Multilevel Pyramid Feature Extraction and Task Decoupling Network for SAR Ship Detection

Yanshan Li, Wenjun Liu, and Ruo Qi

***Abstract*—Synthetic aperture radar (SAR) target detection plays a crucial role in both military and civilian fields, attracting significant attention from researchers globally. CenterNet, a single-stage target detection method, is known for its high detection speed and accuracy by eliminating anchor-related calculations and nonmaximum suppression. However, directly applying CenterNet to SAR ship detection poses challenges due to the distinctive characteristics of SAR images, including lower resolution, lower signal-to-noise ratio, and larger ship aspect ratios. To address these challenges, we propose MPDNet. which introduces a multilevel pyramid feature extraction module (MP-FEM) to replace the encoding–decoding structure in CenterNet. MP-FEM employs multilevel pyramid and channel compression to fuse multiscale SAR image features and acquire deep features quickly. Second, we propose the convolution channel attention module, which improves the multilayer perceptron in the common pooling attention mechanism into a multistage and 1-D convolution. Therefore, the feature extraction capability of MP-FEM is further refined. Furthermore, we propose the detection task decoupling module (DTDM), which considers the characteristics of SAR ships and effectively detects smaller targets of different sizes, distinguishing the centers and sizes of densely arranged ships. DTDM extracts task-related features from the original feature map before inputting it into the three detection headers, thereby addressing the problem of task coupling in CenterNet's detection header module for SAR ship detection. Finally, the experimental results on SSDD dataset and SAR-ship-dataset show that the proposed network can significantly improve the SAR target detection accuracy.**

***Index Terms*—CenterNet, multilevel feature pyramid, synthetic aperture radar (SAR) image, target detection.**

I. Introduction

SYNTHETIC aperture radar (SAR), due to its unique imaging mechanism, enables data acquisition under all-weather and all-day conditions, unaffected by factors, such as weather and lighting [1], [2], [3]. Therefore, target detection algorithms based on SAR images find extensive applications in military fields, such as situation analysis and strategic defense, as well as in civilian fields, including marine monitoring, maritime search and rescue, and disaster monitoring [4]. Numerous scholars worldwide have conducted research on target detection methods based on SAR images [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. These algorithms improve the YOLO [9], [10], [11] and CenterNet [12], [13] networks on RGB images that are adopted for SAR image target detection [4], [5], [6], [7]. However, due to the fundamental differences in imaging principles, shooting angles, and shooting distances between SAR images and conventional optical images, the research on SAR image target detection presents numerous challenges beyond those in typical target detection. Zhang et al. [8] proposed a miniaturized plug-and-play module to select target areas from SAR images and filter out large areas of ocean and coastal backgrounds with minimal computation. Qu et al. [15] introduced transformer encoding and mask guidance modules to address issues in traditional methods, effectively learning dependencies between ship targets and reducing false alarms from complex backgrounds. Ma et al. [16], through the design of an anchor-free framework, key-point estimation module, and channel attention module, successfully alleviated challenges in detecting multiscale and dense ship targets in SAR images. Fig. 1. illustrates typical SAR images with significant challenges in target detection.

1) SAR images exhibit a vast range of target sizes and substantial variations in aspect ratios, as evidenced by the contrasting example image sets in Fig. 1(a), (b), and (e)–(h).
2) SAR images feature complex target environments with background interference from suspected targets and dense target arrangements. Examples include ships near the coastline and coastal structures, ships at sea, and small islands, as shown in Fig. 1(g) and (h).
3) SAR images suffer from low resolution and low signal-to-noise ratios, as evident in Fig. 1(c) and (d). Therefore, conventional target detection algorithms cannot be directly applied to ship detection in SAR images, highlighting the significance of developing target detection networks tailored to SAR image characteristics.

CenterNet, known for its low model complexity, fast inference speed, and capacity to extract distinct features for detecting large objects, holds great promise for SAR ship detection. Based on the characteristics of SAR images, this article proposes multilevel pyramid feature extraction and task decoupling network (MPDNet), which is a single-stage and anchor-free ship detection network. MPDNet can effectively detect SAR ships

Manuscript received 8 August 2023; revised 1 November 2023 and 13 December 2023; accepted 14 December 2023. Date of publication 17 January 2024; date of current version 24 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62076165, in part by the Innovation Team Project of Department of Education of Guangdong Province under Grant 2020KCXTD004, and in part by the Planning Project of Guangdong Province Education Sciences under Grant 2022GXJK367. *(Corresponding author: Yanshan Li.)*

Yanshan Li, Wenjun Liu, and Ruo Qi are with the ATR National Key Laboratory of Defense Technology, Shenzhen University, Shenzhen 518060, China, and also with the Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China (e-mail: lys@szu.edu.cn; axiao.boy@qq.com; qiruo123@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3347454

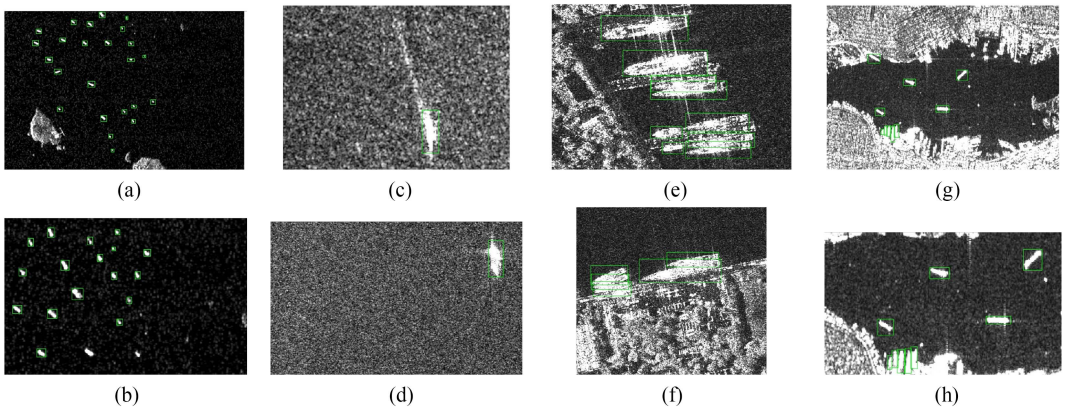

Fig. 1. (a), (b) and (e)–(h) correspond to images with significant variations in target size and large aspect ratios, where (g) and (h) depict images with complex target environments, dense target presence, and severe background interference, while (c) and (d) illustrate images with low resolution and low signal-to-noise ratio.

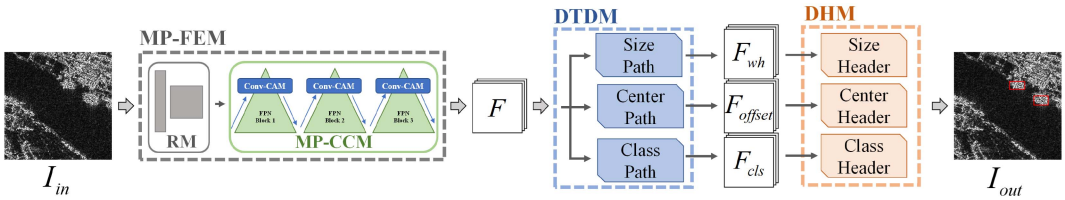

Fig. 2. Architecture of the proposed MPDNet.

with large aspect ratio and dense arrangement in SAR images featured by low resolution and low signal-to-noise ratio. The main contributions of this article are as follows.

1) A new ship detection network in SAR images (MPDNet) is proposed. The MPDNet consists of multilevel pyramid feature extraction module (MP-FEM), detection task decoupling module (DTDM) and detection header module (DHM). MP-FEM combined with convolution channel attention module (Conv-CAM) can effectively extract SAR image features with low resolution and low signal-to-noise ratio. DTDM and DHM are used to accurately detect ships with large aspect ratio and dense arrangement.
2) Considering SAR image is characterized by low resolution, low signal-to-noise ratio, and large aspect ratio of ships, MP-FEM is introduced to extract SAR image features with strong representational power. The MP-FEM is composed of the residual module (RM) and the multilevel pyramid channel compression module (MP-CCM).
3) Taking the lack of selectivity during the channel compression of MP-CCM into account, we design Conv-CAM to improve the feature extraction ability of MP-FEM.
4) DTDM is put forward to accurately detect ships with large aspect ratio and dense arrangement in SAR images. DTDM decouples the target size, center point, and class prediction tasks, respectively, thus effectively improving the detection accuracy of ships.

II. MPDNet Network

A. Structure of MPDNet

MPDNet mainly consists of three parts: MP-FEM, DTDM, and DHM. The overall network structure is shown in Fig. 2.

The MP-FEM is composed of the RM and the MP-CCM, as shown in the gray box in Fig. 2. First, the input SAR image feature I_{in} is extracted by RM, and then the feature map P with 256 channels is obtained. Second, the feature map P is further extracted and enhanced by MP-CCM, and the feature channel is compressed step by step at the same time. Each level of pyramid compression helps reduce the number of channels by half, resulting in a more refined feature map. After MP-CCM compresses the feature through the three-level pyramid structure, the final feature map F with the number of channels compressed to 64 is output.

DTDM involves size path, center path, and class path, as shown in the blue box in Fig. 2. These three paths, respectively, carry on the shunt adaptive optimization processing to the feature map F according to the task dimension so as to achieve the effect of task decoupling. After decoupling by those three paths, DTDM outputs three feature maps—F_{wh}, F_{offset}, F_{cls}, which are used to predict the size of the target, the offset distance of the center point and the target class, respectively.

The DHM consists of size header, center header, and class header, as shown in the orange box in Fig. 2. Unlike in CenterNet, where the header module has only one input, MPDNet's header has three feature maps inputs. The size header takes the output F_{wh} of size path as input, and calculates the length and width of the target by regression. The center header takes the center path output F_{offset} in DTDM as the input, and calculates the horizontal and vertical offset distance of the target center by regression. The class header takes the class path output F_{cls} in DTDM as input and calculates the probability of the target belonging to each class by regression. Finally, MPDNet aggregates the predicted output of the three headers and sorts and filters them to get the final target detection result I_{out}.

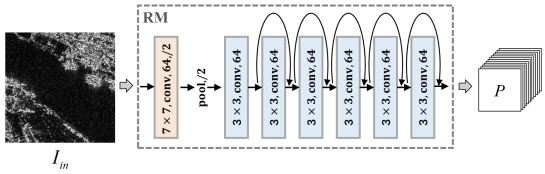

Fig. 3. (a)(b)and(e)-(h) correspond to images with significant variations in target size and large aspect ratios, where (g) and (h) depict images with complex target environments, dense target presence, and severe background interference, while (c) and (d) illustrate images with low resolution and low signal-to-noise ratio.

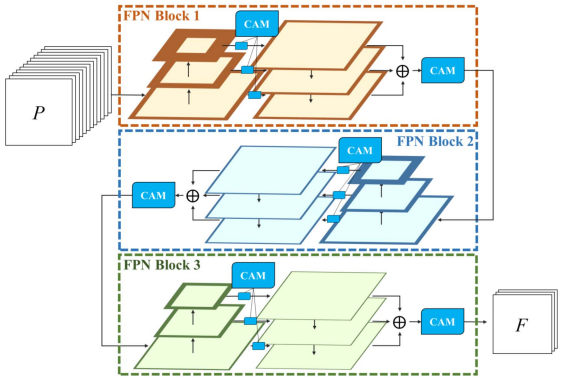

Fig. 4. Architecture of the proposed MP-CCM.

B. Multilevel Pyramid Feature Extraction Module

Considering SAR images with low resolution, low signal-to-noise ratio and ships with large aspect ratio, MP-FEM is proposed to improve feature representational ability by extracting and fusing multiscale features. The MP-FEM consists of the RM and the MP-CCM. Among them, RM contains a 7×7 convolution and a ResNet residual block [17]. The detailed network structure is shown in Fig. 3. RM carries out the feature extraction of input SAR image I_{in} in two stages, and outputs feature map P. The formula is expressed as follows:

$$P = \mathrm{f}_{\text{res2}}(\mathrm{f}_{\text{res1}}(I_{\text{in}})) \tag{1}$$

where $\mathrm{f}_{\text{res1}}(\cdot)$ and $\mathrm{f}_{\text{res2}}(\cdot)$ are feature extraction functions of two stages, respectively, and their major convolution kernel sizes are 7×7 and 3×3.

In order to further extract, enhance and compress features, we design an MP-CCM, whose network structure is shown in Fig. 4. MP-CCM takes the output feature map of RM as input, and outputs the feature map after three feature pyramid networks blocks (FPN Blocks) [18]. In addition to feature extraction and enhancement, MP-CCM also completes the feature map compression. That is to say, the channel of the feature map is compressed by half after each FPN Block. Through the compression of three FPN Blocks, the feature map with 256 channels is compressed into with 64 channels. The processing of feature map by FPN Block consists of five parts: feature extraction from bottom up, feature extraction from top down, lateral connection operation, additive fusion feature, and channel attention emphasis feature. To begin with, the feature extraction capability of FPN Block is mainly realized by its bottom-up operation. One block contains two bottom-up operations, each of which doubles the number of channels of the feature map while reduces the length and width of the feature map to half, as described in the following:

$$\begin{cases} P_{i\_1} = P_{i-1} \\ P_{i\_k} = \mathrm{f}_{\text{bottom2up}}\left(P_{i\_k-1}\right) \end{cases} \quad i=1,2,3; \quad k=1,2,3; \quad P_0 = P \tag{2}$$

where $\mathrm{f}_{\text{bottom2up}}(\cdot)$ is the bottom-up transformation function, consisting of the convolution layer, the batch normalization layer (BN layer) and the pooling layer; and P_{i_k} represents the kth layer feature diagram of the bottom-up process of the ith FPN Block.

Then, the top-down and lateral connection operation in FPN blocks fuse high-level semantic features with high-resolution spatial features to achieve feature map enhancement. Among them, the topmost feature map C_{i_3} is obtained from the topmost feature map P_{i_3} in the bottom-up process after one lateral change. It is shown in the following:

$$C_{i\_3} = \mathrm{f}_{\text{lateral}}(P_{i\_3}) \tag{3}$$

where $\mathrm{f}_{\text{lateral}}(\cdot)$ is the lateral change function, the embodiment of the essential difference between the whole FPN Block and the ordinary FPN. The input feature map is scaled up by $\mathrm{f}_{\text{lateral}}(\cdot)$ to the same size as P_{i_1}. Therefore, in addition to common convolution, transposed convolution is also involved. Besides, $\mathrm{f}_{\text{lateral}}(\cdot)$ is the core structure to complete channel compression, which compresses the number of channels to half of that of P_{i_1}. The proposed Conv-CAM also participates in this process, which will be detailed in Section C. Different from the topmost feature map, the generation of the middle layer and the bottom layer feature map in the top-down process requires both lateral connection and top-down feature extraction, as shown in the following:

$$C_{i\_k} = \mathrm{f}_{\text{latera1}}(P_{i\_k}) + \mathrm{f}_{\text{up2down}}(C_{i\_k+1}) \quad i=1,2,3; \quad k=1,2 \tag{4}$$

where $\mathrm{f}_{\text{up2down}}(\cdot)$ is a top-down operation function, mainly composed of a 3×3 convolution, BN layer and ReLu activation layer, and C_{i_k} represents the kth layer feature diagram of the top-down process in the ith FPN Block.

Finally, FPN Block completes the aggregation and fusion of various scale information. Specifically, the three feature maps C_{i_1}, C_{i_2}, and C_{i_3} generated from the top-down process are added pixel by pixel, and the calculated results are further emphasized by Conv-CAM. It is illustrated in the following:

$$F_i = \mathrm{f}_{\text{CAM}}\left(\sum_{k=1}^{3} C_{i\_k}\right) \tag{5}$$

where $\mathrm{f}_{\text{CAM}}(\cdot)$ represents the channel attention function, F_i is the final output feature map of the ith FPN Block. MP-CCM consists of three FPN blocks, and its feature map and channel changes are as follows:

$$\begin{cases} P \to F_1 \to F_2 \to F_3 \to F \\ 256 \to 256 \to 128 \to 64 \to 64. \end{cases} \tag{6}$$

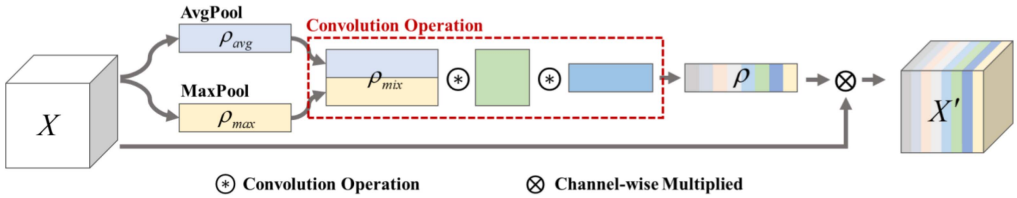

Fig. 5. Architecture of the proposed Conv-CAM.

It is worth noting that the first FPN Block of MP-CCM does not carry out channel compression, which ensures that the network will not be compressed until sufficient features are extracted. As a result, it is beneficial to avoid inadequate feature extraction. The feature map F is the final output of MP-CCM and also the final output of MP-FEM.

C. Convolution Channel Attention Module

In order to selectively extract features, emphasize useful features, and eliminate interfering features when MP-CCM compresses channels, Conv-CAM is put forward, whose structure is shown in Fig. 5.

First, Conv-CAM conducts mean-pooling and max-pooling operations on the input feature map X to obtain ρ_{avg} and ρ_{max}. Different from adding the pooling feature map of CBAM [19], this article concatenates ρ_{avg} and ρ_{max} to obtain mixed weight map ρ_{mix}, as shown in the following:

$$\rho_{\text{mix}} = \text{concat}(\rho_{\text{avg}}, \rho_{\text{max}}) = \text{concat}(\text{f}_{\text{avg\_pool}}(X), \text{f}_{\text{max\_pool}}(X)) \tag{7}$$

where $\text{concat}(\cdot)$ denotes the concatenation function according to the vertical direction of the tensor, $\text{f}_{\text{avg\_pool}}(\cdot)$ and $\text{f}_{\text{max\_pool}}(\cdot)$ are the average channel pooling function and the maximum channel pooling function, respectively.

Second, Conv-CAM replaces the MPL layer in the channel attention module of CBAM with two 1-D convolutions. This change not only reduces the number of parameters and the amount of computation, but also makes full use of the prior knowledge of relevant information of adjacent channels. Therefore, it is more conducive to mining important feature information. After two convolution operations, Conv-CAM maps ρ_{mix} into a 1-D weight map ρ. It is shown in the following:

$$\rho = \text{conv1d}_3(\text{conv1d}_7(\rho_{\text{mix}})) \tag{8}$$

where $\text{conv1d}_k(\cdot)$ is the 1-D convolution function, whose subscript k is the convolution kernel size.

Finally, Conv-CAM multiplies the weight map ρ with the original input feature map, and then emphasizes and cull the original feature map in channel dimension to obtain the final output feature map X', as expressed in the following:

$$X' = \rho \times X. \tag{9}$$

D. Detection Task Decoupling Module

SAR ships are characterized by small size, dense arrangement and large aspect ratio, so it is required that the target detection network can effectively detect the smaller targets of different sizes and distinguish the center and size of the densely arranged ships. However, the input feature maps of the size header, center header, and class header in CenterNet are the same. And the DHM only performs simple convolution operation with a convolution kernel of 3 before the output. As a result, the DHM has serious task coupling, and the headers of different tasks have a great influence on each other during parameter updating. Therefore, it is not conducive to the regression convergence of each task and it reduces the detection accuracy of the targets, especially of the small targets. This article holds that although the prediction of these three tasks is regression calculation, there are great differences among them for they belonging to different types of regression tasks. Extracting corresponding features for different tasks can effectively improve the detection accuracy of each task.

Therefore, we introduce detection task decoupling model (DTDM). Before the feature map is input into three detection headers, task-related features will be further extracted to send to the size header, center header, and class header, respectively, to achieve task decoupling.

As shown in the blue boxes in Fig. 2, the DTDM has three paths that modify, align, and optimize the input feature maps of the size header, center header, and class header according to the different prediction tasks. Fig. 6 expresses the specific structure of the three paths. Block A is center path, block B the size path, block C class path, and block D is the concrete structure display of the modules used in the first three network structures. There are similarities and differences among the three detection tasks. Accordingly, the three paths of DTDM share similarities and differences in structural design. In terms of common ground, the three paths all use multistream structure, residual connections and feature fusion mechanism. First, DTDM enhances the feature map by distinguishing different angles and receptive fields through multistream structures. Then, the multistream structure is aggregated and fused through the channel dimension concatenation and fusion, spatial pixel dimension addition and fusion, and residual jump connection. Finally, the comprehensive and refined feature map is output. In regard to the difference, these three paths do corresponding decoupling design according to the characteristics of their specific tasks.

1) *Center path:* For the offset distance predication task of the target center point, the size, span, offset sensitivity, and the size of the receptive field of the targets to be detected are different. That is to say, the size of the convolution kernel used to extract the feature should also be different. Therefore, the scheme of multiscale convolution kernel is adopted, as shown in block A of Fig. 6.

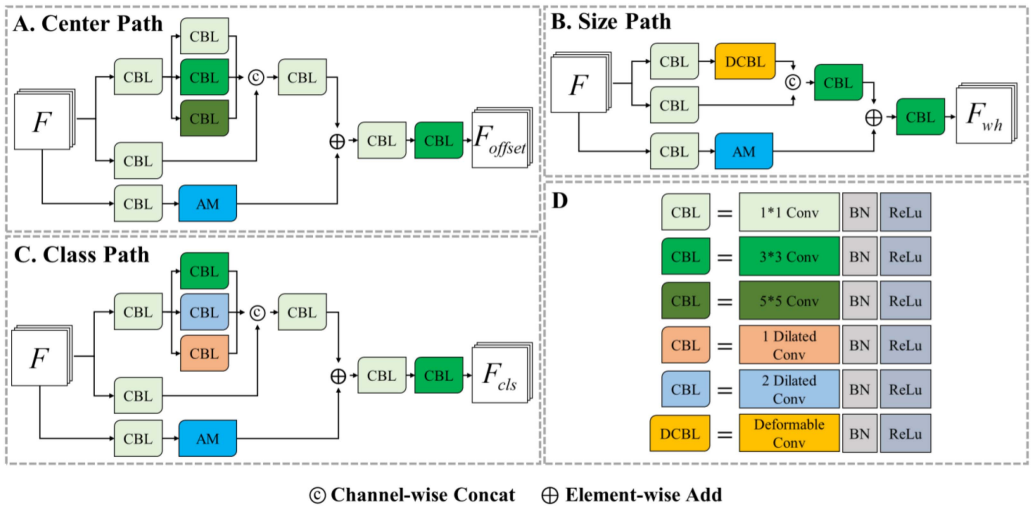

Fig. 6. Structure of three paths of DTDM.

First of all, in order to reduce the amount of computation, it is necessary to downsample the channel of input feature map F to obtain the lightweight feature map F_{cp} for center path. It is shown in the following:

$$F_{\text{ctp}} = \text{f}_{\text{chn\_down}}(F) \tag{10}$$

where $\text{f}_{\text{chn\_down}}(\cdot)$ is a channel downsampling function formed by a 2-D convolution with the kernel size of 1.

Then, center path uses convolution kernels with sizes of 1, 3, and 5 in the multistream structure to shunt, and obtains three shunt feature maps $F_{\text{ctp\_}s1}$, $F_{\text{ctp\_}s2}$, and $F_{\text{ctp\_}s3}$, which has small, medium, and large receptive fields, respectively, as expressed in the following:

$$\begin{cases} F_{\text{ctp\_}s1} = \text{conv}_{1\times1}(F_{\text{ctp}}) \\ F_{\text{ctp\_}s2} = \text{conv}_{3\times3}(F_{\text{ctp}}) \\ F_{\text{ctp\_}s3} = \text{conv}_{5\times5}(F_{\text{ctp}}) \end{cases} \tag{11}$$

where $\text{conv}_{k\times k}(\cdot)$ denotes the combination of the convolutional layer, BN layer, and ReLu activation layer, and its subscript k denotes the size of the convolutional kernel.

Next, the shunt structure is aggregated through the concatenation of channel dimensions, and the multistream aggregation feature map $F_{\text{ctp\_ms}}$ is obtained, which has a wide receptive field, as shown in the following:

$$F_{\text{ctp\_ms}} = \text{f}_{\text{chn\_adjust}}(\text{concat}(F_{\text{ctp\_}s1}, F_{\text{ctp\_}s2}, F_{\text{ctp\_}s3}, F_{\text{ctp}})) \tag{12}$$

where $\text{f}_{\text{chn\_adjust}}(\cdot)$ is the channel adjustment function and $\text{concat}(\cdot)$ is the channel dimension concatenation function.

Finally, the original lightweight feature map F_{ctp} is enhanced by the attention mechanism. Besides, it is aggregated and fused with the multistream aggregation feature map $F_{\text{ctp\_ms}}$. F_{offset} is output, as shown in the following:

$$F_{\text{offset}} = \text{f}_{\text{fuse}}(F_{\text{ctp\_ms}} + \text{f}_{\text{am}}(F_{\text{ctp}})) \tag{13}$$

where $\text{f}_{\text{fuse}}(\cdot)$ is the feature alignment fusion function composed of 3×3 convolution, $\text{f}_{\text{am}}(\cdot)$ is the attention module composed of CBAM, and F_{offset} is the output of the center path.

2) *Size path:* For the target size prediction task, the detector needs to obtain the target boundary information, whose specific structure is shown in block B of Fig. 6. First, like center path, size path downsamples the channel dimension of input feature map F to obtain the lightweight feature map F_{sp}. It is expressed in the following:

$$F_{\text{sp}} = \text{f}_{\text{chn\_down}}(F). \tag{14}$$

Second, the method of deformable convolution [20] is used to obtain the enhanced feature map that can adapt to targets of different shapes. The conact function is adopted to concatenate the enhanced feature map with F_{sp}. And the convolution layer is utilized to align their fusion features. Thus, the multistream aggregation feature map $F_{\text{sp\_ms}}$ of Size Path is obtained. It is shown in the following:

$$F_{\text{sp\_ms}} = \text{conv}_{3\times3}(\text{concat}(\text{dfconv}_{3\times3}(F_{\text{sp}}), F_{\text{sp}})) \tag{15}$$

where $\text{dfconv}_{k\times k}(\cdot)$ denotes deformable convolution functions, and its subscript k denotes the size of the convolution kernel.

Finally, the lightweight feature map F_{sp} is enhanced by the attention mechanism and aggregated as well as fused with $F_{\text{cp\_ms}}$, as shown in the following:

$$F_{\text{wh}} = \text{f}_{\text{fuse}}(F_{\text{sp\_ms}} + \text{f}_{\text{am}}(F_{\text{sp}})) \tag{16}$$

where F_{wh} is the final output feature map of size path.

3) *Class path:* For the target class prediction task, the detector needs to obtain the texture and contour of the target and other specific details, whose specific structure is shown in the block C in Fig. 6. The reasoning process of class path is similar to that of center path.

First, the class path undersamples the channel dimension of the input feature map F to obtain the lightweight feature

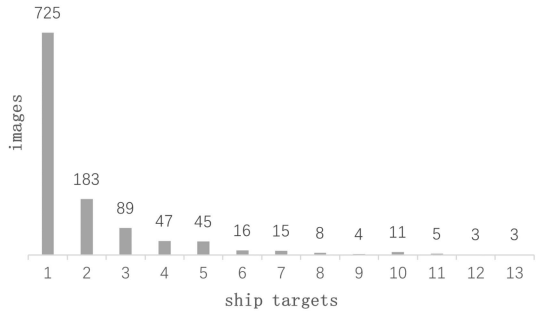

Fig. 7. Number of images and ships in SSDD dataset.

map F_{clsp}. It is expressed in the following:

$$F_{\text{clsp}} = \text{f}_{\text{chn\_down}}(F). \tag{17}$$

Second, dilated convolution [21] with a dilated rate of 0, 1, and 2 is used to build a multistream structure of class path, which is conducive to obtaining high-frequency spatial structure information of targets, as shown in the following:

$$\begin{cases} F_{\text{clsp\_}s1} = \text{dlconv}_0(F_{\text{clsp}}) \\ F_{\text{clsp\_}s2} = \text{dlconv}_1(F_{\text{clsp}}) \\ F_{\text{clsp\_}s3} = \text{dlconv}_2(F_{\text{clsp}}) \end{cases} \tag{18}$$

where $\text{dlconv}_r(\cdot)$ represents the dilated convolution function, and its subscript r represents the dilated rate. Then, the multistream aggregation feature map $F_{\text{clsp\_ms}}$ is obtained by aggregating the shunt structures. It is expressed in the following:

$$F_{\text{clsp\_ms}} = \text{f}_{\text{chn\_adjust}}(\text{concat}(F_{\text{clsp\_}s1}, F_{\text{clsp\_}s2}, F_{\text{clsp\_}s3}, F_{\text{clsp}})). \tag{19}$$

Finally, the feature map F_{clsp} emphasized by the attention mechanism is aggregated and fused with $F_{\text{clsp\_ms}}$, as shown in the following:

$$F_{\text{cls}} = \text{f}_{\text{fuse}}(F_{\text{clsp\_ms}} + \text{f}_{\text{am}}(F_{\text{clsp}})) \tag{20}$$

where F_{cls} is the final output feature map of the class path.

III. Experiment Results and Analysis

A. Datasets and Evaluation Indicators

Considering the diversity and universality of the dataset, SSDD dataset [22] is selected to test the effect of our proposed algorithm. The characteristics of SSDD dataset are as follows.

1) Abundant data sources. The images in SSDD dataset are from RadarSat2, TerraSARX, and Sentinel-1 data sources. It contains 1160 images and 2456 ships in total, meaning that each image contains 2.12 ships on average. The specific distribution is shown in Fig. 7.
2) Large resolution span and large size span. Image resolution varies from 1 to 15 m, with the smallest object having 7×2 pixels and the largest object having 368×69 pixels.
3) Diverse target background and distribution. Such as: near shore, far shore, multiship dense arrangement, multiship sparse distribution, and so on.
4) Wide versatility. At present, lots of relevant studies are carried out on SSDD dataset to verify the validity of the different proposed models. Mao et al. [23] verified the effectiveness of the advanced algorithms on SSDD dataset. Our experiments will be based on their findings.

Referring to the work of Mao et al. [23], in this article, images with file names ending in numbers 1 and 9 are used as the test set, while the remaining images are used as the training set. We get a test set of 232 images and a training set of 928 images.

This article adopts the same evaluation indicator as Mao et al.: MS COCO evaluation matrix [24]. In order to verify the effectiveness of the proposed method in detecting multiscale targets, especially to verify the ability of the MP-FEM to extract multiscale features, four indicators from the MS COCO evaluation matrix, AP^{50}, AP^S, AP^M, and AP^L, are selected as the experimental validation criteria, among which the most important indicator is AP^{50}.

TABLE I
Results of the Comparison Experiment Between MPDNet and Baseline

Method	AP^{50}	AP^S	AP^M	AP^L	Params(M)
CenterNet (baseline)	0.904	0.393	0.602	0.647	34.7
MPDNet (ours)	**0.950**	**0.470**	**0.636**	**0.687**	**30.7**

B. Setting

In this article, pytorch framework is adopted to implement the proposed algorithm. The version of torch is 1.10.1+cu111 and torchvision is 0.11.2+cu111. The program is trained and tested on a 64-bit Linux system and accelerated using a 24 G GeForce RTX 3090 GPU. In the training process, the minimum value of learning rate is 2.5×10^{-6}, the maximum value is 5×10^{-4}, and the initial value is 2.5×10^{-4}. The learning rate is updated by combining the exponential trend and the cos function trend. At the initial stage of training, that is, the first 5% of the total iteration, the model belongs to the "warmup stage" [25]. The learning rate increases exponentially until it reaches the maximum learning rate, and then the maximum learning rate is maintained for further training. At the middle and late stages of training, that is, the last 95% of the total iterations, the models belongs to the "annealing stage" [25]. The learning rate decreases in monotonously decreasing tendency with respect to cos function until it reaches the minimum learning rate, which is maintained for all subsequent training iterations.

C. Performance Evaluation

1) Comparison Experiments Between MPDNet and CenterNet: We conduct a comparison experiment between MPDNet and CenterNet to verity the effectiveness of our proposed algorithm. The experimental results are shown in Table I, where the bold column of Method represents the method in this article, and each bold indicator is the optimal result in the experiment.

As shown in Table I, MPDNet greatly improves in all indicators compared with CenterNet. The most important indicator AP^{50} increases by 4.6%, the average precision of small targets increases by 7.7%, the average precision of medium targets AP^M increases by 3.4%, and the average precision of large targets AP^L increases by 4%. It is obvious that compared with

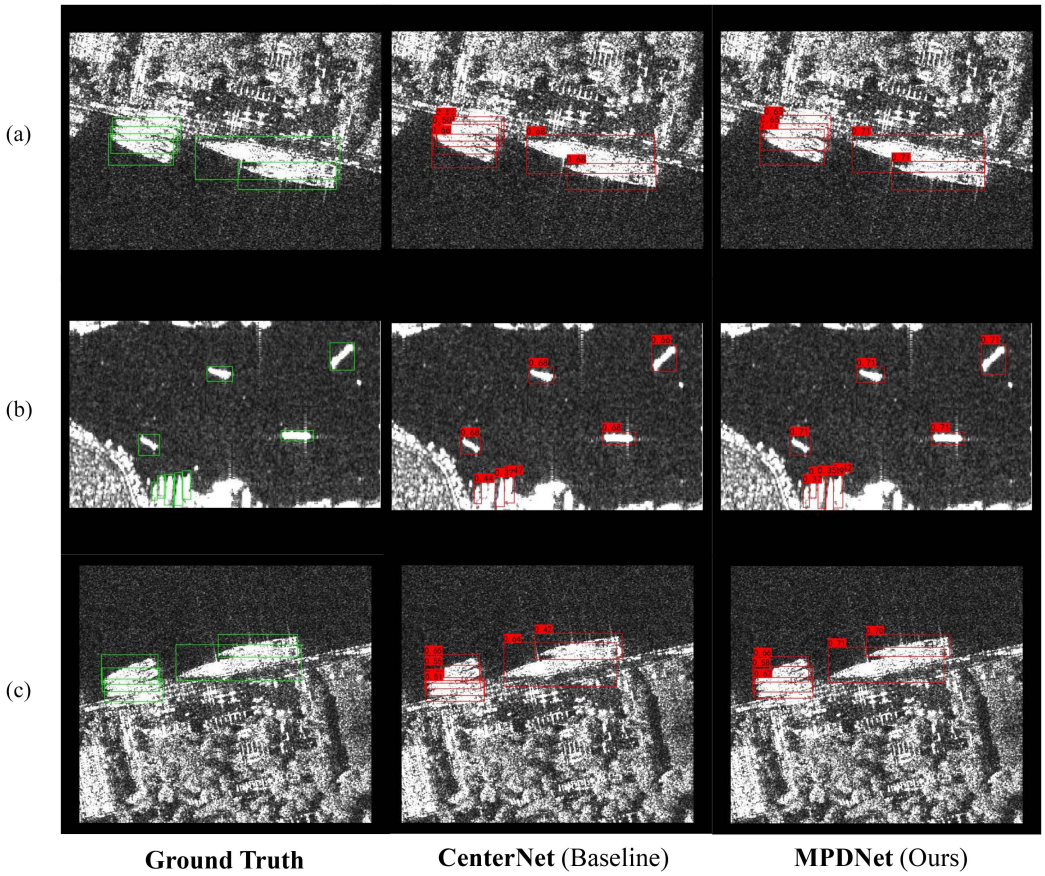

Fig. 8. Visual comparison of test results in the nearshore and densely arranged scenes in SAR images.

CenterNet, the proposed MPDNet effectively solves the difficulties of target detection in SAR images and comprehensively improves the detection accuracy. In order to further verify and analyze the effectiveness of the proposed method, visual analysis of detection results is also carried out on five representative SAR images in this section, as shown in Figs. 8 and 9.

Fig. 8 shows the detection visualization results in the nearshore and densely arranged scene of the targets. There are three rows (a), (b), and (c) from top to bottom, showing three representative images, respectively, and three columns from left to right, representing detection results of ground-truth, CenterNet, and MPDNet. Ground-truth uses green boxes to mark the target location. CenterNet and MPDNet use red boxes to mark the target location, and the upper left corner of the red box shows the target confidence, ranging from 0 to 1. When CenterNet is dealing with nearshore and densely arranged targets, its detection results produce false alarms due to the interaction between suspected target buildings on shore and target detection. The three ships on the left in Fig. 8(a) and (c) are detected as four ships by CenterNet, while MPDNet can accurately detect three ships. In addition, when the size of nearshore and densely arranged targets is small, the detection results of CenterNet is missed due to the lack of feature extraction capability of backbone. In Fig. 8(b), CenterNet misses the middle two targets, while MPDNet could accurately detect the five targets.

Fig. 9 shows the visualization results of detection under the multiscale target scene. Fig. 9(a) contains a large number of small targets, and (b) contains two large targets, both of which jointly verify the detection results of MPDNet for multiscale targets. Due to CenterNet's weak feature extraction ability and lack of feature enhancement process, it is difficult to deal with multiscale scenes, leading to false alarm detection. CenterNet detects false alarms in the upper right corner of the target in Fig. 9(a) and false alarms in the upper left corner of the target in Fig. 9(b). In contrast, our proposed method handles those problems well. But MPDNet also has defects. Because of the influence of suspected target buildings on shore, the detection results of MPDNet also have false alarms, as shown in the lower right corner of the MPDNet detection results in Fig. 9(b).

2) Comparison Experiments Between MPDNet and Other Single-Stage Methods: MPDNet is a single-stage target detection network. We conduct comparison experiments between MPDNet and other representative single-stage methods,. The results are shown in Table II. The comparison methods are: FCOS [27], SSD [28], YOLOv3 [29], YOLOv7 [30], RetinaNet GA [31], Reppoints Moment [32], Fovea Align [33], Deformable_DETR [34], PVT [35], and PyCenterNet [36]. PyCenterNet, proposed by Duan et al., is an enhanced bottom-up CenterNet variant that detects each object as a triplet of keypoints, enabling it to locate objects with arbitrary geometries and perceive global information within objects.

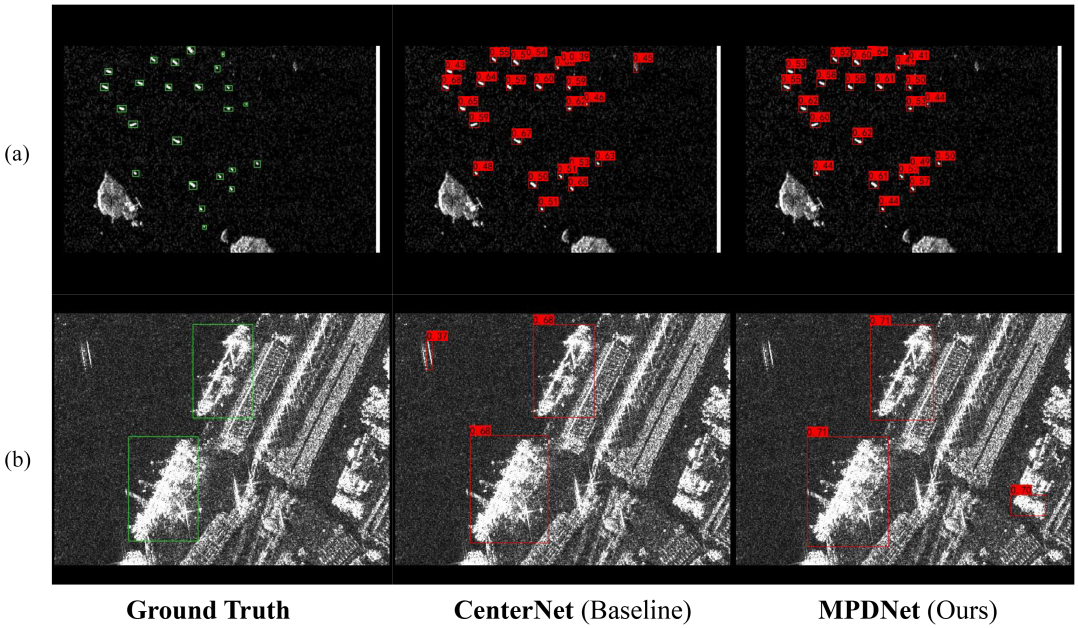

Fig. 9. Visual comparison of test results in multiscale target scene in SAR image.

TABLE II
COMPARISON BETWEEN MPDNET AND OTHER SINGLE-STAGE ALGORITHMS (THE BOLD REPRESENTS THE OPTIMAL RESULT AND THE UNDERLINED REPRESENTS THE SUBOPTIMAL RESULT)

Method	AP^{50}	AP^{S}	AP^{M}	AP^{L}	Params(M)
FCOS[26]	0.820	0.429	0.411	0.264	31.8
Reppoints Moment[31]	0.897	0.521	0.576	0.484	36.6
Retinanet GA[30]	0.900	0.462	0.554	0.444	36.1
Fovea Align[32]	0.933	**0.550**	0.630	0.452	36.0
SSD[27]	0.937	0.523	0.618	0.635	**23.8**
Deformable_DETR[33]	0.901	0.436	0.623	0.524	40.0
PVT[34]	0.925	0.517	0.591	0.551	32.6
YOLOv3[28]	0.942	0.500	0.565	0.382	62.0
YOLOv7[29]	0.946	0.524	0.630	0.511	37.6
PyCenterNet[35]	0.925	0.452	0.614	0.663	33.6
CenterNet (baseline)	0.904	0.393	0.602	0.647	34.7
MPDNet (ours)	**0.950**	0.470	**0.636**	**0.687**	30.7

As shown in Table II, in terms of the detection of small targets, although MPDNet has a great improvement compared with CenterNet, CenterNet and MPDNet are slightly less effective than other algorithms. Lin et al.'s method, Fovea Align, gets the best results in AP^S. CenterNet decreases 15.7% and MPDNet falls 8% compared to Fovea Align. This is because CenterNet takes targets as points for detection. When the network depth is too large, small targets will be lost on the feature map, resulting in missing detection. Compared with CenterNet, MPDNet improves by 7.7% in AP^S, which indicates that the proposed algorithm alleviates the problem of small target missing detection.

Meanwhile, MPDNet achieves 95% in AP^{50}, the highest of all comparison methods. This shows that detection effect of MPDNet is greatly improved thanks to the refinement, enhancement and compression of feature maps by multilevel pyramid and channel attention mechanism, as well as the alignment optimization of task decoupling modules. Therefore, MPDNet is very competitive compared with other single-stage methods in terms of the comprehensive performance.

In addition, for medium and large targets detection, MPDNet also achieves the best results. AP^M and AP^L are 63.6% and 68.7%, respectively. It should be noted that compared with other single-stage methods, MPDNet does the best job in multiscale feature extraction and balance, showing strong comprehensive detection performance.

3) Comparative Experiments Between MPDNet and Other Representative Single-Stage Methods in the SAR-Ship-Datasets: To further validate the model's generalization performance, we conducted a comparative analysis of MPDNet against several representative single-stage methods using the extensive SAR ship detection dataset, SAR-ship-dataset. The SAR-ship-dataset, created by researchers from the Institute of Electronics at the Chinese Academy of Sciences, is designed for deep learning-based SAR ship detection.[1] This dataset comprises 102 images from the Gaofen-3 (GF-3) satellite and 108 images from the Sentinel-1 satellite, all of which have been meticulously annotated. Within this dataset, you will find 39 729 ship chips, each measuring 256 pixels, showcasing variations in scale and background. The dataset is thoughtfully partitioned into a training set and a test set, with a 4:1 ratio between them.

To verify the superiority of the MPDNet proposed in this article, we compared it to several representative approaches, including the transformer-based method Deformable_DETR, YOLOv7, the baseline method CenterNet, and other enhanced techniques, such as PyCenterNet. These methods are all anchor-free object detection approaches. The results are presented in Table III.

MPDNet achieved an outstanding result of 95.9% on the AP^{50} metric. Thanks to a series of enhancements specifically designed for SAR images, MPDNet demonstrates a

TABLE III
COMPARISON BETWEEN MPDNET AND OTHER REPRESENTATIVE SINGLE-STAGE ALGORITHMS USING SAR-SHIP-DATASET (THE BOLD REPRESENTS THE OPTIMAL RESULT AND THE UNDERLINED REPRESENTS THE SUBOPTIMAL RESULT)

Method	AP^{50}	AP^{S}	AP^{M}	AP^{L}	Params(M)
Deformable_DETR[33]	0.922	0.489	0.619	0.573	40.0
YOLOv7[29]	0.925	**0.562**	0.671	0.674	37.6
PyCenterNet[35]	<u>0.947</u>	0.481	<u>0.630</u>	<u>0.702</u>	<u>33.6</u>
CenterNet (baseline)	0.933	0.424	0.597	0.671	34.7
MPDNet (ours)	**0.959**	<u>0.535</u>	**0.674**	**0.741**	**30.7**

TABLE IV
VERIFICATION EXPERIMENT RESULTS OF THE PROPOSED MP-FEM (THE BOLD REPRESENTS THE OPTIMAL RESULT AND THE UNDERLINED REPRESENTS THE SUBOPTIMAL RESULT)

Backbone	AP^{50}	AP^{S}	AP^{M}	AP^{L}
ConvNeXt-S	0.900	<u>0.411</u>	0.580	0.633
ResNet50 (baseline)	<u>0.904</u>	0.393	<u>0.602</u>	<u>0.647</u>
MP-FEM (ours)	**0.928**	**0.413**	**0.633**	**0.734**

TABLE V
VERIFICATION EXPERIMENT RESULTS OF THE PROPOSED CONV-CAM (THE BOLD REPRESENTS THE OPTIMAL RESULT AND THE UNDERLINED REPRESENTS THE SUBOPTIMAL RESULT)

Method	AP^{50}	AP^{S}	AP^{M}	AP^{L}
ResNet50 (baseline)	0.904	0.393	0.602	0.647
MP-FEM (ours)	<u>0.928</u>	<u>0.413</u>	<u>0.633</u>	**0.734**
MP-FEM (ours) + Conv-CAM (ours)	**0.929**	**0.417**	**0.637**	<u>0.714</u>

TABLE VI
VERIFICATION EXPERIMENT RESULTS OF THE PROPOSED DTDM (THE BOLD REPRESENTS THE OPTIMAL RESULT AND THE UNDERLINED REPRESENTS THE SUBOPTIMAL RESULT)

Method	AP^{50}	AP^{S}	AP^{M}	AP^{L}
CenterNet	0.904	0.393	0.602	0.647
CenterNet + DTDM (ours)	**0.910**	**0.409**	**0.613**	**0.699**

substantial improvement in detection performance compared to the baseline network. In the domain of small object detection, YOLOv7 yielded the best results, with MPDNet achieving the second-best performance. Compared to CenterNet, MPDNet exhibited an 11.1% improvement in AP^S, addressing the issue of small object feature loss that was prevalent in CenterNet. For medium and large object detection, MPDNet also delivered the best results with AP^M and AP^L reaching 67.4% and 74.1%, respectively. MPDNet excelled in multiscale feature extraction and balance compared to other single-stage methods. Consequently, when considering comprehensive performance, MPDNet exhibits strong competitiveness when compared to other single-stage methods.

D. Ablation Experiments

1) Verification Experiments of MP-FEM: In order to verify the effect of MP-FEM, two backbone networks, ResNet50 and ConvNeXt-S[37], are chosen for comparison. ResNet50 is the backbone adopted by CenterNet original text. ConvNetXt was put forward in 2022, which refers to the structure design and training method of transformer network. It makes a series of improvements on the basis of ResNet50. With a very small number of parameters and computation, ConvNetXt achieves better results than transformer on the ImageNet-1 K dataset. Here, we choose ConvNeXt-S, which has the same number of ResNet50 parameters, for comparison.

The experimental results demonstrating the impact of the MP-FEM enhancement module on the final outcomes are presented in Table IV. These results unequivocally showcase the outstanding detection capabilities of MP-FEM when applied to the SSDD dataset. Notably, the most critical metric, AP^{50}, exhibits a noteworthy increase of 2.4%, while AP^S and AP^M show respective improvements of 2% and 3.1%. Impressively, the metric AP^L demonstrates a substantial increase of 8.7%.

The above-mentioned experimental results show that MP-FEM has very strong feature extraction ability. Specifically, MP-FEM can mine multiscale target information from images. With MP-CCM, the extracted multiscale target feature map is extracted and compressed by multistage FPN Block, so that the final output feature map contains multiscale information. It effectively solves the detection difficulties caused by the ships with large aspect ratio and complex background in SAR images. Besides, it also addresses the problem of missing detection of small targets and false alarm of large targets caused by the backbone network of single-scale feature extraction. Therefore, the comprehensive detection effect of the network has been greatly improved.

2) Verification Experiments of Conv-CAM: Experimental results of Conv-CAM are shown in Table V. The table reveals the impact of integrating Conv-CAM into the MP-FEM, resulting in a significant improvement in the final detection performance indicators. Specifically, it is observed that the inclusion of Conv-CAM results in a 0.1% increase in AP^{50}, a 0.4% increase in AP^S, an additional 0.4% gain in AP^M, while there is a 2% decrease in AP^L.

In conclusion, Conv-CAM enables MP-FEM to selectively screen and fuse multiscale features in channel dimensions, and extract features of various scales more efficiently and accurately. Therefore, it makes the overall network performance more balanced and the detection effect better.

3) Verification Experiments of DTDM: The experimental results of DTDM are shown in Table VI. In this context, CenterNet is the baseline model, and CenterNet with DTDM refers to the task-decoupled CenterNet based on our proposed DTDM. The effects of our proposed enhancements on the model's final quantitative results are demonstrated in Table VI, showing notable improvements in the performance metrics of CenterNet when influenced by the task decoupling impact of DTDM. The most crucial metric, AP^{50}, has increased by 0.6%, while AP^S has seen a 1.6% improvement, AP^M has increased by 1.1%, and AP^L has witnessed a significant increase of 5.2%.

The obvious performance improvement shows that the DTDM can distinguish different types of prediction tasks better. Before the feature map is input to the corresponding detection header, the feature map can be modified and optimized according to the characteristics of the prediction task. Therefore, the feature map in line with the characteristics of the task can be generated

for different detection headers to maximize the role of the detection header. The proposed DTDM can effectively detect smaller targets of different sizes and distinguish center and size of densely arranged ships.

IV. Discussion

We conducted research on target detection algorithms in SAR images, starting from the detection concept of "treating targets as points" in CenterNet, and designed the MPDNet detection network. MPDNet exhibits several advantages as follows.

1) MPDNet possesses robust feature extraction and enhancement capabilities. It can extract multiscale features from SAR images, effectively addressing the challenge of detecting targets with varying sizes in SAR images.
2) MPDNet can acquire more accurate target features. In addition, it considers the contextual information of target backgrounds, enabling it to handle cases where SAR targets near the coastline may be affected by suspected shore-based objects.
3) MPDNet offers higher resolution, allowing for precise detection of densely arranged targets. In terms of overall performance, MPDNet outperforms CenterNet and other mainstream single-stage target detection algorithms when applied to SAR images. However, MPDNet still has some limitations.

One of the significant advantages of center-point-based detection methods is their lightweight nature. However, they often exhibit poorer performance in detecting small targets. SAR images are frequently captured from high-altitude platforms or satellites, resulting in smaller target pixel sizes, posing a considerable detection challenge. In the future, we intend to address this specific challenge by conducting further research.

V. Conclusion

In this article, MPDNet is proposed. It could effectively solve the problem that the CenterNet-based model is still difficult to achieve good results under the conditions of images with low resolution and low signal-to-noise ratio and ships with large aspect ratio and dense arrangement. The proposed MPDNet mainly consists of MP-FEM, Conv-CAM, and DTDM. First, MP-FEM carries out feature extraction, enhancement and compression of SAR images, and extracts multilevel features. It deals with the problems that SAR image is characterized by low resolution, low signal-to-noise ratio, and large aspect ratio. Second, Conv-CAM is embedded into the channel compression process of MP-CCM, making the process of refining and compression feature map more selective. Third, DTDM decouples the target size prediction task, the target center offset distance prediction task and the target class prediction task. Therefore, the proposed network can effectively detect the smaller targets of different sizes and distinguish the center and size of the densely arranged ships. Finally, all proposed methods are experimentally verified on SSDD dataset and are compared with other single-stage methods. Furthermore, additional validation was conducted using the SAR-ship-dataset. The results show that the detection performance of MPDNet is significantly improved compared with CenterNet. MPDNet also achieves the best results in several indicators compared with other mainstream algorithms, among which AP^{50} reaches 95.0%.

References

[1] Q. Huang, W. Zhu, Y. Li, B. Zhu, T. Gao, and P. Wang, "Survey of target detection algorithms in SAR images," *Proc. IEEE 5th Adv. Inf. Technol., Electron. Automat. Control Conf.*, Chongqing, China, 2021, pp. 1756–1765, doi: 10.1109/IAEAC50856.2021.9390728.

[2] Y. Li, W. Zhu, and Q. Huang, "A review of SAR image object detection method," *Ordnance Ind. Automat.*, vol. 12, no. 40, 2012, Art. no. 12.

[3] X. Hou, G. Jin, and L. Tan, "A review of ship target detection in SAR images based on deep learning," *Adv. Laser Optoelectron.*, vol. 58., no. 4, pp. 53–64, 2021.

[4] T. Zhang et al., "Balance learning for ship detection from synthetic aperture radar remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 182, pp. 190–207, 2021, doi: 10.1016/j.isprsjprs.2021.10.010.

[5] T. Zhang et al., "HyperLi-Net: A hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 123–153, 2020, doi: 10.1016/j.isprsjprs.2020.05.016.

[6] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Nov. 2019, doi: 10.1109/TGRS.2019.2923988.

[7] Y. Gui, X. Li, and L. Xue, "A multilayer fusion light-head detector for SAR ship detection," *Sensors*, vol. 19, no. 5, pp. 1124, 2019, doi: 10.3390/s19051124.

[8] J. Zhang, W. Sheng, H. Zhu, S. Guo, and Y. Han, "MLBR-YOLOX: An efficient SAR ship detection network with multilevel background removing modules," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5331–5343, 2023, doi: 10.1109/JSTARS.2023.3280741.

[9] Y L. Chang et al., "Ship detection based on YOLOv2 for SAR imagery," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 786, doi: 10.3390/rs11070786.

[10] T. Zhang et al., "High-speed ship detection in SAR images by improved YOLOv3," in *Proc. 16th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process.*, 2019, pp. 149–152, doi: 0.1109/ICCWAMTIP47768.2019.9067695.

[11] J. Jiang et al., "High-speed lightweight ship detection algorithm based on YOLO-V4 for three-channels RGB SAR image," *Remote Sens.*, vol. 13, no. 10, 2021, Art. no. 1909, doi: 10.3390/rs13101909..

[12] H. Guo et al., "A CenterNet model for ship detection in SAR images," *Pattern Recognit.*, vol. 112, 2021, Art. no. 107787, doi: 10.1016/j.patcog.2020.107787.

[13] Y. Jiang, W. Li, and L. Liu, "R-CenterNet : Anchor-free detector for ship detection in SAR images," *Sensors*, vol. 21, no. 17, 2021, Art. no. 5693, doi: 10.3390/s21175693.

[14] L. Bai, C. Yao, Z. Ye, D. Xue, X. Lin, and M. Hui, "Feature enhancement pyramid and shallow feature reconstruction network for SAR ship detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1042–1056, 2023, doi: 10.1109/JSTARS.2022.3230859.

[15] H. Qu, L. Shen, W. Guo, and J. Wang, "Ships detection in SAR images based on anchor-free model with mask guidance features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 666–675, 2022, doi: 10.1109/JSTARS.2021.3137390.

[16] X. Ma, S. Hou, Y. Wang, J. Wang, and H. Wang, "Multiscale and dense ship detection in SAR images based on key-point estimation and attention mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022, doi: 10.1109/TGRS.2022.3141407.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Patt. Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[18] 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.*, Honolulu, HI, USA, 2017, pp. 936–944. [Online]. Available: https://ieeexplore.ieee.org/document/8099589/

[19] S J. Woo et al., "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[20] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 764–773, doi: 10.1109/ICCV.2017.89.

[21] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Representations*, 2016, doi: 10.48550/arXiv.1511.07122.
[22] 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, doi: 10.1109/BIGSARDATA.2017.8124934.
[23] Y. Mao, X. Li, H. Su, Y. Zhou, and J. Li, "Ship detection for SAR imagery based on deep learning: A benchmark," in *Proc. IEEE 9th Joint Int. Inf. Technol. Artif. Intell. Conf.*, 2020, pp. 1934–1940.
[24] T Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Springer, 2014, pp. 740-755.
[25] L. Liu et al., "On the variance of the adaptive learning rate and beyond," in *Proc. Int. Conf. Learn. Representations*, 2021, doi: 10.48550/arXiv.1908.03265.
[26] Z X. Tian et al., "Fully convolutional one-stage 3D object detection on lidar range images," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 34899–34911, 2022.
[27] W D. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands,, Springer Int. Publishing, 2016, pp. 21–37.
[28] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, doi: 10.48550/arXiv.1804.02767.
[29] 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. Patt. Recognit.*, Vancouver, BC, Canada, 2023, pp. 7464–7475, doi: 10.1109/CVPR52729.2023.00721.
[30] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 2960–2969. [Online]. Available: https://ieeexplore.ieee.org/document/8953540/
[31] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*. Seoul, Korea (South), 2019, pp. 9656–9665. [Online]. Available: https://ieeexplore.ieee.org/document/9009032/
[32] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyond anchor-based object detector," *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020, doi: 10.1109/TIP.2020.3002345.
[33] X. Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2021, doi: 10.48550/arXiv.2010.04159.
[34] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *IEEE/CVF Proc. Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 548–558, doi: 10.1109/ICCV48922.2021.00061.
[35] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet++ for object detection," *IEEE Trans. Patt. Anal. Mach. Intell.*, 2022, doi: 10.1109/TPAMI.2023.3342120.
[36] S. Woo et al., "ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recognit.*, Vancouver, BC, Canada, 2023, pp. 16133–16142, doi: 10.1109/CVPR52729.2023.01548.

Yanshan Li received the M.Sc. degree in computer applied technology from the Zhejiang University of Technology, Hangzhou, China, in 2005, and the Ph.D. degree in traffic information engineering and control from the South China University of Technology, Guangzhou, China, in 2015.

He is an Associate Professor with the ATR National Key Laboratory of Defense Technology, Shenzhen University, Shenzhen, China. His research interests include computer vision, machine learning, and image analysis.

Wenjun Liu received the B.E. degree in electronic and information engineering from the Guangdong University of Technology, Guangzhou, China, in 2020, and the M.E degree in electronic information from Shenzhen University, Shenzhen, China, in 2023.

His research interests include image object detection and computer vision.

Ruo Qi received the M.E. degree in electronic information from the Harbin University of Science and Technology, Harbin, China, in 2023. She is currently working toward the D.Eng. degree in electronic information with Shenzhen University, Shenzhen, China.

Her research interests include remote sensing image processing and object detection.