A Mixed-Scale Self-Distillation Network for Accurate Ship Detection in SAR Images

Shuang Liu, Dong Li, *Senior Member, IEEE*, Renjie Jiang, Qinghua Liu, Jun Wan, *Member, IEEE*, Xiaopeng Yang, *Senior Member, IEEE*, and Hehao Liu

***Abstract*—Ship detection in synthetic aperture radar (SAR) images has attracted extensive attention due to its promising applications. While numerous methods for ship detection have been proposed, detecting ships in complex scenarios remains challenging. The main factors contributing to the lower detection accuracy are SAR image characteristics, such as blurred outlines, and similar scattering intensities between actual ship targets and background environment, induced by the special imaging mechanism. To alleviate these issues, we propose a mixed-scale self-distillation network (MSNet) for accurate ship detection in SAR images. First, the zoom strategy is used to obtain more ship target information, and differentiated information between ship targets and background environments at different scales is aggregated through the designed search module. Then, the consistency self-distillation module is proposed to match feature attention maps at different scales, which forces the model to capture the potential semantic attributes of ship targets through a self-distillation fashion. After that, the refinement module is developed to further enhance the discriminative semantics among different hierarchical features under mixed scales. Furthermore, to alleviate the uncertainty arising from indistinguishable background interference in SAR images, we introduce an uncertainty perception loss to facilitate the model to make accurate judgments in candidate regions. Extensive experiments are performed on the SAR ship detection dataset from the Gaofen-3, RadarSat-2, Sentinel-1, and TerraSAR satellites. The experimental results consistently demonstrate the superiority of our method over the existing state-of-the-art methods. Besides, detailed model analysis experiments further validate the effectiveness of our proposed method in SAR image ship detection tasks.**

***Index Terms*—Mixed-scale, synthetic aperture radar (SAR) ship detection, search and refinement network, self-distillation.**

I. Introduction

SYNTHETIC aperture radar (SAR) is a remote sensing technique that uses microwave signals reflected from objects to image objects [1]. SAR has become an important means for marine monitoring due to its ability to overcome disturbances, such as lighting conditions, weather, and occlusions, and provide high-quality ocean observation data all-day and all-weather [2]. Analysis of ship targets in SAR images can provide crucial maritime information, which is valuable for applications, such as maritime safety, resource management, and navigation regulation [3].

Numerous ship detection methods have been developed, broadly categorized into traditional and deep learning-based detection methods [4]. Traditional SAR ship detection algorithm mainly includes methods based on target scattering [5], polarization [6], [7], and geometric characteristics [8]. Among them, the constant false alarm rate (CFAR) [9] and its variants [10], [11], [12] are commonly used algorithms based on target scattering characteristics. These approaches focus on modeling the statistical distribution of background clutter and applying predefined thresholds to judge whether the target is present. However, accurately modeling background clutter, especially in inshore scenarios, is challenging. In addition, the method based on target polarization characteristics leverages the differences between ship target and background clutter under different polarization modes [13], [14]. Nonetheless, this approach depends on a prior database of polarimetric scattering characteristics and is sensitive to changes in background sea clutter. In contrast, the geometric-based method identifies the ship targets from the backgrounds in SAR images by analyzing the features, such as the shape, length, and contour [15], [16]. These geometric attributes directly reflect the physical characteristics of SAR ship targets [17], which enables reliable predictions. However, this method requires many training samples and prior template

Manuscript received 3 August 2023; revised 17 September 2023 and 8 October 2023; accepted 10 October 2023. Date of publication 13 October 2023; date of current version 30 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61971075, Grant 62371079, and Grant 62201099, in part by the Defense Industrial Technology Development Program under Grant JCKY2022110C171, in part by the Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education under Grant CRKL220202, in part by the Opening Project of the Guangxi Wireless Broadband Communication and Signal Processing Key Laboratory under Grant GXKL06200214 and Grant GXKL06200205, in part by the Sichuan Science and Technology Program under Grant 2022SZYZF02, in part by the Engineering Research Center of Mobile Communications, Ministry of Education under Grant cqupt-mct-202103, in part by the Fundamental Research Funds for the Central Universities Project under Grant 2023CDJXY-037, and in part by the Natural Science Foundation of Chongqing, China under Grant cstc2021jcyj-bshX0085. *(Corresponding author: Dong Li.)*

Shuang Liu, Dong Li, Renjie Jiang, Jun Wan, and Hehao Liu are with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China, and also with Chongqing Key Laboratory of Space Information Network and Intelligent Information Fusion, Chongqing University, Chongqing 400044, China (e-mail: shuangliumax@cqu.edu.cn; lidongcuit@126.com; 202112131089t@cqu.edu.cn; xidianwanjun@163.com; eryu523@163.com).

Qinghua Liu is with Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing, Guilin 541004, China (e-mail: qhliu@guet.edu.cn).

Xiaopeng Yang is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: xiaopengyang@bit.edu.cn).

The code will be available at: https://github.com/shuangliumax/MSNet/.

Digital Object Identifier 10.1109/JSTARS.2023.3324496

information, resulting in poor generalization capability for different ship target types and shapes. While traditional SAR ship detection algorithms may perform well in specific simple scenarios, due to the particularity of SAR images, their performance is limited by factors such as target complexity and background interference. Therefore, a more accurate and robust SAR images ship target detection method needs to be further developed.

Recently, deep learning-based object detection (OD) algorithms have demonstrated remarkable performance in the computer vision community [18]. These algorithms show high accuracy, scalability, and automatic feature learning ability when dealing with OD tasks in complex scenes. Therefore, SAR researchers began to introduce the deep learning-based OD method into the SAR field [19], [20], [21]. Currently, the most commonly used detection methods in SAR images are based on two-stage mechanism and region-based convolutional neural network (R-CNN) [22] frameworks. This approach divides the OD task into two stages: candidate box generation and object classification and bounding box regression. In the first stage, candidate object regions are generated using a region proposal network (RPN) or other region-generation algorithms. The second stage includes a classifier and a bounding box regressor for object classification and position adjustment on each candidate box. The classic two-stage detection methods include Faster R-CNN [23], feature pyramid networks (FPN) [24], Cascade R-CNN [18], Mask RCNN [25], etc [26]. Due to the advantages of Faster RCNN, such as simple structure, high detection accuracy, and strong versatility, most SAR ship detection algorithms are improved based on it [27], [28]. For instance, Zhao et al. [27] proposed a top–down fine-grained feature pyramid with the receptive field (RF) block and a convolutional attention module to capture the features of ships with large aspect ratios. Similarly, Li et al. [29] employed a skip connection structure to extract multiscale ship target features in SAR images. Yue et al. [30] gradually integrates the semantic strong features and low-level high-resolution features to mitigate false alarms. Shin et al. [31] combined candidate proposals from the raw SAR image and the synthetically denoised SAR image to reduce the impact of noise on ship detection. Su et al. [32] adopted focal loss to adjust the weights of hard negative and simple samples. Wang et al. [33] utilized a feature enhancement module based on a self-attention mechanism and performs an extended region-of-interest pooling operation on the potential proposals to improve model detection accuracy. Gong et al. [34] leverages the scale enhancement module, scale selection module, and context attention module to encourage the model to focus more on crucial regions within the image, thus improving the detection accuracy of small ships in complex environments. Moreover, since ship targets often exhibit irregular shapes and diverse orientations, some researchers have introduced the oriented bounding box (OBB)-based deep learning algorithm into the SAR field to detect ship targets with arbitrary orientations. For instance, Zhang et al. [35], an anchor-free and keypoint-based approach is presented for oriented ship detection in multiresolution SAR images. To overcome the boundary discontinuity problem in predicting bounding box angles and keypoint regression. Gao et al. [36] proposed ellipse encoding to effectively exploit the ship target's geometry and scattering characteristics, thus mitigating the negative impact of boundary discontinuity. Furthermore, to deal with the challenge of strong scattering interference in inland areas. Sun et al. [37] proposed to detect the strong scattering points on ships and then combine their positions to obtain an arbitrary orientation bounding box for the ship target. Guo et al. [38] proposed a new encoding method for describing OBB and incorporates a feature adaptive module to learn the shape and direction information of arbitrarily oriented ships, which effectively alleviates the challenge of detecting ships with arbitrary orientations in SAR images. Beyond these methods, some one-stage detection algorithms have also been developed, such as [39], which proposes a strong scattering point aware network to identify and detect ships by capturing the strong scattering points in the ship area.

In general, the current SAR ship detection approaches based on deep learning have shown promising results in various scenarios. However, detecting ship targets with complex backgrounds from SAR images remains a challenging task due to the SAR image characteristics, e.g., blurred outlines, and the similarity in scattering intensity between actual ship targets and the background. These challenges result in three main algorithmic issues. The first issue is feature extraction. In the SAR ship detection task, we are concerned with how to accurately identify a specific class of targets, ships, from SAR images while distinguishing them from the surrounding background environment. Nevertheless, actual ship targets often exhibit a similar appearance and texture to the background environment, making the extracted features indistinguishable. The second is semantic information understanding, as actual ship targets may be hidden within complex backgrounds, making it difficult to explore the discriminative and subtle semantic clues to identifying real ship targets. The third issue pertains to uncertainty, where the image may have some regions of uncertainty (low confidence) or ambiguity due to the concealment of the actual target. These issues make detecting ship targets with complex backgrounds from SAR images challenging.

Considering these, we summarize the SAR ship detection issue in complex scenarios into two points: 1) how to effectively capture the differentiated information between the ship target and the background environment, and accurately locate the real ship target in a confused environment? The differentiated information here refers to the differences between the ship target and its surrounding environment in terms of visual features, textures, shapes, etc., which can help the model to better distinguish the ship target from the background within a chaotic environment, thus facilitating accurate ship target detection and 2) how to suppress obvious background interference and identify ship targets more reliably. Taking inspiration from human behavior, we can observe the differences between the object and the background by zooming in or out of an image, to identify blurred or hidden objects. With this, we propose a novel approach that simulates the zooming in and out strategy to capture the subtle differences between the actual ship target and the background environment.

Based on this inspiration, we propose a MSNet as a novel two-stage detection approach for accurate ship detection in SAR images with complex scenarios, which significantly improves ship detection performance. First, to achieve accurate ship object

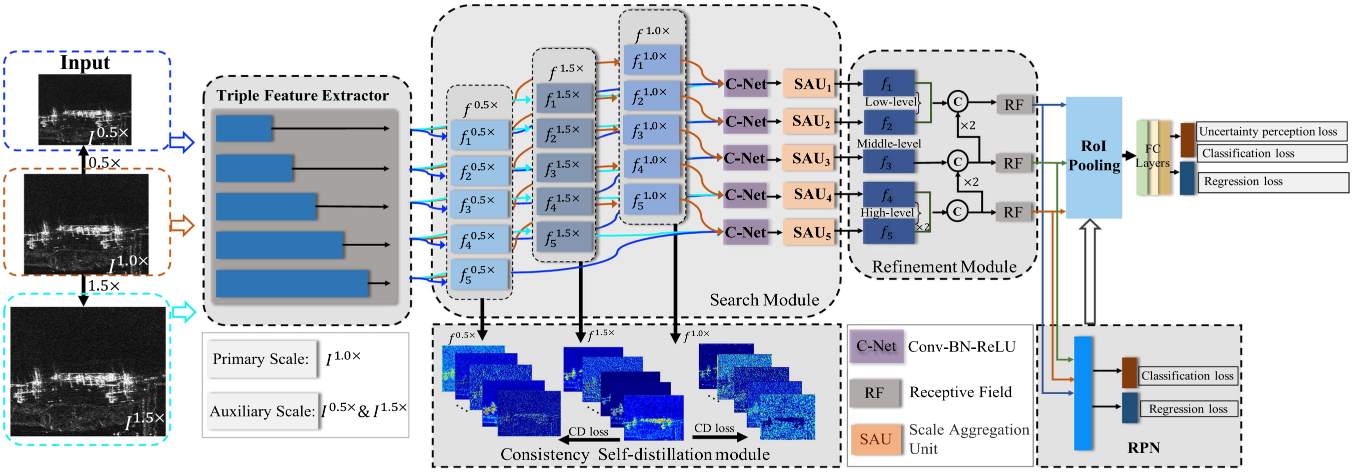

Input
$I^{0.5\times}$
0.5×
$I^{1.0\times}$
1.5×
$I^{1.5\times}$
Triple Feature Extractor
Primary Scale: $I^{1.0\times}$
Auxiliary Scale: $I^{0.5\times}$ & $I^{1.5\times}$
$f^{0.5\times}$
$f_1^{0.5\times}$ $f_2^{0.5\times}$ $f_3^{0.5\times}$ $f_4^{0.5\times}$ $f_5^{0.5\times}$
$f^{1.5\times}$
$f_1^{1.5\times}$ $f_2^{1.5\times}$ $f_3^{1.5\times}$ $f_4^{1.5\times}$ $f_5^{1.5\times}$
$f^{1.0\times}$
$f_1^{1.0\times}$ $f_2^{1.0\times}$ $f_3^{1.0\times}$ $f_4^{1.0\times}$ $f_5^{1.0\times}$
C-Net
SAU_1 SAU_2 SAU_3 SAU_4 SAU_5
Search Module
f_1 f_2 f_3 f_4 f_5
Low-level
Middle-level
High-level
C
×2
RF
Refinement Module
RoI Pooling
FC Layers
Uncertainty perception loss
Classification loss
Regression loss
$f^{0.5\times}$ $f^{1.5\times}$ $f^{1.0\times}$
CD loss
Consistency Self-distillation module
C-Net Conv-BN-ReLU
RF Receptive Field
SAU Scale Aggregation Unit
Classification loss
Regression loss
RPN

Fig. 1. Overall framework of the proposed MSNet. The triple feature extractor with shared weights extracts multilevel features for three input scales. The SM integrates critical cues among the same level at different scales. The CSM matches feature attention maps of different scales with a self-distillation mechanism. The RM integrates the different hierarchical features under mixed scales. The UPL is used to facilitate the model to make more accurate judgments in uncertain or ambiguous regions.

localization, we adopt a zoom-in and out strategy, and integrate the differentiated information between the ship target and the background environment at different scales through the designed search module (SM). In the following, we develop a consistency self-distillation module (CSM) to match feature attention maps at different scales, which forces the model to capture the semantic attributes of ship targets through a self-distillation fashion. This facilitates extracting and transmitting crucial semantic cues of ship targets from images at different scales, thus prompting accurate predictions. Second, for reliable predictions, we design a refinement module (RM) to enhance the discriminative representations among different hierarchical features under mixed scales. Moreover, considering that indistinguishable background environments in SAR images may negatively affect the model learning, we introduce an uncertainty perception loss (UPL) to encourage the model to make more accurate judgments in candidate regions. We conduct extensive experiments on multiple SAR ship detection datasets (SSDDs), including Gaofen-3, RadarSat-2, Sentinel-1, and TerraSAR satellites. Experimental results consistently show that compared with other OD algorithms, the proposed method has obvious advantages in quantitative and qualitative performance. Furthermore, detailed model analysis experiments further validate the effectiveness of our proposed method in SAR ship detection tasks.

In summary, our contributions are as follows.

1) A MSNet for SAR ship detection is proposed in challenging scenarios. MSNet focuses on mining semantic clues of ship targets at different scales and combines purposeful optimization strategies to reliably detect ship targets in complex scenarios.
2) A SM and a CSM are proposed. With the combined effect of the two, the model can comprehensively search for crucial clues about ship targets from different scale images within chaotic scenarios. These modules also facilitate the aggregation of more discriminative feature representations, thus improving the model's detection accuracy.
3) Extensive experiments are conducted on the SSDD from the Gaofen-3, RadarSat-2, Sentinel-1, and TerraSAR satellites. The experimental results consistently show that our method achieves good detection performance in both quantitative and qualitative evaluation.

The rest of the article is organized as follows. The proposed MSNet is elaborated in Section II. The experimental results and discussions are presented in Section III. Finally, Section IV concludes this article.

II. Proposed Method

In this section, we start by introducing the overall architecture of the proposed MSNet. Then, we provide a detailed explanation of each module and the loss function of the proposed method.

A. Overall Architecture

The overall framework of the proposed MSNet is shown in Fig. 1. As mentioned above, due to the SAR image characteristics, blurred outlines, and similar scattering intensities between actual ship targets and background surrounding, result in ship targets that are difficult to identify accurately. Inspired by the idea from [40] about human beings adopting a zoom strategy when observing confused or complex scenarios, we realize that varying scaling factors often preserve their specific information. Therefore, detecting ships with complex backgrounds from SAR images, aggregating the differentiated information between ship targets and background interferences at different scales can facilitate the model to capture valuable ship target clues from complex environments, thus promoting accurate ship detection. To achieve this, we take the original input image as the primary scale, and two auxiliary scales are obtained through zoom-in and zoom-out operations. We use a triple feature extractor for

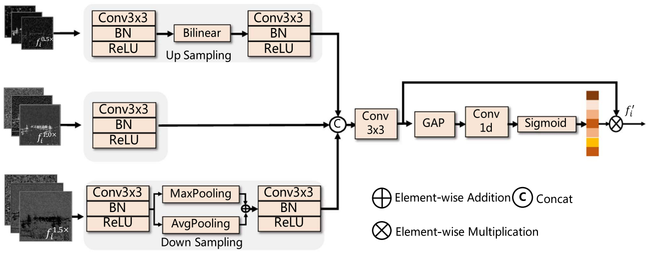


Fig. 2. Illustration of the SAU.

multilevel feature extraction on images at three scales. This triple feature extractor is three identical feature extractors, and the weights are shared. The multilevel features corresponding to these three scale images are then sent to a SM designed to integrate scale-specific information. After that, we build a CSM to match feature attention maps across different scales. This largely forces the model to extract the crucial ship semantic cues through a self-distillation mechanism. Moreover, we design a RM to gradually integrate feature representations at different hierarchies under the mixed scale. This further increases the semantic representations of diversity. Finally, to overcome the uncertainty induced by the inherent complexity of SAR images, we introduce an UPL. This allows the model to produce more reliable predictions for these regions of uncertainty or ambiguity.

B. SM

The SM is designed to integrate and enhance the differentiated information between ship targets and background surroundings. Specifically, input images of different scales are fed into a triple feature extractor. This triple feature extractor represents three identical feature extractors with shared weights. These are used for multilevel feature extraction corresponding to three input images of different scales. The feature extractor here consists of a commonly used ResNet-50 [41], which extracts features from five layers: $\{\text{conv}, \text{layer1}, \text{layer2}, \text{layer3}, \text{layer4}\}$. It is worth noting that the primary and two auxiliary scales are set to $1.0\times$, $1.5\times$, and $0.5\times$, respectively, to better balance effectiveness and efficiency. After passing through the feature extractor, the features of different levels corresponding to each scale are output as $\{f_i^k\}_{i=1}^5, k \in \{0.5, 1.0, 1.5\}$. These features are then passed through a cascaded channel compression network (C-Net) to obtain more compact feature representations. C-Net consists of an independent "Conv3×3-BN-ReLU" unit. Subsequently, the features at different levels are fed into the designed scale aggregate unit (SAU) to combine those scale-specific information, as shown in Fig. 2.

Before scale aggregation, the features $f_i^{0.5}$ and $f_i^{1.5}$ are first resized to have the same resolution as the primary scale feature $f_i^{1.0}$. For $f_i^{1.5}$, we down-sample it using a hybrid structure of average-pooling and max-pooling to preserve discriminative semantic cues related to ship targets. For $f_i^{0.5}$, we directly up-sample it using bilinear interpolation. We combine these features at the same layer by using a concatenation operation, i.e., $f_i = \text{Concat}(f_i^{0.5}, f_i^{1.0}, f_i^{1.5})$, and adjust the channel dimension using a 3×3 convolution. The features are then aggregated using global average pooling (GAP), and the corresponding channel attention (weight) is obtained through 1-D convolutions followed by a sigmoid function. Finally, the channel attention is multiplied with the input features to obtain the final output f_i'. The procedure is expressed as follows:

$$f_i' = \sigma(F_{1D}^k(GAP(f_i))) \otimes f_i \tag{1}$$

where F_{1D}^k is 1-D convolution, σ denotes sigmoid function, $\otimes$ denotes element-wise multiplication operation. Thus, through the SM, the scale-specific features can be fully integrated.

C. CSM

After passing through the triple feature extractor, we design a CSM to facilitate the model to further explore the semantic clues of the ship target. The motivation behind this module is that features at different scales often contain scale-specific information. Intuitively, we believe that large-scale image features contain richer structural information, which can provide more obvious differentiation between the ship target and the complex background. Leveraging large-scale features to assist small-scale feature learning can help the model identify inconspicuous but valuable semantic clues in indistinguishable complex scenarios. Inspired by the self-knowledge distillation mechanism [42], [43], which employs the model itself as a teacher and student model to distill its own knowledge by comparing and matching features at different levels, we take the large-scale feature as the "teacher" and the small-scale features as "students." By comparing and matching the attention maps of the same hierarchical features corresponding to different scales, critical semantic information can be effectively shared and transmitted. Specifically, let $F_i^{\text{scale}} \in R^{C\times H\times W}$ represent an output feature with a specific scale, consisting of C feature channel with spatial dimensions $H \times W$. The scales are 1.5, 1.0, and 0.5, respectively. We first calculate the absolute value of the activation value of each feature map in the channel dimension. Then, we weight each absolute value using a power operation and sum each feature map element-wise in the channel dimension to generate a spatial attention map, which can be formalized as follows:

$$A_i^{1.5} = \sum_{j=1}^{C} \left| \left(F_i^{1.5}\right)_j \right|^p$$

$$A_i^{0.5} = \sum_{j=1}^{C} \left| \left(F_i^{0.5}\right)_j \right|^p$$

$$A_i^{1.0} = \sum_{j=1}^{C} \left| \left(F_i^{1.0}\right)_j \right|^p \tag{2}$$

where $(F_i^{1.5})_j$, $(F_i^{0.5})_j$, and $(F_i^{1.0})_j$ represent the jth featrure map of the ith layer with scales $1.5\times$, $0.5\times$, and $1.0\times$, respectively. $A_i^{1.5}$, $A_i^{0.5}$, and $A_i^{1.0}$ represent the obtained spatial attention maps with scale $1.5\times$, $0.5\times$, and $1.0\times$, respectively. The parameter p represents a power operation on the absolute value, which assigns more weight to the spatial location corresponding to the region with the highest activation value (i.e., assigns more weight to the more discriminative part). In other words, the larger

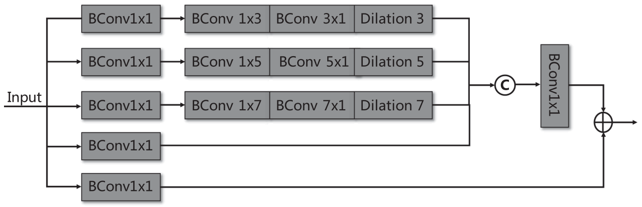


Fig. 3. Illustration of the RF) component.

the value of p, the more the model focuses on those parts with the highest activation values, that is, the more prominent the regions that are important for distinguishing the target. This mechanism amplifies the crucial clues of ship targets at different scales, and enhances the differentiated information between actual ship targets and background surroundings. Using $A_i^{1.5}$, which contains richer structural information to guide the learning of $A_i^{0.5}$ and $A_i^{1.0}$, encourages the model to excavate similar semantic clues with different scales. This can be implemented as part of the loss function, referred to as the consistency self-distillation loss ($\mathcal{L}_{cd}$), which can be formalized as follows:

$$\mathcal{L}_{cd}^1 = \text{RMSE}\left(\frac{A_i^{1.5}}{\|A_i^{1.5}\|_2}, \frac{A_i^{0.5}}{\|A_i^{0.5}\|_2}\right)$$
$$\mathcal{L}_{cd}^2 = \text{RMSE}\left(\frac{A_i^{1.5}}{\|A_i^{1.5}\|_2}, \frac{A_i^{1.0}}{\|A_i^{1.0}\|_2}\right) \quad (3)$$

where $\|\cdot\|_2$ represents L_2 normalization. RMSE$(\cdot)$ represents the root mean square error function, which is employed to measure the difference of spatial attention maps at different scales. By optimizing the minimization of $\mathcal{L}_{cd}^1$ and $\mathcal{L}_{cd}^2$, ship semantic clues of different scales are fully explored and transmitted. The differentiated information between ship targets and confused backgrounds is effectively captured, thus improving the model detection performance in complex backgrounds.

D. Refinement Module

After SAU, scale-specific information is aggregated, and five different mixed-scale hierarchical features are obtained. It is known that the inherent property of neural networks, namely, low-level features in shallow layers, preserve spatial details for constructing object boundaries, while high-level features in deep layers retain semantic information for locating targets [44]. For this purpose, we design a RM to learn discriminative semantics among different hierarchical to achieve more accurate predictions. Specifically, we first divide the extracted features into low-level features $\{f_1, f_2\}$, middle-level features $\{f_3\}$, and high-level features $\{f_4, f_5\}$. We then fuse the high-level and low-level features separately using the concatenation operation Concat$(\cdot)$ and up-sampling operations $Up_{2\times}(\cdot)$. That is, $f_h = \text{Concat}(f_4, Up_{2\times}(f_5))$ and $f_l = \text{Concat}(f_1, f_2)$. After that, we utilize the f_h to guide middle-level features, i.e., $f_{hm} = \text{Concat}(Up_{2\times}(f_h), f_3)$. Similarly, the f_{hm} is used to guide the low-level features f_l, i.e., $f_{ml} = \text{Concat}(Up_{2\times}(f_{hm}), f_l)$. Further, we adopt the modified RF component [45], as shown in Fig. 3, to integrate more discriminative feature representations by enlarging RFs. The procedure is expressed as follows:

$$f_h' = \text{RF}(f_h), f_{hm}' = \text{RF}(f_{hm}), f_{ml}' = \text{RF}(f_{ml}) \quad (4)$$

where RF$(\cdot)$ is RF component. Thus, f_h', f_{hm}', and f_{ml}' as the final output prediction layers are fed into the RPN and RoI Pooling layer to generate candidate region proposals for potential targets and extract the regions of interest features, respectively.

RF: The RF component is used to enlarge the RFs of the feature map, which can facilitate the model to capture more contextual information from the input data. It consists of five branches, denoted as $\{r_b, b = 1, \ldots, 5\}$, each with a different convolutional (Bconv) layer and Dilation rate. Within each branch, the first Bconv layer to adjust the channel dimensions with 1×1 operation, followed by a $1 \times (2b + 1)$ Bconv layer and a $(2b + 1) \times 1$ Bconv layer with a specific dilation rate $(2b + 1)$ when $b < 4$. The outputs of these four branches are concatenated, and a 1×1 Bconv operation is used to adjust their channel size. Finally, the fifth branch is added, and the entire module outputs the final features.

E. Loss Functions

In this article, the loss function of the proposed MSNet consists of multiple components, including the RPN loss, OD loss, consistency self-distillation loss, and UPL. Among them, the RPN loss comprises two parts: the anchor box classification loss and the anchor box bounding regression loss. The anchor box classification loss is used to distinguish whether a candidate region contains the ship target or belongs to the background, while the anchor box bounding regression loss is utilized to predict the position shift between the candidate region and the real target. The RPN loss can be formalized as follows:

$$L_{\text{rpn}}(\{p_i\}, \{t_i\}) = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(p_i, p_i^*) + \frac{1}{N_{\text{reg}}} \sum_i p_i^* L_{\text{reg}}(t_i, t_i^*) \quad (5)$$

where p_i^* and p_i are the ground-truth label and predicted probability of the ith anchor in a mini-batch being an object, respectively. p_i^* is 1 when the anchor is positive, otherwise p_i^* is 0. t_i is the parameterized coordinates vector of the predicted bounding box, and t_i* is the vector representing the ground-truth box corresponding to the positive anchor. By minimizing the loss sum of L_{cls} and L_{reg}, the model can produce high-quality region proposals.

This is followed by OD loss, which consists of the classification loss and the bounding box regression loss. The classification loss determines which category the candidate region belongs to, and the bounding box regression loss predict the precise location of the target, which can be formalized as follows:

$$L_{\text{cls}} = -\frac{1}{N_{\text{cls}}} \sum_{i=1}^{N_{\text{cls}}} \sum_{c=1}^{C} y_i^c \log(p_i^c) + (1 - y_i^c) \log(1 - p_i^c)$$

$$L_{\text{reg}} = \frac{1}{N_{\text{reg}}} \sum_i^{N_{\text{reg}}} \text{SmoothL1}(t_i - t_i^*)$$

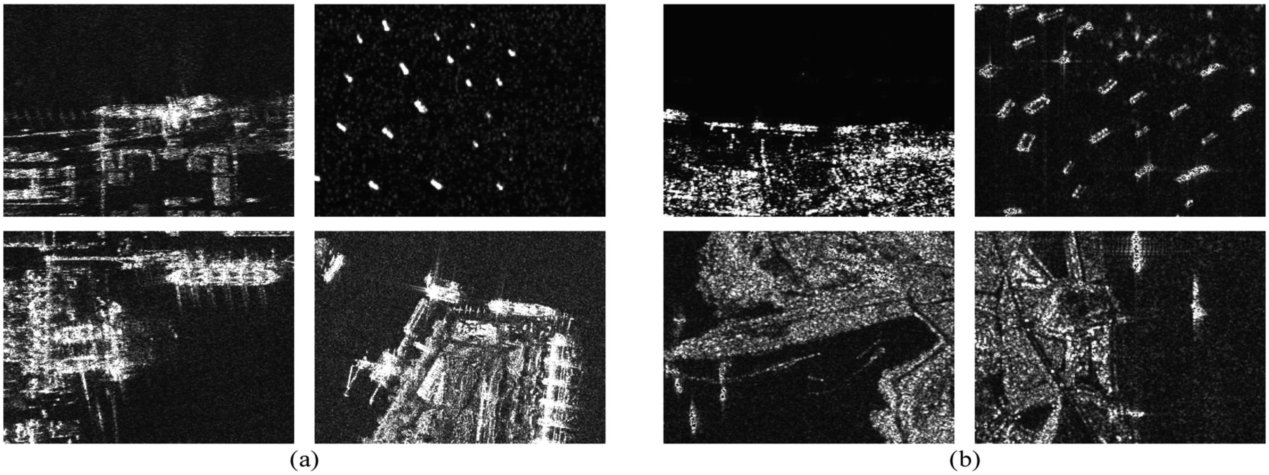


Fig. 4. Sample images in the two experiment datasets. (a) SSDD. (b) Gaofen-3.

TABLE I
INFORMATION OF GAOFEN-3 AND SSDD DATASETS

Dataset	Gaofen-3	SSDD
Satellite	Gaofen-3	RadarSat-2,TerraSAR-X, Sentinel-1
Polarization	HH,HV,VV,VH	HH,HV,VV,VH
Resolution	1m-3m	1m-15m
scenes	inshore and offshore	inshore and offshore
Image Size	256×256	190-668
image number	2832	1160
ship number	4386	2456

$$L_{od} = L_{\text{cls}} + L_{\text{reg}} \tag{6}$$

where L_{cls} is the cross-entropy loss, and L_{reg} is the SmoothL1 funciton. N_{cls} and N_{reg} represent the total number of samples, respectively. C is the number of categories, y_i^c is a binary indicator variable indicating whether the i sample belongs to the category c, and p_i^c is the model predicts probability that the ith sample belongs to the class c. t_i is the bounding box parameter predicted by the model, and t_i^* is the true bounding box parameter. The final OD loss function L_{od} is the sum of L_{cls} and L_{reg}.

In addition, considering that the uncertainty stems from the SAR image characteristic, that is, blurred outlines and similar scattering intensities between actual targets and background targets. Inspired by [40], we introduce an UPL to encourage the model to produce accurate predictions in those uncertain or ambiguous regions. Specifically, for the final predicted probability score of the ship target, the probability score range from $[0, 1]$, where 0 indicates that the feature belongs to the background, and 1 indicates it belongs to the ship target. The closer the prediction score is to 0.5, the more uncertain the attribute of the feature is. That is, it is difficult for the model to distinguish between the real ship target and the background. To optimize it, the ambiguity is used as the supplementary loss for these hard samples. The ambiguity measure of prediction score x is defined, which maximizes at $x = 0.5$ and minimizes at $x = 0$ or $x = 1$. Thus, the UPL can be formulated as $L_{\text{upl}}^{i,j} = 1 - |2\text{p}_{i,j} - 1|^2$.

Finally, the overall loss function is formulated as follows:

$$L_{\text{overall}} = L_{\text{rpn}} + L_{\text{od}} + L_{\text{cd1}} + L_{\text{cd2}} + L_{\text{upl}}. \tag{7}$$

By optimizing L_{overall}, our model can achieve accurate and robust ship detection in complex SAR scenarios.

III. EXPERIMENTS

In this section, we present the evaluation of our proposed MSNet. The dataset, evaluation criteria, implementation details are described in Section III-A. Comparisons with other state-of-the-art detection methods are presented in Section III-B. Model analysis and discussion is presented in Section III-C. More details are described as follows.

A. Experiments Setup

Datasets: In this article, the SSDD [46] and Gaofen-3 dataset are utilized to assess the performance of the proposed MSNet. The details of these two datasets are shown in Table I, and several sample images for each dataset are illustrated in Fig. 4. These datasets exhibit rich diversity, including complex backgrounds, dense distribution, small-size ships, and arrangement near the wharf. *SSDD*: The SSDD dataset images are primarily from RadarSat-2, TerraSAR-X, and Sentinel-1 sensors. The image resolution ranges from 1 to 15 m. The SSDD contains 1160 images of different image sizes with four polarization modes of HH, HV, VV, and VH, such as 501 × 349 and 500 × 403, as shown in Fig. 4(a). The training set consists of 928 images (2041 ships), while the test set comprises 232 images (546 ships). *Gaofen-3*: The Gaofen-3 dataset images are from the Gaofen-3 sensor, with four polarization modes of HH, HV, VV, and VH. The dataset is collected and produced by [47]. These images are cropped from large-scale scene images with a resolution between 1 and 3 m for SAR image ship detection tasks, as shown in Fig. 4(b). A total of 2832 images with the image size 256 × 256 is used in this article to evaluate the model. Among these, 2346 images (3614 ships) are employed as the training set, and 486 images (772 ships) are employed as the test set.

Evaluation Criteria: This article uses Precision, Recall, F1-Score, Average Precision (AP), and Inference time to assess the performance of the detection model. Positive examples refer to ship targets, while negative examples are nontargets (background). The Precision measures the detection accuracy of the model, indicating the proportion of correctly detected ships among all detected ships. It is defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

where TP represents true positive, i.e., the number of positive examples correctly predicted as ship targets, and FP represents false positives, i.e., the number of negative examples incorrectly predicted as positive examples. The Recall measures the ability of the model to identify ship targets, which indicates the proportion of correctly predicted ships relative to all ground truth ships

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

where FN represents false negative, i.e., the number of samples that are positive examples but are misclassified as negative examples. The AP measures the average of Precision and Recall at different intersection over Union (IoU) thresholds to evaluate the overall performance of the model on the ship class. It is defined as follows:

$$\text{AP} = \int_0^1 \text{P}(r)dr \tag{10}$$

where $\text{P}(r)$ represents the Precision at recall r. The F1-Score integrates Precision and Recall to provide a comprehensive evaluation of the model's performance

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

Precision, Recall, F1-Score, and AP values range from $[0, 1]$. Higher Precision, Recall, and AP values mean that the model has a stronger ability to predict and identify ship targets. The larger F1-Score value indicates that the model performs better on both Precision and Recall simultaneously and vice versa.

Inference time refers to the time required for the model to perform inference operations on each image, typically measured in milliseconds (ms), to assess the real-time performance of a model. Less inference time indicates that the model is more efficient on the given hardware. With these metrics, we can comprehensively evaluate the performance of the SAR ship detection model.

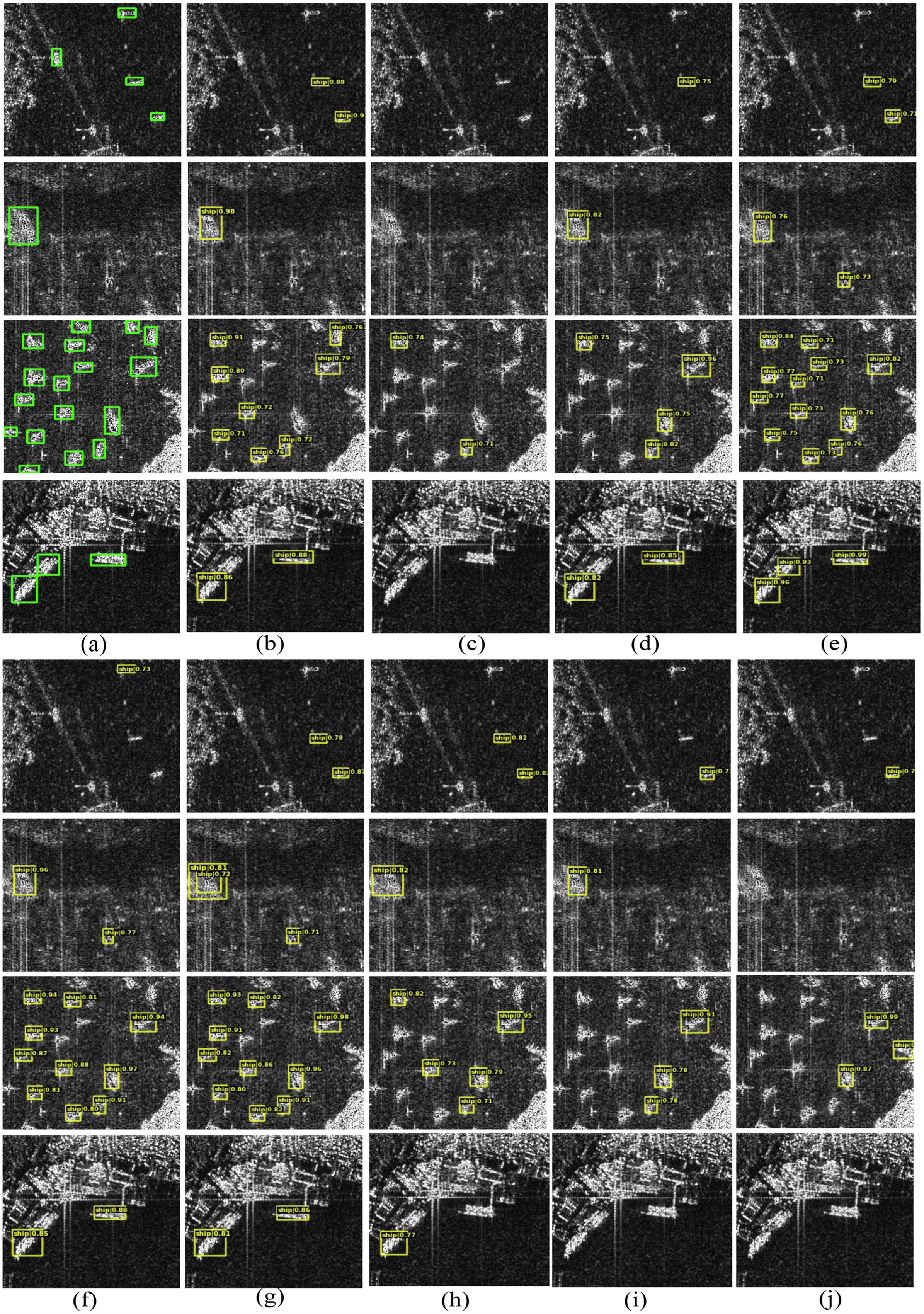
Implementation Details: The proposed MSNet is implemented with PyTorch [48]. The feature extractor is initialized with the parameters of ResNet-50 [41] pretrained on ImageNet [49]. The remaining part is randomly initialized. We utilize the SGD with momentum 0.9 and weight decay 0.0005 for optimization, and the learning rate is set to 0.005. The entire model is trained for 30 epochs with a batch size of 4 in an end-to-end manner on an NVIDIA GeForce RTX 4090 GPU. During training and inference, the primary scale is fixed 512 × 512, and the hyperparameter p is fixed to 4 in (2) for all experiments.

B. *Comparisons With State-of-the-Arts*

To demonstrate the superiority of the proposed MSNet, we compare MSNet with eight state-of-the-art OD methods, including Baseline [24], NAS-FCOS [50], RetinaNet [51], VarifocalNet [52], SABL [53], GA-RPN [54], ARPN-SAR* [27], and DAPN-SAR* [55]. The Baseline is Faster RCNN with FPN [24], which is based on a two-stage detection framework, Faster RCNN, and adopts a FPN for OD. Notably, due to the lack of publicly available source code, we reproduced the ARPN-SAR and DAPN-SAR methods and marked them with asterisks. We conduct experiments on both the Gaofen-3 and SSDD datasets. For a fair comparison, all methods use Resnet50 as the backbone network. We quantitatively evaluate the performance of the detection model using Precision, Recall, F1-Score, AP, and Inference time metrics and also provide qualitative evaluation by visualizing the detection results of different algorithms. The following section presents detailed experimental results.

Quantitative Evaluation: Tables II and III report the detection results of various algorithms on the Gaofen-3 and SSDD datasets, respectively. On the Gaofen-3 dataset, it can be seen that the proposed MSNet achieves the best results in Precision (76.55%), F1-Score (83.50%), and AP (61.86%), showing the best detection accuracy. RetinaNet performs the best result on the Recall metric (94.81%). This is due to its use of a dense anchor box distribution. These denser anchor boxes can facilitate the model to capture targets more comprehensively. However, MSNet performs better overall. Notably, compared with deep learning-based SAR ship detection algorithms, ARPN-SAR*, and DAPN-SAR*, MSNet achieves significant advantages in Precision, F1-Score, Recall, and AP metrics. Furthermore, it can be seen that MSNet requires more inference time on each image compared to other methods. This is because the proposed MSNet generates the final prediction result by simultaneously capturing the ship target semantic clues on multiple scale images, which leads to more computing time. On the SSDD dataset, MSNet shows the best performance in Precision (91.24%), Recall (96.33%), F1-Score (93.72%), and AP (75.93%), which further proves the superiority of MSNet in SAR target detection accuracy. This emphasizes that aggregating differentiated information between ship targets and background environments at different scales is helpful for the model to capture crucial semantic information about ship targets, thereby achieving better target recognition ability. Overall, MSNet shows better performance on the Gaofen-3 and SSDD datasets compared to other detection algorithms, although with a slightly increased inference time requirement.

Qualitative Evaluation: Figs. 5 and 6 report the visualization results of different algorithms on the Gaofen-3 and SSDD datasets, respectively. In Fig. 5(a), the green rectangles represent the ground truths of SAR ship targets. Fig. 5(b)–(h) are the ship targets detected by the proposed MSNet, Baseline, NAS-FCOS, RetinaNet, VarifocalNet, SABL, GA-RPN, ARPN-SAR*, and DAPN-SAR* methods on the Gaofen-3 dataset. These detected targets are marked with yellow rectangles containing predicted target categories and confidence scores. In Fig. 5, it can be observed that in inshore scenarios and scenes with confused



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 5. Detection results of different methods on Gaofen-3 dataset. (a) Ground truth. (b) MSNet (Ours). (c) Baseline. (d) NAS-FCOS. (e) RetinaNet. (f) VarifocalNet. (g) SABL. (h) GA-RPN. (i) ARPN-SAR*. (j) DAPN-SAR*. The image is best viewed by zooming in on the electronic version.

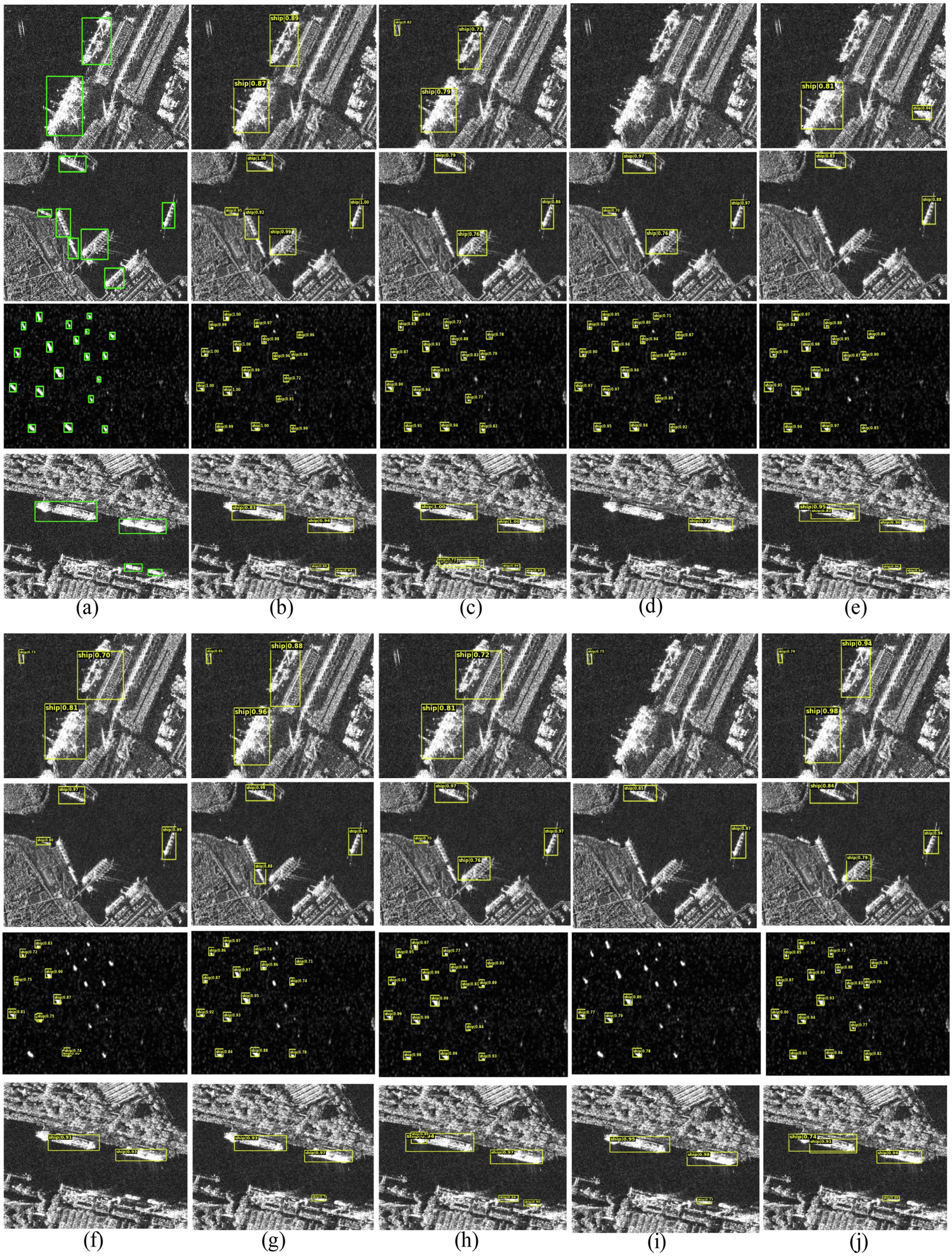


Fig. 6. Detection results of different methods on SSDD dataset. (a) Ground truth. (b) MSNet (Ours). (c) Baseline. (d) NAS-FCOS. (e) RetinaNet. (f) VarifocalNet. (g) SABL. (h) GA-RPN. (i) ARPN-SAR*. (j) DAPN-SAR*. The image is best viewed by zooming in on the electronic version.

TABLE II
COMPARISON OF DETECTION RESULTS FOR DIFFERENT ALGORITHMS ON GAOFEN-3 DATASET

Method	**Precision**(%)	**Recall**(%)	**F1-Score**(%)	**AP**(%)	**Times** (ms/img)
Baseline [24]	62.98	80.70	70.75	53.08	18
NAS-FCOS [50]	74.27	88.75	80.87	60.59	34
RetinaNet [51]	73.40	**94.81**	82.74	59.86	**12**
VarifocalNet [52]	74.61	93.74	83.09	60.24	13
SABL [53]	70.41	94.62	80.74	58.60	21
GA-RPN [54]	73.49	88.32	80.23	60.25	14
ARPN-SAR* [27]	66.23	86.31	74.95	56.22	26
DAPN-SAR* [55]	69.41	83.03	75.61	56.14	24
MSNet(Ours)	**76.55**	91.84	**83.50**	**61.86**	43

TABLE III
COMPARISON OF DETECTION RESULTS FOR DIFFERENT ALGORITHMS ON SSDD DATASET

Method	**Precision**(%)	**Recall**(%)	**F1-Score**(%)	**AP**(%)	**Times** (ms/img)
Baseline [24]	87.80	96.20	91.81	74.06	25
NAS-FCOS [50]	89.93	93.47	91.67	74.35	41
RetinaNet [51]	88.13	94.31	91.12	75.10	**18**
VarifocalNet [52]	89.51	92.73	91.09	73.40	19
SABL [53]	90.13	92.76	91.43	74.92	27
GA-RPN [54]	90.37	94.90	92.58	73.52	21
ARPN-SAR* [27]	88.45	92.12	90.25	72.15	33
DAPN-SAR* [55]	91.21	94.69	92.92	74.89	30
MSNet(Ours)	**91.24**	**96.33**	**93.72**	**75.93**	49

backgrounds (the first and second row), although the SAR ship targets and the background interference have a high degree of similarity in appearance, the MSNet can still locate the ship target relatively accurately with higher confidence than other methods. While other methods have obvious false detection problems, such as Fig 5(e)–(g). In a complex scenario with multiple targets (third row), RetinaNet shows good prediction performance, thanks to its dense anchor box distribution strategy, which allows it to detect targets more comprehensively. On the SSDD dataset, as shown in Fig. 6, in the inshore area near the wharf, compared with other methods, MSNet can more accurately detect the target with higher confidence, while other methods have an obvious false detection problem, as shown in the first row. Combining the results from Figs. 5 and 6, it is evident that the proposed MSNet shows superior SAR ship detection performance in complex backgrounds.

C. Ablation Studies

In this section, we conduct comprehensive ablation analyses on different components of the proposed MSNet. Since the Gaofen-3 dataset has a rich diversity, including complex background, dense distribution, and arrangement near the wharf, all subsequent ablation experiments are performed on it.

Effectiveness of SM and CSM: In the proposed method, the SM is used to integrate differentiated information between ship targets and background environments at different scales. The CSM is used to encourage the model to capture crucial semantic cues by matching and transmitting salient information at different scales. As these two modules are interactive, we simultaneously remove them from the full model to evaluate their effectiveness. The quantitative results are shown in Table IV. It can be seen that removing SM and CSM, i.e., Ours w/o SM & CSM, results in a significant decrease in Precision, Recall, and F1-Score, dropping by 13.22%, 12.44%, and 13.04%, respectively. Moreover, it can be seen from Fig. 7(b), Ours w/o SM & CSM, results in the model cannot accurately distinguish between the actual ship target and the background interferer, i.e., the features in the nontarget area are also activated with a larger value. These results highlight the effectiveness of the SM and CSM components in helping the model mine and refine valuable semantic information of ship targets, thereby improving detection ability in complex backgrounds.

Effectiveness of CSM: After passing the images of different scales through the feature extractor, we design a CSM to match spatial attention maps of different scales through a self-distillation fashion. We remove the CSM from the full model, i.e., Ours w/o CSM, to evaluate its effectiveness. In Table IV, it can be seen that Ours w/o CSM led to a significant decrease in the Precision, Recall, and F1-Score, dropping by 7.74%, 8.25%, and 8.02%, respectively. Notably, the Ours w/o CSM outperforms Precision, Recall, and F1-Score, by 5.48%, 4.19%, 5.02%, respectively, compared to the Ours w/o SM & CSM. This aspect provides evidence for the effectiveness of the SM module. In addition, Fig. 7(c) illustrates that Ours w/o CSM causes the model to disregard regions where the actual ship target exists, which makes the model prone to making wrong decisions. Overall, these results show that the CSM module can effectively extract the semantic properties of ships by matching the saliency information of images at different scales.

Effectiveness of RM: In the proposed model, a RM is used to enhance the discriminative semantics between different levels

TABLE IV
ABLATION STUDIES ON EFFECTIVENESS OF DIFFERENT COMPONENTS ON GAOFEN-3 DATASET

Method	SM	RM	CSM	UPL	Precision(%)	Recall(%)	F1-Score(%)	Param.(M)	Times (ms/img)
Baseline	✘	✘	✘	✘	62.98	80.70	70.75	41.30	18
Ours w/o SM & CSM	✘	✔	✘	✔	63.33	79.40	70.46	41.04	23
Ours w/o RM	✔	✘	✔	✔	70.70	91.06	79.60	37.45	30
Ours w/o CSM	✔	✔	✘	✔	68.81	83.59	75.48	42.49	39
Ours w/o UPL	✔	✔	✔	✘	75.52	87.29	80.98	42.85	42
Ours (Full Model)	✔	✔	✔	✔	**76.55**	**91.84**	**83.50**	42.85	43

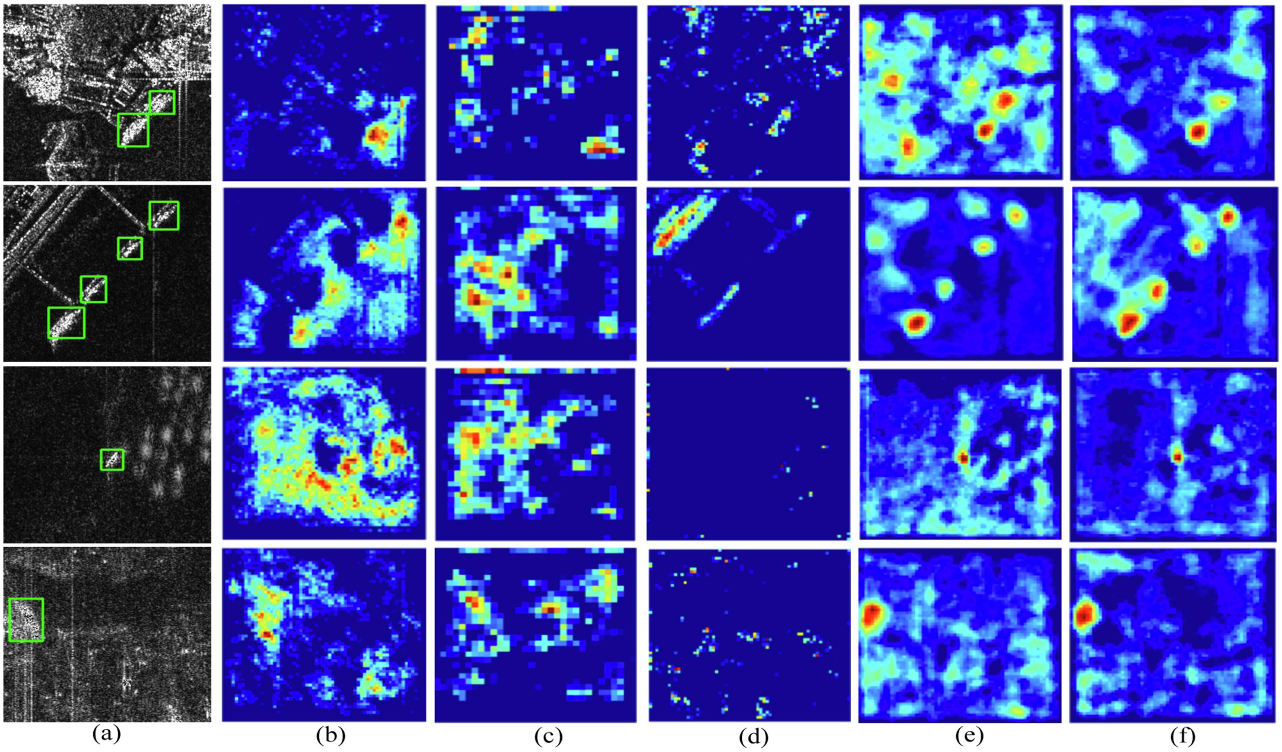


Fig. 7. Visualization of attention maps for different methods on the Gaofen-3 dataset. (a) Ground truth. (b) Ours w/o SM & CSM. (c) Ours w/o CSM. (d) Ours w/o RM. (e) Ours w/o UPL. (f) Ours (Full Model).

at mixed scales. We remove RM from the full model, i.e., Ours w/o RM, to evaluate its effect. From Table IV, it can be seen that Ours w/o RM leads to a significant decrease in Precision and F1-Score, dropping by 5.85% and 3.90%, respectively. This indicates that RM can improve the accuracy of model detection by enhancing the features between different levels at mixed scales. Further, as shown in Fig. 7(d), Ours w/o RM causes the model to fail to perceive ship targets accurately. These positive results indicate that RM has an important contribution to the detection performance of the model.

Effectiveness of UPL:. Considering that indistinguishable background environments easily bring negative effects to model learning, an uncertain perception loss (UPL) is introduced to facilitate the model to produce more reliable predictions. We remove the UPL from the full model, i.e., the Ours w/o UPL. From Table IV, it can be seen that the Ours w/o UPL leads to a significant decline in Recall and F1-Score. Moreover, as shown in Fig. 7(e), it can be observed that the Ours w/o UPL causes the model to not accurately focus on the actual ship target, especially in areas where the contour of the ship target is blurred and uncertain, as shown in the first row in Fig. 7(e). This further demonstrates the importance of UPL in improving the model's ability to deal with complex samples and focus on the actual ship target regions.

Effect of Backbone: Different backbone networks may exhibit different predictive capabilities when handling the same task. Therefore, in this section, we investigate the performance of the proposed MSNet and Baseline on different backbone networks, as shown in Table V. From Table V, it can be seen that as the backbone network complexity increases, the detection performance of both Baseline and MSNet on the Gaofen-3 dataset improves. Second, for all backbone networks (ResNet18, ResNet50, and ResNet101), the MSNet approach significantly improved over the Baseline regarding Precision, Recall, and F1-Score. This shows the MSNet method can better localize and identify ship targets in the SAR ship detection task. Overall, the proposed MSNet demonstrates better detection performance than Baseline, with its advantages particularly pronounced when

TABLE V
DETECTION RESULT OF MSNET AND BASELINE WITH DIFFERENT BACKBONES ON GAOFEN-3 DATASET

Backbone	**Method**	**Precision(%)**	**Recall(%)**	**F1-Score(%)**
ResNet18	Baseline	59.58	78.76	67.84
	MSNet	**62.68**	**78.89**	**69.86**
ResNet50	Baseline	62.98	80.70	70.75
	MSNet	**76.55**	**91.84**	**83.50**
ResNet101	Baseline	69.93	85.23	76.83
	MSNet	**72.70**	**90.14**	**80.49**

TABLE VI
IMPACT OF CSM AT DIFFERENT LEVELS ON GAOFEN-3 DATASET

Method	**Precision(%)**	**Recall(%)**	**F1-Score(%)**
Baseline	68.81	83.59	75.48
Layer1	**76.55**	**91.84**	**83.50**
Layer2	71.98	86.33	78.50
Layer3	70.19	83.75	76.37
Layer1 & 2	68.85	84.14	75.73
Layer1 & 3	68.84	83.53	75.48
Layer2 & 3	66.69	86.27	75.23
Layer1 & 2 & 3	65.89	83.79	73.77

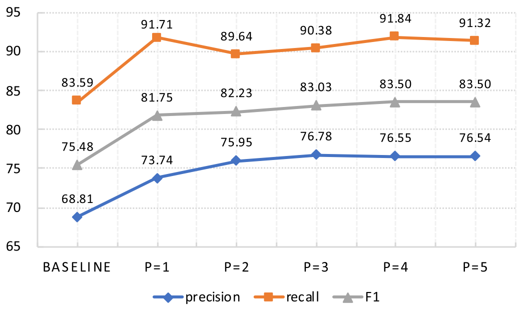


Fig. 8. Evaluation on the hyperparameter p on Gaofen-3 dataset.

using the ResNet50 backbone network. This further confirms the effectiveness of the proposed MSNet in enhancing the detection performance of SAR ships.

Where to apply CSM: We investigate the impact of CSM at different levels. Given that a standard ResNet model has four residual blocks defined as Layer1-4. For notation, Layer1 means CSM is applied after the first residual block; Layer1&2 means CSM is applied after both the first and second residual blocks; and so forth. It is worth noting that since Layer4 is the closest prediction layer, CSM is not applied to Layer4. The baseline means the CSM module is removed from the full model. The results are shown in Table VI. We make the following observations. 1) Applying CSM to multiple layers leads to poor detection results. This suggests that self-distillation learning at multiple levels prevents the model from focusing on critical information about the ships. 2) Applying CSM to a single layer can bring obvious performance gain to the model, especially in Layer1. This is reasonable since Layer1 contains larger-scale features, which encompass richer structural information. This structural information can provide more differentiated information between the ship target and the background environment. By applying CSM after Layer1, the model can be promoted to learn the critical features of the ship target, thus improving the accuracy of model detection. In conclusion, applying the CSM module at different levels has varying effects on model performance. Applying CSM at lower layers, particularly after Layer1, improves performance in the SAR ship detection task.

Sensitivity of Hyperparameter: As mentioned in (2), the parameter p is used to put more weights on the spatial position corresponding to the region with the highest activation value (i.e., put more weights on the more discriminating parts). Here, we evaluate the impact of different p values on model performance. The baseline indicates that the CSM is removed from the whole model. As shown in Fig. 8, Precision, Recall, and F1-Score gradually increase as the p value increases and reach a peak at $p = 4$. This shows that increasing the p can help improve the performance of the model within a certain range. However, the performance begins to decline when the p value increases beyond a certain point. This may be due to an excessive focus on local features, leading to a decrease in the model's generalization ability in other regions. Based on the above analysis, we suggest that when using the CSM, a larger p (e.g., $p = 4$) can be selected for better detection performance. Meanwhile, it is necessary to balance the model's detection performance and generalization ability when selecting the p, to avoid overfitting caused by excessive attention to local features.

RM versus FPN: To further explore the effectiveness of the RM proposed in this article, we discuss the impact of the RM and the FPN [24] module on model performance. The core idea of the FPN is to up-sample high-level features and connect them with low-level features from top to bottom to achieve scale enhancement. Different from this, our proposed RM first divides the feature layer into high-level, mid-level, and low-level. Then, high-level features and mid-level features are connected to integrate hierarchical information. The connected features are further enhanced through the RF [45] module. RF module can facilitate the model to capture more discriminative feature representations by enlarging RFs. We introduce the FPN and RM modules into the model to evaluate their impact on model performance. The experimental results are shown in Table VII. In Table VII, the *Baseline* indicates the model without RM and FPN modules. *Ours with FPN* means that our proposed method uses the FPN module. *Ours with RM* means that our proposed method uses the RM module. As shown in Table VII, compared with *Baseline*, both *Ours with RM* and *Ours with FPN* improve the Precision and AP metrics. However, *Ours with RM* improves is more obvious in Precision (+5.85%) and AP (+3.47%). Notably, *Ours with RM* introduces more parameters. This is because of the inclusion of the RF module. Nevertheless, this resource increase enables the model to capture ship target information across various scales better, leading to enhanced ship detection accuracy. Overall, compared with the FPN, the RM can significantly improve the ship detection performance of the model, although it brings a slight increase in parameters.

Effect of Different Layers: In the proposed method, we employ a triple feature extractor for multilevel feature extraction on inputs at three different scales. Subsequently, we aggregate the multilevel features corresponding to images from

TABLE VII
COMPARISON OF THE IMPACT OF FPN AND RM ON THE GAOFEN-3 DATASET

Method	Precision(%)	Recall(%)	F1-Score(%)	AP(%)	Params.(M)
Baseline	70.70	91.06	79.60	58.39	37.45
Ours with FPN	71.29	86.95	78.35	57.51	37.59
Ours with RM	76.55	91.84	83.50	61.86	42.85

TABLE VIII
COMPARISON OF THE IMPACT OF FEATURE AGGREGATE ON DIFFERENT LAYERS ON THE GAOFEN-3 DATASET

Method	Precision(%)	Recall(%)	F1-Score(%)	AP(%)
C1,C2,C3	56.30	75.52	64.51	48.42
C2,C3,C4	65.56	81.18	72.54	54.57
C3,C4,C5	66.93	82.28	73.82	55.01
C2,C3,C4,C5	69.25	83.30	75.63	57.05
C1,C2,C3,C4,C5	**76.55**	**91.84**	**83.50**	**61.86**

different scales at the same level. In this section, we discuss the impact of aggregating features at different levels on model performance. Specifically, ResNet50 [41] is used as our backbone, which contains five output feature layers, i.e., $\{\text{conv}, \text{layer1}, \text{layer2}, \text{layer3}, \text{layer4}\}$. Here, we label these five output feature layers as C1, C2, C3, C4, and C5, respectively. We conduct multiple sets of experiments to explore the impact of aggregating features at different levels on model performance. The experimental results are shown in Table VIII. From Table VIII, it can be evident that when we aggregate features from all feature output layers, the model exhibits the best performance in terms of Precision, Recall, F1-Score, and AP metrics. This indicates that ship semantic information from different scales images are fully integrated during the feature aggregate process, thereby improving model detection accuracy.

IV. CONCLUSION

This article proposes a MSNet for accurate ship detection in SAR images with complex scenarios. MSNet leverages the zoom strategy to capture more ship target information and aggregates differentiated information between ship targets and background environments at different scales through the designed SM. The CSM is proposed to match feature attention maps at different scales, which forces the model to capture the potential semantic attributes of ship targets through a self-distillation fashion. The RM is developed to enhance further the discriminative semantics among different hierarchical features under mixed scales. Additionally, considering that uncertainty stems from indistinguishable background interference, an UPL is introduced to facilitate the model to produce more reliable predictions. Extensive experiments are performed on the SSDDs from the Gaofen-3, RadarSat-2, Sentinel-1, and TerraSAR satellites. The experimental results consistently demonstrate the superiority of our method over the existing approaches. Furthermore, detailed model analysis experiments further validate the effectiveness of the proposed MSNet approach in SAR image ship detection tasks. However, it is worth noting that since MSNet searches and aggregates crucial semantic clues related to ship targets across multiple scales, it results in a slight increase in model parameters and inference time. In future research, we aim to address this concern and optimize the model's efficiency.

REFERENCES

[1] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou, "A tutorial on synthetic aperture radar," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 1, pp. 6–43, Mar. 2013.

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

[3] A. Reigber et al., "Very-high-resolution airborne synthetic aperture radar imaging: Signal processing and applications," *Proc. IEEE*, vol. 101, no. 3, pp. 759–783, Mar. 2013.

[4] 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, pp. 1–16, 2023.

[5] G. Gao, "A Parzen-window-kernel-based CFAR algorithm for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 557–561, May 2011.

[6] T. Zhang, J. Ji, X. Li, W. Yu, and H. Xiong, "Ship detection from PolSAR imagery using the complete polarimetric covariance difference matrix," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2824–2839, May 2019.

[7] T. Zhang, S. Quan, Z. Yang, W. Guo, Z. Zhang, and H. Gan, "A two-stage method for ship detection using PolSAR image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[8] S. Jiang, C. Wang, B. Zhang, and H. Zhang, "Ship detection based on feature confidence for high resolution SAR images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2012, pp. 6844–6847.

[9] F. C. Robey, D. R. Fuhrmann, E. J. Kelly, and R. Nitzberg, "A CFAR adaptive matched filter detector," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 28, no. 1, pp. 208–216, Jan. 1992.

[10] A. C. Frery, H.-J. Muller, C. d. C. F. Yanasse, and S. J. S. Sant'Anna, "A model for extremely heterogeneous clutter," *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 3, pp. 648–659, May 1997.

[11] C. P. Schwegmann, W. Kleynhans, and B. P. Salmon, "Manifold adaptation for constant false alarm rate ship detection in South African oceans," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 8, no. 7, pp. 3329–3337, Jul. 2015.

[12] X. Qin, S. Zhou, H. Zou, and G. Gao, "A CFAR detection algorithm for generalized gamma distributed background in high-resolution SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 806–810, Jul. 2013.

[13] C. Wang, Z. Wang, H. Zhang, B. Zhang, and F. Wu, "A PolSAR ship detector based on a multi-polarimetric-feature combination using visual attention," *Int. J. Remote Sens.*, vol. 35, no. 22, pp. 7763–7774, 2014.

[14] G. Atteia and M. J. Collins, "On the use of compact polarimetry SAR for ship detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 80, pp. 1–9, 2013.

[15] C. Wang, F. Bi, L. Chen, and J. Chen, "A novel threshold template algorithm for ship detection in high-resolution SAR images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 100–103.

[16] J. Zhu, X. Qiu, Z. Pan, Y. Zhang, and B. Lei, "Projection shape template-based ship target recognition in TerraSAR-X images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 222–226, Feb. 2017.

[17] D. Li, Q. Liang, H. Liu, Q. Liu, H. Liu, and G. Liao, "A novel multidimensional domain deep learning network for SAR ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

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

[19] 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 Obs. Remote Sens.*, vol. 16, pp. 1042–1056, 2023.

[20] J. Cui, H. Jia, H. Wang, and F. Xu, “A fast threshold neural network for ship detection in large-scene SAR images,” *IEEE J. Sel. Topics Appl. Earth Observ Remote Sens.*, vol. 15, pp. 6016–6032, 2022.

[21] Y. Zhou, F. Zhang, Q. Yin, F. Ma, and F. Zhang, “Inshore dense ship detection in SAR images based on edge semantic decoupling and transformer,” *IEEE J. Sel. Topics Appl. Earth Observ Remote Sens.*, vol. 16, pp. 4882–4890, 2023.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

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

[24] 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.

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[26] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, “Light-head R-CNN: In defense of two-stage object detector,” 2017, *arXiv:1711.07264*.

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

[28] C. Zhao, X. Fu, J. Dong, R. Qin, J. Chang, and P. Lang, “SAR ship detection based on end-to-end morphological feature pyramid network,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4599–4611, 2022.

[29] 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, pp. 1–5, 2022.

[30] B. Yue, W. Zhao, and S. Han, “SAR ship detection method based on convolutional neural network and multi-layer feature fusion,” in *Proc. Int. Conf. Natural Comput. Fuzzy Syst. Knowl. Discov.*, Springer, 2020, pp. 41–53.

[31] S. Shin, Y. Kim, I. Hwang, J. Kim, and S. Kim, “Coupling denoising to detection for SAR imagery,” *Appl. Sci.*, vol. 11, no. 12, 2021, Art. no. 5569.

[32] H. Su, S. Wei, M. Wang, L. Zhou, J. Shi, and X. Zhang, “Ship detection based on RetinaNet-plus for high-resolution SAR imagery,” in *Proc. Asia-Pac. Conf. Synth. Aperture Radar*, 2019, pp. 1–5.

[33] 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, pp. 044522–044522, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229934967

[34] 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.

[35] J. Zhang, M. Xing, G.-C. Sun, and N. Li, “Oriented Gaussian function-based box boundary-aware vectors for oriented ship detection in multiresolution SAR imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[36] F. Gao, Y. Huo, J. Sun, T. Yu, A. Hussain, and H. Zhou, “Ellipse encoding for arbitrary-oriented SAR ship detection based on dynamic key points,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–28, 2022.

[37] Y. Sun, X. Sun, Z. Wang, and K. Fu, “Oriented ship detection based on strong scattering points network in large-scale SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[38] P. Guo, T. Celik, N. Liu, and H.-C. Li, “Break through the border restriction of horizontal bounding box for arbitrary-oriented ship detection in SAR images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[39] Y. Sun, Z. Wang, X. Sun, and K. Fu, “SPAN: Strong scattering point aware network for ship detection and classification in large-scale SAR imagery,” *IEEE J. Sel. Topics Appl. Earth Observ Remote Sens.*, vol. 15, pp. 1188–1204, 2022.

[40] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, “Zoom in and out: A mixed-scale triplet network for camouflaged object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2150–2160.

[41] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[42] N. Komodakis and S. Zagoruyko, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–13.

[43] M. Ji, S. Shin, S. Hwang, G. Park, and I.-C. Moon, “Refine myself by teaching myself: Feature refinement via self-knowledge distillation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10664–10673.

[44] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, “Camouflaged object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2774–2784.

[45] S. Liu et al., “Receptive field block net for accurate and fast object detection,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 385–400.

[46] 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.

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

[48] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, Art. no. 721.

[49] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.

[50] N. Wang et al., “NAS-FCOS: Fast neural architecture search for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11943–11951.

[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.

[52] H. Zhang, Y. Wang, F. Dayoub, and N. Sünderhauf, “VarifocalNet: An iou-aware dense object detector,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8510–8519.

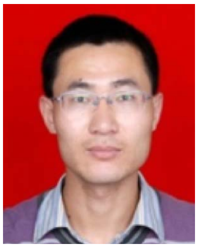
[53] J. Wang et al., “Side-aware boundary localization for more precise object detection,” in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K.: Springer, 2020, pp. 403–419.

[54] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, “Region proposal by guided anchoring,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2965–2974.

[55] 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.

Shuang Liu received B.S. and M.S. degrees in electronics and information engineering from South-Central Minzu University, Wuhan, China, in 2018 and 2021, respectively. She is currently working toward the Ph.D. degree in information and communication engineering from Chongqing University, Chongqing, China.

Her research interests include image processing, deep learning, transfer learning, and object detection.



Dong Li (Senior Member, IEEE) received the B.S. degree in automation from Chengdu University of Information Technology, Chengdu, China, the M.S. degree in signal processing from Sichuan University, Chengdu, and the Ph.D. degree in signal and information processing from the National Laboratory of Radar Signal Processing, Xidian University, Xi’an, China, in 2007, 2010, and 2014, respectively.

He is currently a Professor with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China. His research interests include target detection and tracking, and parameter estimation.

Renjie Jiang received the bachelor’s degree in electronics and information engineering from Ningxia University, Yinchuan, China, in 2021. He is currently working toward the master’s degree in electronic information from Chongqing University, Chongqing, China.

His research interests include deep learning and object detection.

Qinghua Liu received the B.S. degree in physics from Sichuan Normal University, Chengdu, China, the M.S. degree in signal and information processing from Guilin University of Electronic Technology, Guilin, China, and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 1995, 2001, and 2014, respectively.

She is currently a Professor with the School of Information and Communication, Guilin University of Electronic Technology, Guilin. Her research interests include target detection and active noise control.

Jun Wan (Member, IEEE) received the B.Eng. degree in electrical engineering from Fuzhou University, Fuzhou, China, in 2015, and the Ph.D. degree in signal processing from Xidian University, Xi'an, China, in 2020.

In 2020, he joined the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China, where he is currently a Lecturer. His current research interests include ground moving target indication and imaging, moving/maneuvering target detection, and time–frequency analysis.

Xiaopeng Yang (Senior Member, IEEE) received B.E. and M.E. degrees from Xidian University, Xi'an, China, in 1999 and 2002, respectively, and the Ph.D. degree in electrical and communication engineering from Tohoku University, Sendai, Japan, in 2007.

He was a Post-Doctoral Research Fellow at Tohoku University from 2007 to 2008 and a Research Associate at Syracuse University, Syracuse, NY, USA, from 2008 to 2010. He has been working with Beijing Institute of Technology (BIT), Beijing, China, where he is currently a Full Professor and the chair of the Radar research laboratory. His current research interests include radar signal processing, ground penetrating radar, and through the wall radar.

Hehao Liu received the bachelor's degree in electronic information engineering from Tiangong University, Tianjin, China, in 2022. He is currently working toward the master's degree in electronic information with Chongqing University, Chongqing, China. His research interests include deep learning and object detection.