Art. no. 871.

[27] K. Zhou, M. Zhang, H. Wang, and J. Tan, "Ship detection in SAR images based on multi-scale feature extraction and adaptive feature fusion," *Remote Sens.*, vol. 14, 2022, Art. no. 755.

[28] H. Tang et al., "A Lightweight SAR image ship detection method based on improved convolution and YOLOv7," 2023.

[29] M. Yasir et al., "Instance segmentation ship detection based on improved Yolov7 using complex background SAR images," *Front. Mar. Sci.*, vol. 10, 2023, Art. no. 1113669.

[30] F. Zhou et al., "SAR target detection based on improved SSD with saliency map and residual network," *Remote Sens.*, vol. 14, no. 1, 2022, Art. no. 180.

[31] L. Li, Y. Du, and L. Du, "Vehicle target detection network in SAR images based on rectangle-invariant rotatable convolution," *Remote Sens.*, vol. 14, no. 13, 2022, Art. no. 3086.

[32] M. Zhu, G. Hu, S. Li, S. Liu, and S. Wang, "An effective ship detection method based on RefineDet in SAR images," in *Proc. Int. Conf. Commun., Inf. Syst. Comput. Eng.*, 2021, pp. 377–380.

[33] L. Jin and G. Liu, "An approach on image processing of deep learning based on improved SSD," *Symmetry*, vol. 13, no. 3, 2021, Art. no. 495.

[34] Y. Du, L. Du, and L. Li, "An SAR target detector based on gradient harmonized mechanism and attention mechanism," *IEEE Geosci. Remote Sens. Lett*, vol. 19, 2021, Art. no. 4017005, doi: 10.1109/LGRS.2021.3103378.

[35] Z. Hou, Z. Cui, Z. Cao, and N. Liu, "An integrated method of ship detection and recognition in SAR images based on deep learning," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1225–1228.

[36] C. Wang, W. Su, and H. Gu, "Two-stage ship detection in synthetic aperture radar images based on attention mechanism and extended pooling," *J. Appl. Remote Sens.*, vol. 14, 2020, Art. no. 044522.

[37] Y. Li, S. Zhang, and W. Q. Wang, "A lightweight faster R-CNN for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art. no. 4006105, doi: 10.1109/LGRS. 2020.3038901.

[38] W. Hu, Z. Tian, S. Chen, R. Zhan, and J. Zhang, "Dense feature pyramid network for ship detection in SAR images," in *Proc. 3rd Int. Conf. Image, Video Process. Artif. Intell.*, 2020.

[39] S. Sujin et al., "Coupling denoising to detection for SAR imagery," *Appl. Sci.*, vol. 11, no. 12, 2021, Art. no. 5569.

[40] Z. Wang, Y. Ma, and Y. Zhang, "Review of pixel-level remote sensing image fusion based on deep learning," *Inf. Fusion*, vol. 90, pp. 36–58, 2023.

[41] Y. Xu et al., "Infrared small target detection based on local contrast-weighted multidirectional derivative," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5000816, doi: 10.1109/TGRS. 2023.3244784.

[42] Q. Wang et al., "A fast facet-based SAR imaging model and target detection based on YOLOv5 with CBAM and another detection head," *Electronics*, vol. 12, no. 19, 2023, Art. no. 4039.

[43] L. Zhu et al., "BiFormer: Vision transformer with bi-level routing attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10323–10333.

[44] P. Quan et al., "Research on identification and location of charging ports of multiple electric vehicles based on SFLDLC-CBAM-YOLOV7-Tinp-CTMA," *Electronics*, vol. 12, no. 8, 2023, Art. no. 1855.

[45] M. Yasir et al., "Instance segmentation ship detection based on improved Yolov7 using complex background SAR images," *Front. Mar. Sci.*, vol. 10, 2023, Art. no. 1113669, doi: 10.1109/JSTARS.2022.3177235.

[46] 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, 2023, doi: 10.1109/JSTARS.2023.3316309.

[47] W. Wu et al., "Ship detection and recognition based on improved YOLOv7," *Comput., Mater. Continua*, vol. 76, no. 1, 2023.

[48] C. Yu et al., "Pay attention to local contrast learning networks for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3512705, doi: 10.1109/LGRS. 2022.3178984.

[49] L. Xu et al., "Remote sensing image segmentation of mariculture cage using ensemble learning strategy," *Appl. Sci.*, vol. 12, no. 16, 2022, Art. no. 8234.

[50] A. Shafique et al., "Deep learning-based change detection in remote sensing images: A review," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 871.

[51] M. Amani et al., "Ocean remote sensing techniques and applications: A review (Part II)," *Water*, vol. 14, no. 21, 2022, Art. no. 3401.

[52] X. Zhang, S. Feng, C. Zhao, Z. Sun, S. Zhang, and K. Ji, "MGSFA-Net: Multi-scale global scattering feature association network for SAR ship target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4611–4625, 2024, doi: 10.1109/JSTARS.2024.3357171.

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

[54] S. Wang, Z. Cai, and J. Yuan, "Automatic SAR ship detection based on multi-feature fusion network in spatial and frequency domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4102111, doi: 10.1109/TGRS.2023.3267495.

[55] 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, 2023, Art. no. 5205516, doi: 10.1109/TGRS.2023.3264231.

[56] 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, 2023, doi: 10.1109/dsadJSTARS.2023.3302575.

[57] M. Kang et al., "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 860.

[58] Z. Sun et al., "BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 21, 2021, Art. no. 4209.