A Single-Stage Arbitrary-Oriented Detector Based on Multiscale Feature Fusion and Calibration for SAR Ship Detection

Shuang Zhao, Qi Liu, Weidong Yu, Member, IEEE, and Jiyu Lv

Abstract—Ship detection is a challenging task in the synthetic aperture radar (SAR) automatic target recognition due to the large aspect ratio, arbitrary orientation, and dense arrangement of ships and severe background interference in the inshore scenes. Although considerable progress has been made in recent research, there have still been certain challenges in achieving fast and efficient detection of arbitrary-oriented ships in SAR images. To address these challenges, this article proposes a single-stage detection method based on multiscale feature fusion and calibration. The proposed detection method can detect arbitrary-oriented ships in SAR images with high accuracy and speed. Specifically, a head network with the stepwise regression from the coarse- to fine-grained detection is designed to detect arbitrary-oriented ships accurately. In addition, a feature enhancement module is constructed to fuse and refine shallow texture features and deep semantic features, aiming to obtain multiscale fusion features containing sufficient contextual information. Finally, an attention module is used to calibrate multiscale fusion features to highlight the ship information while suppressing interference from the surrounding background. The effectiveness of the proposed method is verified by experiments on two public SAR ship datasets and a panoramic SAR image. The experimental results show that compared with the other rotation detectors, the proposed detection method has competitive detection results and can achieve state-of-the-art detection performance.

Index Terms—Arbitrary-oriented detector, feature enhancement, ship detection, synthetic aperture radar (SAR).

I. INTRODUCTION

SYNTHETIC aperture radar (SAR) has the ability of all-day, all-weather operation [1]. It can acquire high-resolution images uninterruptedly and monitor a marine environment dynamically and continuously for a long time, and it plays an important role in marine applications [2]. Ship detection is an important part of the SAR marine applications, and it is of great significance to maritime transportation, fishery management, and maritime rescue [3]. However, due to the large aspect ratio, arbitrary orientations, dense arrangement of ships, as well as severe background interference in the inshore scenes, fast and accurate ship detection from massive SAR images has still been a challenge.

Traditional SAR ship detection methods can be roughly divided into four types: methods based on statistical characteristics [2], [4], [5], methods based on visual saliency [6], [7], methods based on wavelet transform [8], [9], and methods based on polarization information [10], [11]. Among them, the constant false alarm rate (CFAR) method based on statistical characteristics has been the most widely researched algorithm. The CFAR method has been broadly applied to practical ship detection systems, and it is easy to operate. Specifically, it detects ships by comparing the gray value of pixels to be detected with an adaptive threshold, which is set through statistical modeling of sea clutter. The CFAR method can provide good detection results in an offshore scene. However, the CFAR method has certain disadvantages. First, the adaptive threshold strongly depends on the sea clutter distribution, so it is challenging to select a suitable statistical model of sea clutter for ship detection in multiple scenes. Second, the detection performance can deteriorate significantly when performing the ship detection in the background of a heterogeneous clutter. Therefore, the traditional ship detection methods represented by the CFAR method are not suitable for modern SAR ship detection tasks.

Due to the powerful feature representation and extraction capabilities of deep convolutional neural networks (CNNs) [12], the CNN-based object detection methods [13], [14], [15], [16], [17], [18], [19], [20], [21] have made significant progress for natural images. Inspired by this, the two-stage detector named the Faster RCNN [15] has been introduced into the field of SAR ship detection in recent years. A series of modified detectors based on the Faster RCNN have been explored [22], [23], [24], [25], [26], [27]. Although the two-stage detector can achieve high accuracy, it is difficult to obtain a large improvement in detection speed due to the need to generate proposals in a region proposal network. However, compared with a two-stage detector, a single-stage detector has a simpler structure but excellent speed and can meet the real-time requirements of SAR ship detection in practical applications. Therefore, recent studies have paid more attention to single-stage SAR ship detectors. For instance, Wang et al. [28] used the RetinaNet [21] to detect ships in multiresolution GF-3
images automatically and excellent detection performance was achieved. Fu et al. [29] used a scene classification network before a single-shot multibox detector (SSD) [20] to address the problem of low detection efficiency in large-scale images in near-shore regions. Chang et al. [30] improved the YOLOv2 model [19] by combining convolutional layers to achieve fast ship detection in SAR images. In [31] and [32], the authors redesigned a feature extraction network of a single-stage detector, which solved the problem of insufficient representation of ship features in SAR images to a certain extent and improved detection performance. However, the abovementioned methods all represent and locate ships in the form of horizontal bounding boxes. In fact, ships in SAR images usually have arbitrary orientations and are densely arranged, so using horizontal bounding boxes in the SAR ship detection can cause certain problems. First, a horizontal bounding box usually contains many background areas, resulting in an inaccurate representation of the target range. Second, ships in ports and near-shore are usually densely arranged, so they cannot be effectively distinguished, resulting in numerous false negatives.

To solve the aforementioned problems, oriented bounding boxes have been proposed to detect arbitrary-oriented ships in SAR images. An et al. [33] proposed a rotation detector using a multilevel anchor generation strategy and a modified RBox encoding method to improve detection accuracy, but this method is mainly applicable to optical remote sensing images. Pan et al. [34] adopted a network structure similar to the RRPN [35] and combined it with a multistage rotation detection network to solve the problem of false positives in a complex environment and improve the SAR ship detection accuracy. Chen et al. [36] proposed a cascaded refined feature alignment detector combining horizontal and oriented anchors, which can address the problem of background clutter in complex environments. Yang et al. [37] developed an improved oriented framework named the R-RetinaNet from the perspective of positive sample distribution and feature matching. This framework can solve the problems of an unbalanced distribution of positive samples and a mismatch of ship features between SAR images and a single-stage detector.

Although many rotation detectors have been proposed in recent years, there have still been many problems in arbitrary-oriented ship detection in SAR images. There is a lack of arbitrarily oriented ship detectors for SAR images that can balance efficiency and accuracy well and where a network can converge easily. In the existing rotation ship detectors, detection performance has been mainly improved in two ways: by generating densely oriented anchors [33], [34], [37], where multiple anchors with different angles are generated at each pixel to cover the target as much as possible, or by using a multistage cascade network to perform multiple classification and regression in the detection head [34], [36]. The first method produces a large number of anchors, which increases the difficulty of network convergence significantly. Also, many similar predicted boxes may be output for the same target, which reduces detection accuracy. The second method improves the detection accuracy using a multilevel cascade scheme but increases network complexity and reduces detection efficiency. Therefore, it is necessary to develop a rotation SAR ship detector that can easily converge and balance efficiency and accuracy well. In addition, feature representations of most of the existing arbitrary-oriented SAR ship detectors are insufficient and rough to achieve precise ship positioning. The existing arbitrary-oriented SAR ship detectors mainly use a feature pyramid network (FPN) [38] to obtain multiscale fusion features. However, the FPN up-samples high-level features only through interpolation, which is a rough operation. Moreover, ships in SAR images often can have various geometric shapes and large aspect ratios, but ship features obtained by an FPN using a fixed single-scale convolution kernel are local and incomplete. Sun et al. [39] used a structure similar to the PAFPN [40] for feature extraction, introducing an additional bottom-up path. Compared with optical remote sensing images, SAR images have lower resolution and obvious speckle noise, and a too-long propagation path may cause losses of ship texture details. Su et al. [41] introduced deformable convolution into the backbone network and proposed a lightweight fully convolutional network to improve the network ability to extract rotation information. Fu et al. [42] designed a contextual feature selection module to learn local and contextual features in the channel dimension dynamically, and significantly reduced the number of false positives. However, contextual information learned by the aforementioned methods has been limited, and ships in SAR images have the characteristics of multiple scales, large aspect ratio, arbitrary orientations, and are usually embedded in complex backgrounds (e.g., wharves, buildings, islands, and ship wakes). Therefore, more contextual information is required to enhance ship features while suppressing interference from the surrounding background. Thus, in-depth research on feature enhancement methods is urgently needed to improve the performances of SAR ship detectors.

To address the abovementioned problems, this article proposes a single-stage arbitrary-oriented SAR ship detection framework based on multiscale feature fusion and calibration named the SaDet. First, a coarse-to-fine detection strategy is designed to detect arbitrary-oriented ships in SAR images. This strategy is a stepwise regression approach from coarse-to-fine-grained detection. In the coarse-detection stage, similar-to-horizontal anchors are used to obtain enough proposals rapidly. In the fine-detection stage, boxes predicted in the coarse-detection stage are first refined, and boxes with the highest score at each of the pixels are retained to improve detection efficiency. At the same time, a feature refinement module is designed to refine features, which are then input to the fine-detection head for further detection. In addition, a multilevel feature enhancement module (MFFM), which aims to make full use of the features extracted by a backbone network and can obtain multiscale fusion features that contain sufficient contextual information, is constructed. This module uses a multilevel feature fusion module (MFFM) that performs deconvolution to improve the resolution of high-level features with a low resolution and strong semantics. Compared with the feature fusion methods based on interpolation [38], [40], the deconvolution operation can learn more nonlinear relations dynamically. The obtained features are fused with low-level features that have a high-resolution but weak semantic to obtain multiscale fusion features. Next, a feature enhancement module (FEM), which consists of multiple
branches of dilated convolutions with different kernel sizes and dilation rates, is adopted to capture rich contextual information to enhance the fused multiscale features. Finally, the spatial attention module (SAM) is used to calibrate the features to distinguish ships from the surrounding background, highlighting the ship features while suppressing interference from the surrounding backgrounds. The proposed method is verified by experiments on two public SAR ship datasets and a panoramic SAR image. Experimental results show that the proposed method can achieve state-of-the-art detection performance.

The major contributions of this work can be summarized as follows.

1) A single-stage detector based on multiscale feature fusion and calibration, which aims to detect arbitrary-oriented ships in SAR images with high accuracy and speed, is proposed. The proposed method adopts a coarse-to-fine detection strategy. In the coarse-detection stage, similar-to-horizontal anchors are used to obtain enough proposals rapidly. The refined oriented anchors and features are used in the refinement stage to accommodate the arbitrary-oriented ships in a SAR image. Compared with other rotation detectors, the proposed method can significantly reduce the number of generated oriented anchors, making the network converge easily and balancing the efficiency and accuracy well.

2) A multilevel feature enhancement strategy is developed to fuse and refine shallow texture features and deep semantic features to obtain multiscale fusion features that contain sufficient contextual information. This strategy can remove false positives from the detection results effectively, thus improving the detection accuracy of multiscale ships.

3) A feature calibration module, which uses the attention mechanism to calibrate features in the spatial dimension, is constructed. It highlights the important ship features and suppresses the interference from the surrounding background, which improves the anti-interference performance of the proposed method.

4) Extensive experiments on two public datasets are performed. The proposed method is compared with other rotation detectors, and the comparison results indicate that the proposed method can achieve state-of-the-art detection performance.

The rest of this article is organized as follows. Section II describes the proposed method in detail. Section III presents the experimental results. Finally, Section IV concludes this article.

II. METHODOLOGY

This section introduces the overall structure of the proposed method, describes the internal modules of the proposed method in detail, and calculates the overall loss of the proposed method.

A. Proposed Method Structure

The proposed method is a single-stage detector, and its overall architecture is shown in Fig. 1. In the proposed method, the top–down path is followed by the backbone network, feature enhancement network, and detection head, which are explained in detail in the following.

Backbone: The ResNet50 model is used as a backbone to extract basic features from input SAR images. The outputs of the last four stages of the backbone are used as the input of the feature enhancement network. It should be noted that features at different levels have different sizes. The sizes of the four output features are 1/4, 1/8, 1/16, and 1/32 of the input image.

Feature Enhancement: To improve detection performance for multiscale ships, features extracted by the backbone are enhanced. First, four groups of MFEMs are used to fuse and enhance the features. Three backward propagation paths, which are denoted by red arrows in Fig. 1, are introduced to fuse low-resolution, strong-semantic features with high-resolution, weak-semantic features to obtain features with rich information. Considering that ships in the inshore scenes are susceptible to interference from a complex surrounding background, including wharves, buildings, islands, and ship wakes, four groups of SAMs are used after the MFEM to highlight ship features and suppress interference from the surrounding background.

Coarse-to-Fine Head: The proposed method adopts a coarse-to-fine detection strategy to detect arbitrary-oriented ships in SAR images with high accuracy and speed. The proposed method is a stepwise regression approach from coarse- to fine-grained detection. Particularly, the detection head consists of two stages, each of which is divided into two branches, a classification branch and a regression branch. Each branch consists of four groups of 3 × 3 convolutions. In the coarse-detection stage,
similar-to-horizontal anchors are used to obtain enough number of proposals fast. In the fine-detection stage, boxes predicted in the coarse-detection stage are first refined, and then boxes with the highest score at each of the pixels are retained to improve detection efficiency. The refined features are fed to the fine-detection head for further processing.

To optimize the proposed method, a multitask loss with the focal loss and smooth L1 loss is employed to supervise the classification and regression branches of the coarse-to-fine head.

### B. MFEM

To obtain multiscale fusion features with sufficient contextual information, an MFEM is introduced when building a fine-grained feature pyramid, as shown in Fig. 2. The MFEM consists of an MFFM and a FEM. The \( \text{type@}k\times k\_sA\_pB\_cD \) represents a general expression of a specific convolution operation in the MFEM, where \( \text{type} \), \( k \), \( A \), \( B \), and \( D \) denote the convolution type, kernel size, stride, padding pixels, and output channels, respectively; \( C \) represents a basic feature extracted by the backbone; \( M_i \) is an intermediate feature; \( F_i \) represents a final fine-grained feature obtained by the MFEM.

**MFFM:** To obtain features with rich information, low-resolution, strong-semantic features are fused with high-resolution, weak-semantic features. First, a \( 1 \times 1 \) convolution is applied to \( C_i \) and \( C_{i+1} \) to reduce the channel dimension. Next, \( C_{i+1} \) uses a \( 4 \times 4 \) deconvolution with a stride of two to recover the spatial resolution of features to obtain rich semantic features, which helps to detect small-sized ships from shallow features efficiently. Then, transformed \( C_i \) and \( C_{i+1} \) are element-wise added to obtain \( M_i \) that contains rich location information and semantic features. Finally, a \( 3 \times 3 \) convolution is applied to the fused features to reduce the aliasing effect caused by the deconvolution operation. Specifically, \( C_2 \) and \( C_3 \) are used to construct the lowest intermediate feature \( M_2 \). Intermediate features \( M_3 \) and \( M_4 \) are obtained from \( C_3 \) and \( C_4 \), \( C_4 \), and \( C_5 \), respectively. Since \( C_5 \) has the smallest size and the strongest semantic among all features, the top-most intermediate features \( M_5 \) are directly constructed by simple two-dimensional convolution.

**FEM:** The success of inception [43], [44] demonstrates the effectiveness of a multibranch architecture, where features with a different sparsity can be obtained by applying convolution kernels of different sizes on the same feature. In addition, atrous convolution [45], [46] can effectively expand the receptive field and capture multiscale ship features, and it has been shown to be an effective tool for densely arranged ship detection in SAR images [47], [48]. Inspired by this, this study uses multiple branches and applies atrous convolution with different kernel sizes and atrous rates on each branch to obtain multiscale ship features containing sufficient contextual information.

The FEM consists of three convolution branches and a residual branch. This module is applied to the intermediate features. First, each branch adopts a bottleneck structure, and a channel hyperparameter reduction is used to determine the channel number of the bottleneck structure. In addition, each branch adopts atrous convolution with different kernel sizes and atrous rates to learn nonlinear relationships. Multibranched atrous convolutions can capture useful information from larger areas, enhancing local features of ships by establishing global dependencies between pixels. Then, the features of all branches are combined into a convolution array. A \( 1 \times 1 \) convolution is used to obtain features of the same size as the input features. Further, a residual branch is constructed to avoid gradient vanishing during model training while preserving the original information of the input. Finally, the features of all branches are element-wise added, and the final output is obtained by a nonlinear rectified linear unit.

### C. SAM

Although enhanced features can accumulate rich information on ships, the collected data may include certain disturbances caused by complex backgrounds (e.g., wharves, buildings, islands, and ship wakes). Therefore, a SAM [49] is adopted after the MFEM to highlight ship features and suppress the interference from the surrounding background. The SAM adopts the idea of crisscrossing to generate a sparse attention map, which captures long-distance dependence between pixels from all pixels through a loop operation. The SAM structure is shown in Fig. 3.

In the SAM, \( F \in \mathbb{R}^{C \times H \times W} \) applies \( 1 \times 1 \) convolution to reduce the computational cost, and features \( Q \) and \( K \) are obtained, where \( Q, K \in \mathbb{R}^{C \times H \times W}; C \) represents the channel number, which is defined by the hyperparameter \( \text{ratio} \), and \( C = C/ratio \), and it is less than \( C \). Another \( 1 \times 1 \) convolution is applied to \( F \) to obtain \( V \in \mathbb{R}^{C \times H \times W} \) to fit features; \( j \) is a position in the space dimension of \( Q \), and \( Q_j \in \mathbb{R}^C \). All elements in the same row or column as a position \( j \) in \( K \) are used to obtain \( \Omega_j \), where \( \Omega_j \in \mathbb{R}^{(H \times W - 1) \times C}; \Omega_{i,j} \in \mathbb{R}^C \) is an...
element at the $i$th position of $\Omega_j$. The correlation degree $x_{i,j}$ between $Q_j$ and $\Omega_{i,j}$ can be obtained by applying the affinity operation between $Q_j$ and $\Omega_{i,j}$. The affinity operation is defined by
\[
x_{i,j} = Q_j \cdot \Omega_{i,j}^T.
\] (1)

All $x_{i,j}$ form a feature $X$, and $X \in \mathbb{R}^{(H+W-1) \times (H+W)}$. A softmax operation is applied to $X$ to construct a sparse spatial attention map $A \in \mathbb{R}^{(H+W-1) \times (H+W)}$. Then, the spatial attention information is integrated into local features $V$ to enhance the features. Finally, an element-wise sum operation is performed on the original features to obtain the final feature set $P$; $P$ includes context information only in the horizontal and vertical directions, which may be insufficient for accurate ship detection. To obtain richer and denser surrounding information, $P$ is used as the input of the SAM again. After several loop iterations, features that can highlight information on ships while suppressing interference from the surrounding background are obtained.

D. Anchor Design

The coarse-to-fine detection strategy is inspired by the R$^2$Det model [50]. In [50], it has been pointed out that horizontal anchors can achieve higher recall with fewer numbers than oriented anchors. Therefore, the proposed method uses similar-to-horizontal anchors in the coarse-detection stage to obtain enough proposals rapidly. The refined oriented anchors are used in the refinement stage to detect arbitrary-oriented ships in SAR images.

In this study, the classic open-cv method is selected to represent oriented bounding boxes, as shown in Fig. 4. Five parameters, namely, $x_{\text{center}}, y_{\text{center}}, \text{width}, \text{height},$ and $\theta$, are used to represent a rectangular box in any direction, where $(x_{\text{center}}, y_{\text{center}})$ represents the center coordinates of an oriented rectangular box; $\text{width}$ is the length of the first edge that coincides with the rectangular box by rotating it in the positive direction of the $x$-axis counter-clockwise; $\text{height}$ represents the length of the other side; $\theta$ is the acute angle between the width direction and the positive $x$-axis direction.

The anchor design is shown in Fig. 5. In the coarse-detection stage, many similar-to-horizontal anchors with different aspect ratios are generated. They are expressed as $(x, y, w, h, \theta)$, where $\theta = 0$, which means that the angle value is set to zero. In the fine-detection stage, only boxes with the highest score at each of the pixels are retained to improve detection efficiency. It should be noted that all settings are designed for anchors during model training. The ground truth boxes change in neither coarse- nor fine-detection stage. During model training, the ground truth boxes in both detection stages use the oriented box annotation information that is obtained from a dataset. In this study, the scale-invariant parameterizations tules, denoted by $g = (g^x, g^y, g^w, g^h, g^\theta)$ and $v = (v^x, v^y, v^w, v^h, v^\theta)$, of the regression branch can be defined as follows:
\[
g^x = (x - x_a)/w_a, \quad g^y = (y - y_a)/h_a \quad (2)
\]
\[
g^w = \log(w/w_a), \quad g^h = \log(h/h_a), \quad g^\theta = \theta - \theta_a \quad (3)
\]
\[
v^x = (x' - x_a)/w_a, \quad v^y = (y' - y_a)/h_a \quad (4)
\]
\[
v^w = \log(w'/w_a), \quad v^h = \log(h'/h_a), \quad v^\theta = \theta' - \theta_a \quad (5)
\]
where $(x, y, w, h, \theta)$ represent the center point coordinates, width, height, and angle of a box, respectively; it should be noted that $x, x_a, \text{and } x'$ represent the ground-truth, anchor, and predicted boxes, respectively, and the same definition principle holds for $y, y_a, \text{and } y'$. The specific architecture of the coarse-to-fine detection head is shown in Fig. 6, where it can be seen that it consists of two nonshared subnetworks corresponding to the classification and regression tasks. In different detection stages, four 256-dimensional $3 \times 3$ convolutional layers are applied to the input features subsequently, and the detection results of the corresponding stages are output. The feature refinement module connects the coarse- and fine-detection stages. It uses the output features of the coarse-detection stage to predict the offsets of
the sampling grid, which are then used to perform deformable convolution with the original features to make the features more consistent with ships in the fine-detection stage. For the classification branch, the feature refinement module can be defined as follows:

\[ \text{offsets} = f_{18\times1}(f_{256\times3}(\text{cls}_x)) \]

\[ x' = f_{256\times3}(x) \]

\[ \text{cls}_{x'} = \text{DeformableConv}(x', \text{offsets}) \]

where \( f_{a\times a} \) is a filter whose output channel number is \( C \), and the convolution kernel size is \( a \times a \); \( \text{cls}_x \) is the output feature of the classification branch in the coarse-detection stage; \( x \) is the original feature of the detection head; \( \text{cls}_{x'} \) is the input feature of the classification branch in the fine-detection stage.

Similarly, the features of the regression branch are refined.

In addition, different intersection over union (IoU) thresholds are used to assign positive and negative samples in the coarse- and fine-detection stages. Specific details are given in Section III.

F. Loss

The overall loss of the proposed method consists of the losses in the coarse- and fine-detection stages, and can be calculated by

\[ \text{Loss} = L_{\text{coarse\_stage}} + L_{\text{fine\_stage}} \]

\[ = \frac{\lambda_1}{N_{\text{coarse}}} (L_{\text{coarse\_cls}}(C) + L_{\text{coarse\_reg}}(V, G)) + \frac{\lambda_2}{N_{\text{fine}}} (L_{\text{fine\_cls}}(C) + L_{\text{fine\_reg}}(V, G)) \]

where \( N_{\text{coarse}} \) and \( N_{\text{fine}} \) represent the numbers of positive samples in the coarse- and fine-detection stages, respectively; \( \lambda_1, \lambda_2 \) are hyperparameters that control the trade-off of the loss in the two stages, and they are both set to one by default.

As defined in (9), the loss in both the coarse- and fine-detection stages consists of \( L_{\text{loss\_cls}}(C) \) and \( L_{\text{loss\_reg}}(V, G) \), where \( L_{\text{loss\_cls}}(C) \) is the confidence loss produced by the classification branch, and \( C = [c_1, c_2, \ldots, c_N]^T \) is the confidence vector predicted by the model, where \( c_i \) is the confidence of the \( i \)th anchor, and \( N \) is the total number of anchor boxes; \( L_{\text{loss\_reg}}(V, G) \) is defined by

\[ L_{\text{loss\_reg}}(V, G) = \sum_{i \in \text{Pos}} \sum_{t \in x, y, w, h, \theta} \text{smooth}_{L_1}(v_i - g_i) \]

where \( \text{smooth}_{L_1} \) is the smooth \( L_1 \) loss, which can be defined as

\[ \text{smooth}_{L_1}(s) = \begin{cases} 0.5s^2 & \text{if } |s| < 1 \\ |s| - 0.5 & \text{else} \end{cases} \]

III. EXPERIMENTAL RESULTS

A. Datasets

Two public SAR ship datasets labeled with rotated bounding box information, namely, the SAR ship detection dataset (SSDD+) [24] and the high-resolution SAR images dataset (HRSID) [51], were used to evaluate the performance of the proposed method.

SSDD+: The SSDD+ dataset is the first public dataset for ship detection in SAR images, and it has been widely used in the field of SAR ship detection. The SSDD+ dataset includes a total of 1160 images and 2456 ship instances. It covers a variety of images with different polarization modes, sensor types, ship scales, and image scenes and thus is very suitable for the performance evaluation of methods under conditions of complex background and diverse target distribution. The specific statistics on the SSDD+ dataset are given in Table I.
TABLE I
Detailed Description of SSDD+

| Parameter      | Value            |
|----------------|------------------|
| Sensors        | RadarSat-2, TerraSAR-X, Sentinel-1 |
| Polarization   | HH, VV, VH, HV   |
| Resolution (m) | 1-15             |
| Number of images | 1160           |
| Number of ships | 2456           |
| Sea condition  | good and bad condition |
| Scene          | inshore, offshore|

TABLE II
Detailed Description of HRSID

| Parameter      | Value          |
|----------------|----------------|
| Sensors        | TerraSAR-X, Sentinel-1B, TanDEM |
| Imaging mode   | S3-SM, ST, HS  |
| Polarization   | HH, VV, HV     |
| Resolution (m) | 0.5, 1, 3      |
| Number of images | 5604         |
| Number of ships | 16951         |
| Size of images (pixels) | 800×800     |
| Scene          | inshore, offshore|

**HRSID:** The HRSID dataset is a high-resolution SAR ship detection dataset, which consists of 5604 high-resolution SAR images and includes 16951 ship instances. It is characterized by high diversity in sensor types, image resolutions, scenes, ship scales, and polarization modes and thus is ideal for evaluating multiscale ship detection performance in high-resolution images. Important parameters of the HRSID dataset are presented in Table II.

**B. Implementation Details**

In this section, specific details on the experimental verification process, including data preprocessing, hyperparameter settings, and network optimization, are given.

**Data Preprocessing:** The SSDD+ and HRSID data were randomly divided into training and test sets according to the ratio of 8:2. To reduce the long training time caused by large input data size, images in the training set from the HRSID dataset were cropped into subimages with a size of 480 × 480 pixels with an overlap of 120 pixels, while images in the training set from the SSDD+ dataset were cropped into subimages with a size of 300 × 300 pixels with an overlap of 50 pixels. The training set was then flipped horizontally with a probability of 0.5 for data augmentation.

**Hyperparameter Settings:** Considering the characteristics of ships with multiscale and different geometric shapes, anchors of three scales \( \{2^0, 2^{1/3}, 2^{2/3}\} \) and three aspect ratios \( \{1, 1/2, 2\} \) were designed on each level feature in the coarse detection stage. Since the rotated bounding box was more sensitive to the IoU computation than the horizontal bounding box, in the postprocessing stage, the nonmaximum suppression (NMS) threshold was set to 0.2, while the score threshold was set to 0.2.

**Network Optimization:** The model was trained using the stochastic gradient descent (SGD) algorithm, and the weight attenuation of the SGD was set to 0.0001; the momentum was set to 0.9. A total of eight images were input to the network in each training iteration. The learning rate was set to 0.001, and the learning rate adjustment during model training adopted the cyclic strategy. The maximum number of training iterations was set based on the image number of the dataset; for the SSDD+ dataset, there were a maximum of 8000 iterations, and for the HRSID dataset, training included maximum 30,000 iterations.

All experiments were implemented in the deep learning framework PyTorch and performed on a computer equipped with an Intel i9-9700 k processor and an Nvidia GTX3090 GPU. The computer operating system was Ubuntu 20.04.

**C. Evaluation Metrics**

To evaluate the proposed method quantitatively, the average precision (AP) and F1-score were used as evaluation metrics in experiments. The AP value indicates the area between the precision-recall curve consisting of recall and precision and the coordinate axes. The AP is calculated as follows:

\[
AP = \int_0^1 P(R) dR
\]

(14)

where \( P \) represents the precision, which denotes the proportion of correctly detected ships in all detection results, and it is defined by

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

(15)

Further, \( R \) refers to the recall, which denotes the proportion of correctly detected ships in all ground-truth data, and is defined as follows:

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

(16)

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

It should be noted that the AP value is a comprehensive measure of detection performance; the larger the AP value is, the better the detection performance will be.

In addition, the F1-score is a harmonic average of precision and recall, and it has been a common, comprehensive evaluation metric of detection performance; it is calculated by

\[
\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

(17)

**D. Ablation Experiment**

To illustrate the effectiveness of the proposed method, the components of the proposed method were analyzed comprehensively. Specific contributions of the internal modules were
analyzed by progressively adding the components of the proposed method to the baseline, and the results are shown in Table III. First, the MFEM, SAM, and the coarse-to-fine head were removed from the proposed method, and only the basic architecture was maintained to construct a baseline detector that completely depended on the rotated anchor in the detection head. The anchors of the baseline detector were set with six rotation angles of \{-90°, -75°, -60°, -45°, -30°, -15°\}. Furthermore, the MFEM was split into two parts, namely, the MFFM and the FEM, and their individual contributions to the ship feature enhancement in SAR images were analyzed in detail. In addition, considering that the proposed coarse-to-fine detection strategy has been inspired by the R3Det model, the first row in Table III provides the results of the R3Det model. Based on the results in Table III, several insightful findings about the proposed method can be summarized as follows.

1) The experimental results indicated that each improvement in the proposed structure could effectively improve the performance of the baseline, which proved the significance of each component of the proposed method.

2) The coarse-to-fine detection strategy improved the AP by 4.52% compared to the baseline, indicating that the idea of using a coarse-to-fine head to replace a head that completely depends on the rotated anchor was feasible. The inference time increased because of the extra convolution operations added to the coarse-to-fine head. In the experiments, the proposed method was trained for only 8000 iterations, while the baseline detector was trained for 34,000 iterations, which shows that the proposed method could converge easier than the method that completely depended on the rotated anchor.

3) The MFFM provided the largest performance gain for F1-score (an increase of approximately 5%), and it improved the precision significantly (from 75.17% to 82.44%) with a slight increase in the recall value, which showed that the false positives were effectively removed through the multi-level feature fusion. After the introduction of the FEM, the AP and F1-score values increased by approximately 1%. After the introduction of the SAM, the detector achieved the best performance, with an AP value reaching 95.34% and an F1-score reaching 90.70%. The importance of feature enhancement for ship detection in SAR images can be illustrated by comparing the results of progressively adding the MFFM, FEM, and SAM to the baseline.

4) The proposed method performed better on various evaluation metrics than the R3Det model. This was due to two main reasons. First, the proposed method used features that were suitable for ships in SAR images in the detection head. The R3Det model used the P3–P7 level features generated by the FPN network in the detection head. While the proposed method comprehensively considered the characteristics of SAR images and ship targets and selected P2–P5 level features enhanced by the MFEM and SAM blocks for detection, which improved detection performance. But, since the proposed method used shallow features of a larger size, and the feature enhancement method introduced additional convolution operations, the inference time was prolonged. Second, the proposed method used deformable convolution to refine ship features adaptively, while the R3 Det model reconstructed features through interpolation. The specific comparison results of the two feature refinement modules are discussed in detail in Section III.

5) The last column in Table III indicates that the FEM had a great impact on the inference time of the detection model. There were two main reasons for this result. First, the FEM consisted of multibranch convolutions, which were finally integrated through connections. Although these operations improved detection accuracy, they also reduced the parallelization ability of the detection model. Second, additional convolution operations increased both the complexity and computational cost of the detection model, resulting in a longer inference time and affecting the real-time performance of the detector.

1) **Coarse-to-Fine Head Effect:** Both the R3Det model and the proposed method adopt a coarse-to-fine detection strategy, utilizing a feature refinement module to refine the features in the coarse-detection stage, which are then fed to the fine-detection head for further processing. However, the feature refinement schemes of the two methods are different. The R3Det model uses the results of the coarse-detection stage to refine the features through interpolation operations. Meanwhile, the proposed method uses the output features of the coarse-detection stage to predict the offsets of the sampling grid, which are then used to realize deformable convolution with the original features to make features more consistent with ships in the fine-detection stage. Considering that the R3Det model and the proposed method have similar architectures and processing schemes at
the detection head, the impact of these refinement schemes on the model performance were compared. The comparison results are shown in Table IV, where FRM represents the feature refinement module of the R$^3$Det model, and the coarse-to-fine head represents the feature refinement scheme of the proposed method. The results indicate that the feature refinement scheme of the proposed method can achieve higher detection accuracy than the refinement scheme of the R$^3$Det model. This further demonstrates that compared with the refinement scheme based on interpolation, the refinement scheme of the proposed method is more suitable for SAR ship detection.

2) MFEM Effect: Next, the effect of the MFEM was analyzed. First, the contributions of the MFFM and FEM to the ship feature enhancement in SAR images were discussed separately, and then the role of the entire MFEM module was analyzed and compared with the FPN and PAFPN.

MFFM Effect: The detection results of the SSDD+ with and without MFFM in the inshore and offshore scenes are presented in Figs. 7 and 8, respectively. According to the results Figs. 7 and 8, the recall, precision, and AP value of the MFFM tested in the inshore scene were 12.62%, 37.65%, and 15.30% lower than those in the offshore scene, respectively. The main reason was that the ships were affected by complex environments, such as surrounding wharves, buildings, and islands in the inshore scene, resulting in a significant decrease in detection performance.

As shown in Fig. 7, the scores of all evaluation metrics of the MFFM were higher than those without MFFM; particularly, the improvements of 3.10%, 12.28%, and 2.46% were achieved in the recall, precision, and AP value, respectively. These significant improvements in the evaluation metrics show that the multiscale feature fusion can effectively improve the ship detection performance in the inshore scene. Further, as shown in Fig. 8, compared to the detector without MFFM, the precision and AP value of the detector with the MFFM obtained improvements of 2.08% and 0.19%, respectively. In the offshore scene, the sea surface was relatively pure, and the surrounding interference factors were significantly reduced, so the improvement in the detection performance in the offshore scene by the multiscale feature fusion was lower than that in the inshore scene. In addition, the reason for a slight decrease in recall could be the accumulation of uninformative interference during the multiscale feature fusion process. Finally, the MFFM can significantly improve the precision in both inshore and offshore scenes and can also effectively remove false positives in the detection results.

FEM Effect: The detection results on the SSDD+ dataset with and without the FEM in the inshore and offshore scenes are presented in Figs. 9 and 10, respectively. As shown in Figs. 9
Fig. 10. Model performance on SSDD+ with and without FEM in the offshore scene.

TABLE V
INFLUENCE OF REDUCTION ON THE MODEL PERFORMANCE

| Reduction | AP (%) |
|-----------|--------|
| 2         | 94.05  |
| 4         | 94.16  |
| 8         | 94.50  |
| 16        | 94.54  |
| 32        | 93.76  |

Fig. 11. Model performance on SSDD+ with and without SAM in the inshore scene.

Fig. 12. Model performance on SSDD+ with and without SAM in the offshore scene.

and 10, the FEM improved the performance of SAR ship detection in both scenes, which proved the effectiveness of feature enhancement. In the inshore scene, the recall with the FEM was 2.06% higher than that without the FEM, indicating that the FEM significantly reduced the occurrence of false negatives in the inshore scene. In the offshore scene, the precision with the FEM was 1.1% higher than that without the FEM, demonstrating that the FEM could effectively remove the false positives from the detection results in the offshore scene.

For the FEM, the bottleneck structure was applied to the features at each level to reduce the computation cost. The effect of hyperparameter reduction on model performance is shown in Table V. According to the results, increasing reduction could reduce the number of parameters of the FEM, but an excessively large reduction led to insufficient information and inadequate extracted features, which affected the judgment performance of the ship detector and reduced the model performance. In this article, the reduction of 16 was used, which provided the best detection performance in the experiments.

**MFEM Effect:** The MFEM fuses and refines shallow texture features and deep semantic features using the MFFM and FEM, and generates multiscale fusion features that contain sufficient contextual information. As shown in Table III, compared to the detector without the MFEM, the recall, precision, AP value, and F1-score were greatly improved after the introduction of the MFEM. To further demonstrate the effectiveness of the MFEM, the proposed MFEM was compared with the FPN and PAFPN, and the results are presented in Table VI. As shown in Table VI, compared with the FPN, the results of the MFEM were improved in terms of recall, AP value, and F1-score, indicating that the proposed MFEM could construct more comprehensive and informative multiscale fusion features. Compared with the PAFPN, the results of the MFEM were improved in terms of many evaluation metrics, indicating that the modified method that obtains contextual information is more suitable for extracting features of ships with large aspect ratios than the modified method that introduces additional paths, which further proves the effectiveness of MFEM.

**3) SAM Effect:** The detection results on the SSDD+ dataset with and without the SAM in the inshore and offshore scenes are presented in Figs. 11 and 12, respectively. In Figs. 11 and 12, it can be seen that the performance of the detector with the SAM in the inshore scene improved significantly compared to that without the SAM but decreased slightly in the offshore scene. There were two reasons for this phenomenon. First, the SAM paid great attention to the global property and highlighted important regions by exploring relationships between pixels in the whole image. The inshore scene was complex, including not only ships, but also other targets, such as wharves, buildings, and islands. Therefore, ships could be easily disturbed by the surrounding

| Type     | Recall | Precision | AP (%) | F1-score |
|----------|--------|-----------|--------|----------|
| None     | 0.9442 | 0.7517    | 92.91  | 0.8370   |
| FPN      | 0.9500 | 0.8345    | 93.40  | 0.8885   |
| PAFPN    | 0.9519 | 0.8305    | 94.06  | 0.8870   |
| MFEM     | 0.9596 | 0.8317    | 94.54  | 0.8910   |
background. At this time, the SAM could effectively highlight the important regions containing ships while weakening the importance of useless information, such as wharves, buildings, and islands. Therefore, the detection performance could be significantly improved. Meanwhile, the sea surface of the offshore scene was pure, so the features obtained in the previous steps were sufficient to fit the ship. At this time, the SAM might be redundant for the offshore scene, which could result in a slight decrease in the detection performance. Second, the SAM consisted of convolution layers, which increased model complexity. Before adding the SAM block, the model might be in a state of underfitting in the inshore scene. The number of network parameters increased with the introduction of the SAM, which was beneficial to further learning of the model and improved detection performance. However, the detection model performed well in the offshore scene before the introduction of the SAM, achieving the recall, precision, and AP value of 98.11\%, 93.05\%, and 96.90\%, respectively. At this time, adding the SAM block had a slight effect on the detection performance, while the increase in the number of parameters might cause network overfitting, resulting in performance degradation. Therefore, an attention mechanism that adapts to both inshore and offshore scenes should be explored in the future.

Next, the influence of different numbers of SAM blocks on detection performance was investigated. The results are shown in Table VII, where it can be seen that when the number of SAM blocks increased, the detection performance first improved and then reduced. The AP value and F1-score improved by 0.43\% and 2.05\%, respectively, when the number of SAM blocks increased from one to two, illustrating that acquiring richer and denser surrounding information was effective in improving detection performance. However, the detection performance tended to deteriorate when the number of SAM blocks increased from two to three and from three to four. This shows that the detection performance was not directly proportional to the number of SAM blocks. This could be because more SAM blocks introduced more parameters, which could cause network overfitting, thus degrading the detection performance. Based on the above analysis, selecting an appropriate number of SAM blocks can provide the surrounding information on a ship and improve the detection performance effectively.

When designing the SAM block, a 1 × 1 convolution was used for the input features to reduce the computation cost. The effect of the hyperparameter ratio on the model performance was also investigated, and the results are shown in Table VIII. Similar to the effect of the hyperparameter reduction of the FEM, an excessively large ratio led to insufficient information and inadequate extracted features, thus affecting the judgment performance of the ship detector and reducing the model performance. In this study, ratio = 8 was used in the SAM block, which provided the best detection performance in the experiments.

In addition, the SAM module was compared with the other attention modules, such as the nonlocal module [52], the squeeze-and-excitation (SE) module [53], and the convolutional block attention module (CBAM) [54]. The results are shown in Table IX. The nonlocal network is a classical spatial attention method. The experimental results indicated that the precision, AP value, and F1-score were significantly improved after the introduction of the nonlocal attention module. This nonlocal module employed the self-attention mechanism to make a single-pixel feature at any location perceive the influence of features at all other positions to obtain the attention information on the whole image. However, when the feature size was large, the nonlocal attention module was very memory intensive and computationally expensive. It should be noted that the SAR image-based ship detection tasks usually require high-resolution large-scale features, so a nonlocal attention module is not suitable for such tasks. The SE module is a classic channel attention module. The experimental results demonstrated that channel attention had a slight effect on the proposed method’s performance. The CBAM achieved improvements based on the SE attention module. It aggregated spatial information of features through average-pooling and max-pooling operations on the channel attention module. At the same time, a lightweight spatial attention block was introduced to calibrate features in the spatial dimension. However, since the channel attention had a slight effect on the proposed method’s performance, and the spatial attention of the CBAM was achieved using a 7 × 7 convolutions on the aggregated features, the obtained contextual information was limited. Therefore, the improvement in the proposed method’s performance was not obvious after the introduction of the CBAM. The SAM adopts the idea of crisscrossing to generate sparse attention maps, which capture long-distance dependence from all pixels through a loop operation. In addition, the SAM uses two consecutive sparse attention maps to replace a single dense attention map of a nonlocal attention module, which saves computing resources significantly and avoids interference.
from other irrelevant pixels. Therefore, the performance of the proposed method is significantly improved after the introduction of the SAM.

E. Discussion

1) Influence of IoU Threshold in the Coarse-to-Fine Head: As mentioned before, the proposed method adopts a coarse-to-fine detection strategy to detect arbitrary-oriented ships in SAR images accurately. The detection stage consists of the coarse- and fine-detection stages. Therefore, the effect of the IoU threshold used to assign positive and negative samples in the two stages on the model performance was analyzed. In this study, the IoU threshold denoted the foreground threshold, and the background threshold was obtained by the foreground threshold $\sim 0.1$. The results are shown in Table X. In the coarse-detection stage, the IoU threshold was set to 0.5, and in the fine-detection stage, three different threshold values were used, 0.5, 0.6, and 0.7. The results showed that the proposed model’s performance was the worst and all evaluation metrics were minimal when the IoU thresholds of both stages were set to 0.5. With the increase in the IoU threshold in the fine-detection stage, precision, and F1-score increased, while recall and AP value first increased and then decreased. This was because the higher IoU threshold in the fine-detection stage improved the quality of positive samples during model training, which further improved the precision. However, a higher IoU threshold at the beginning of model training led to sparse positive samples, which was not beneficial to training performance. This made the network unable to be trained to the optimal state and resulted in a decrease in the AP value. Consequently, settings with the IoU threshold values of 0.5 and 0.6 have been considered optimal for the coarse- and fine-detection stages, respectively.

2) Backpropagation Effect in MFFM Blocks: As shown by the red arrows in Fig. 1, there are three backpropagation paths in the proposed method. These paths are designed to fuse low-resolution, strong-semantic features with high-resolution, weak-semantic features to obtain features with rich information to improve the detection performance for multiscale ships. In the experiment, different feature backpropagation paths were configured in the proposed method to analyze the MFFM effect on features at different levels, and the results are shown in Table XI. For the purpose of comparison, the first row in Table XI shows the results without the MFFM block. Obviously, compared to the results without the MFFM block, the proposed model’s performance could be significantly improved by applying the MFFM block regardless of which level of features was used. In addition, compared with the $C_3$ and $C_4$ levels, the performance was significantly improved after applying the MFFM block at the $C_2$ level, resulting in an AP value of 95.14%, which was only 0.2% lower than the result of the proposed method. This indicated that the effect of MFFM blocks on low-level features was more pronounced. Furthermore, the precision was generally low when the MFFM block was performed on only a certain level of features. And when the MFFM blocks were applied to $C_2$, $C_3$, and $C_4$, the precision improved significantly with a slight change in the recall value. This shows that the MFFM block can effectively remove false positives when multilevel features influence each other and work together.

3) Influence of FEM Block Location: The FEM blocks have been designed to enhance the fused multiscale features. This section studies the influence of the location of the FEM blocks on the proposed model’s performance. The analysis results are shown in Table XII. Similarly, for comparison purpose, the results without the FEM block are given in the first row of Table XII. Based on the results, the following conclusions were drawn. First, the introduction of FEM blocks could effectively improve the proposed model’s performance. Second, in the case where the FEM block only acted on single-level features, the performance gain was the highest when acting on the $C_5$ level. The recall, AP value, and F1-score reached 96.35%, 95.00%, and 88.99%, respectively. This demonstrated that the enhancement effect of the FEM block was more obvious for the low-level features than for the high-level features. Finally, when the FEM blocks were applied to $C_2$, $C_3$, $C_4$, and $C_5$, the proposed detector achieved excellent performance, with the recall, precision, AP value, and F1-score of 96.54%, 85.57%, 95.34%, and 90.70%, respectively.

4) Influence of SAM Block Location: The SAM blocks are introduced to the proposed method to calibrate enhanced features to highlight ship features and suppress the interference from the surrounding background. In this section, the influence of the SAM block location on the proposed model’s performance

| TABLE X | INFLUENCE OF IOU THRESHOLD OF THE COARSE-TO-FINE HEAD ON THE MODEL PERFORMANCE |
|---------|-----------------------------------|
| IoU Threshold | Recall | Precision | AP (%) | F1-score |
| 0.5, 0.5 | 0.9500 | 0.7671 | 93.83 | 0.8488 |
| 0.5, 0.6 | **0.9654** | **0.8552** | **95.34** | **0.9070** |
| 0.5, 0.7 | 0.9577 | 0.8799 | 94.38 | 0.9171 |

| TABLE XI | INFLUENCE OF DIFFERENT FEATURE BACK-PROPAGATION PATHS IN MFFM ON THE MODEL PERFORMANCE |
|----------|-----------------------------------|
| C_2 C_3 C_4 | Recall | Precision | AP (%) | F1-score |
| ☑ ☑ ☑ | 0.9519 | 0.7477 | 94.02 | 0.8375 |
| ☑ ☑ ☑ | 0.9635 | 0.7602 | 95.14 | 0.8499 |
| ☑ ☑ ☑ | 0.9635 | 0.7601 | 95.02 | 0.8535 |
| ☑ ☑ ☑ | 0.9577 | 0.7650 | 94.38 | 0.8506 |
| ☑ ☑ ☑ | **0.9654** | **0.8552** | **95.34** | **0.9070** |

| TABLE XII | INFLUENCE OF THE LOCATION OF FEM BLOCK ON THE MODEL PERFORMANCE |
|-----------|-----------------------------------|
| C_2 C_3 C_4 C_5 | Recall | Precision | AP (%) | F1-score |
| ☑ ☑ ☑ ☑ | 0.9538 | 0.8105 | 93.95 | 0.8763 |
| ☑ ☑ ☑ ☑ | 0.9635 | 0.8267 | 95.00 | 0.8899 |
| ☑ ☑ ☑ ☑ | 0.9777 | 0.8286 | 95.02 | 0.8535 |
| ☑ ☑ ☑ ☑ | 0.9577 | 0.7650 | 94.59 | 0.8885 |
| ☑ ☑ ☑ ☑ | 0.9558 | 0.8283 | 94.01 | 0.8875 |
| ☑ ☑ ☑ ☑ | **0.9654** | **0.8552** | **95.34** | **0.9070** |
is analyzed. The results are shown in Table XIII. Similarly as before, the detection results without SAM blocks are given in the first row in Table XIII. According to the results, the following conclusions were drawn: the introduction of SAM blocks could improve the proposed model’s performance, and when the SAM block acted only on a single-level feature, the proposed model’s performance was improved. The largest improvement was when it acted on the $C_3$ level, achieving the recall, precision, and F1-score of 95.96%, 84.01%, and 89.58%, respectively. Finally, the performance improvement of the proposed detector was the most obvious when SAM blocks were applied to the features at the $C_2$, $C_3$, $C_4$, and $C_5$ levels.

To demonstrate the effect of SAM blocks on the SAR ship detection result more intuitively, the heatmap comparisons of fine-grained features in different scenes were conducted, and the comparison results are presented in Figs. 13 and 14. As shown in Fig. 13(a), the area in a SAR image that contained a ship (orange box) was the desired area, and most of the remaining areas were background, which was of interest, especially the yellow oval, which contained noisy targets that could be easily misidentified as ships. The heatmap analysis of the features after the introduction of SAM blocks is presented in Fig. 13(b)–(e). Based on the results, the following conclusions can be drawn. First, the ship features were effectively highlighted, and the background noise features were significantly suppressed. This showed that SAM blocks could effectively highlight the target area of an image and reduce the attention on the background by understanding the semantics of both foreground and background from the whole image. Second, in the inshore scene, the higher the level was, the stronger the highlight effect on ships was, and the stronger the suppression effect on backgrounds was. However, the effect of SAM blocks on the offshore scene was not as prominent as that for the inshore scene. As shown in Fig. 14, applying SAM blocks on mid-level features (such as the $C_3$ level) provided the best results. Although $C_2$ and $C_4$ levels also highlighted the target area, the contrast between the background and the target on the $C_2$ level was not as obvious as that on the $C_3$ level, and the highlighted area on the $C_4$ level exceeded the target area. The effect of highlighting target areas decreased significantly at the $C_5$ level. This could be because the sea surface in the offshore scene was relatively pure, and the high-level features obtained by the FEM in the previous step could suit the ships well, while the middle- and low-level features were insufficient. This also explains the results in Table XIII. The performance improvement was the most obvious when the SAM block acted on the $C_3$ level, which was due to the large proportion of offshore images in the dataset.

5) Influence of Feature Refinement Module: As explained before, the feature refinement module connects the coarse- and fine-detection stages in the coarse-to-fine head. The main idea is to predict offsets of the feature sampling grid and then use deformable convolution to refine features. When designing the feature refinement module, different structures were adopted, as shown in Fig. 15. The detection results of different feature refinement modules are presented in Table XIV, where it can be seen that the detector using the feature refinement module designed in this article achieved the best detection performance.

F. Comparison With CNN-Based Rotation Object Detectors

To verify the effectiveness of the proposed method, multiple datasets were used to compare the proposed method with other CNN-based rotation object detectors.
Fig. 15. Architecture of different types of feature refinement module. (a) Type A. (b) Type B. (c) Type C. (d) Type D.

### TABLE XIV
DETECTION RESULTS OF DIFFERENT TYPES OF THE FEATURE REFINEMENT MODULE

| Type    | Recall | Precision | AP (%) | F1-score |
|---------|--------|-----------|--------|----------|
| Type A  | 0.9404 | 0.8246    | 92.14  | 0.8787   |
| Type B  | 0.9308 | 0.8387    | 93.19  | 0.8823   |
| Type C  | 0.9327 | 0.8017    | 91.74  | 0.8622   |
| Type D  | 0.9385 | 0.8576    | 92.02  | 0.8962   |
| The proposed method | **0.9654** | **0.8557** | **95.34** | **0.9070** |

1) **Quantitative Analysis:** The results of different methods on the SSDD+ dataset are presented in Table XV, where R represents the rotated anchor, H denotes the horizontal anchor, and SH is the similar-to-horizontal anchor. The results indicated that the proposed method achieved a competitive advantage in terms of accuracy compared with the rotated box variants of classical detectors, such as RetinaNet and RRCN detectors. However, the proposed method increased the inference time compared to the RetinaNet detector. The main reason for this was that the coarse-to-fine head included the fine detection subnetwork and introduced extra convolution operations. Still, the inference time of the proposed method was significantly shorter than that of the RRCN detector. Compared with the Faster RCNN model, which combines the horizontal and rotated anchors, the proposed method had higher accuracy, but slightly lower speed due to the introduction of the MFEM, SAM, and coarse-to-fine head network. Compared with the R-RetinaNet, DRBox-V2, MSR2N, and R^2FA-Det models, the proposed method achieved
significant improvements in accuracy, having the AP value improvements of 0.68%, 2.53%, 4.47%, and 0.62%, respectively. This could be because the proposed method extracted ship features more fully and avoided the problems caused by completely depending on rotated anchors. Furthermore, the proposed method improved the detection speed while maintaining high accuracy, which demonstrates its feasibility for real-time applications.

To evaluate the performance of the proposed method on the high-resolution SAR image dataset, the experiments were conducted on the HRSID dataset. The proposed method was compared with the RetinaNet, R-RetinaNet, Faster RCNN, and R^2FA-Det models. The comparison results are presented in Table XVI and Fig. 16, where the precision-recall (PR) curves of different methods on the HRSID dataset are plotted. Due to the unique anchor design and the feature enhancement strategy, the proposed method achieved the best performance among all models, having the AP value and F1-score of 89.18% and 84.78%, respectively.

### Table XVI

| Method          | Recall | Precision | AP (%) | F1-score |
|-----------------|--------|-----------|--------|----------|
| RetinaNet       | 0.8576 | 0.7243    | 81.01  | 0.7853   |
| Faster RCNN     | 0.8453 | 0.7278    | 82.34  | 0.7822   |
| R-RetinaNet     | 0.8750 | 0.8188    | 85.60  | 0.8459   |
| R^2FA-Det       | 0.8405 | 0.7644    | 82.41  | 0.8006   |
| The proposed method | **0.9057** | **0.7920** | **89.18** | **0.8478** |

Fig. 16. Precision-recall (PR) curves of different methods on HRSID.

Based on the experimental results, the following conclusions were drawn. First, all methods could obtain good results except for a few false negatives for SAR images for the pure sea surface and moderate-sized ship. Second, the proposed method had higher precision and recall for small-sized ships than the other models. This could be because the other methods used only low-resolution and insufficiently informative features so that small-sized ships could not be matched with suitable features correctly. Third, for inshore SAR images with large-sized ships, the proposed method had no false negatives, and the ship position result was relatively accurate. Further, for SAR images with complex scenes and densely arranged ships, the proposed method used enhanced features to avoid false negatives caused by densely arranged ships. At the same time, the attention mechanism was employed to make the model less susceptible to interference from the surrounding background, and misjudging the background as a ship was avoided to a certain extent. Therefore, the proposed method achieved better performance than the other methods.

However, the proposed method had certain limitations in detecting extremely complex scenes. For instance, for the scene where ships were mixed with man-made facilities with strong scattering [see Fig. 18(c)], although the proposed method could significantly reduce false negatives compared to the other methods, there were still many false positives, indicating that the antiinterference ability of the proposed method needed further improvement. In addition, for SAR images with unclear ship targets and low contrast [see Fig. 18(d)], the proposed method could easily generate many false negatives. There might be due to two main reasons: the proposed method has limited ability to represent the features of such images, and the number of such images in the dataset is small, so it is difficult for the network model to learn the features of such images. Therefore, in the future, ship detection in SAR images with low contrast and extremely unclear ship targets could be further studied.

3) Detection Results on Large-Scale SAR Image: To analyze the detection performance of the proposed method on a large-scale SAR image, the model trained on the HRSID dataset was used to perform the experiment on a panoramic ALOS-2 SAR image with multiple inshore and offshore ships [51]. Detailed parameters of this image are shown in Table XVII.

| Parameter            | Value                  |
|----------------------|------------------------|
| Position             | Tokyo Bay              |
| Waveband             | L                     |
| Resolution (m)       | 3                     |
| Polarization         | HH                    |
| Time                 | 2014-4-10              |
| Size of the image (pixels) | 10389 × 6487 |

The detection results on the panoramic ALOS-2 SAR image are presented in Fig. 19, where it can be seen that the proposed...
method had excellent performance in ship detection in the offshore scene under the conditions of pure sea background and moderate-sized ships; there were no false negatives and false positives. For the complex inshore scene, there were a few false negatives and false positives, which could be because the man-made facilities in the port had similar characteristics to ships, making it difficult for detectors to distinguish ships from the surrounding background, which resulted in the increase in the number of false negatives and false positives. This result shows that the proposed method has certain limitations, so the feature enhancement and anti-interference ability need to be improved further, which will be part of our future work.
Fig. 18. Comparison of some detection results of different models on HRSID. From top to bottom: Ground-truth, RetinaNet, Faster RCNN, R-RetinaNet, R²FA-Det and the proposed method. Note that the purple boxes represent true positives, the yellow boxes represent false positives, and red boxes represent false negatives.
IV. CONCLUSION

In this article, a single-stage detection method based on multiscale feature fusion and calibration named the SaDet is proposed. The SaDet can detect arbitrary-oriented ships in SAR images with high accuracy and speed. The SaDet is mainly composed of three important components: the coarse-to-fine head, the MFEM, and the SAM. The proposed method is verified by experiments and compared with other methods. The experimental results show that it is feasible to apply the coarse-to-fine head to ship detection in SAR images, which can effectively improve the detection accuracy and network convergence ability. The MFEM used in the proposed method can effectively enhance multiscale features and significantly improve the model performance. Further, the SAM can calibrate the enhanced features to highlight the ship features, while suppressing
interference from the surrounding background. Compared with the CNN-based rotation object detectors, the proposed method has better detection results, achieving state-of-the-art detection performance.

REFERENCES

[1] W. Yao, M. Huang, Y. Zhang, Y. Xu, X. Liu, and Y. Xiang, “Boosting ship detection in SAR images with complementary pretraining techniques,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 8941–8954, Sep. 2021.

[2] P. Jervolino and R. Guida, “A novel ship detector based on the generalized-lifetime detection ratio test for SAR imagery,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 10, no. 8, pp. 3616–3630, Aug. 2017.

[3] J. Li, C. Qu, and J. Shao, “Ship detection in SAR images based on an improved faster R-CNN,” in Proc. IAPR Conf. Comput. Vis., 2017, pp. 1–6.

[4] J. Jiao et al., “A densely connected end-to-end neural network for multiscale and multisene SAR ship detection,” IEEE Access, vol. 6, pp. 20881–20892, 2018.

[5] S. Wei et al., “Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet,” Remote Sens., vol. 12, no. 1, 2020, Art. no. 167.

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

[7] Y. Zhang, C. Wang, H. Zhang, Y. Dong, and S. Wei, “Automatic ship detection by RetinaNet using multiscale Gaofen-3 imagery,” Remote Sens., vol. 11, no. 5, 2019, Art. no. 531.

[8] X. Fu and Z. Wang, “Fast ship detection method for SAR images in the inshore region,” in Proc. IEEE Int. Geosci. Remote. Sens. Symp., 2021, pp. 3569–3572.

[9] Y.-L. Chang, A. Anagaw, L. Chang, Y.-C. Hsiao, and W.-H. Lee, “Ship detection based on YOLOv2 for SAR imagery,” Remote Sens., vol. 11, no. 7, 2019, Art. no. 786.

[10] J. Xu, C. Pang, Y. Guo, and Z. Shu, “Combinational fusion and global attention of the single-shot method for synthetic aperture radar ship detection,” Remote Sens., vol. 13, no. 23, 2021, Art. no. 4781.

[11] X. Zhang, Y. Zhang, N. Wang, and X. Gao, “A robust one-stage detector for multiscale ship detection with complex background in massive SAR images,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 521712.

[12] Z. Huang, Y. Yan, C. Zhao, and S. Zhou, “Detect larger at once: R2FDet-RS: A region-based convolutional neural network with multilayer fusion for SAR ship detection,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, pp. 1137–1149.

[13] K. He, X. Zhang, and J. Zhan, “R2FDet: Delving into high-quality rotatable boxes for ship detection in SAR images,” Remote Sens., vol. 12, no. 12, 2020, Art. no. 2031.

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

[15] M. Tello, R. Yang, and Z. Zhang, “MSR2N: Multi-stage rotation-based network for arbitrary-orientation ship detection in SAR images,” Sensors, vol. 20, no. 8, 2020, Art. no. 2340.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for visual recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1–9.

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

[18] J. Redmon and A. Farhadi, “YOLO9000: Better, faster, stronger,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6517–6525.

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

[20] W. Liu et al., “SSD: Single shot MultiBox detector,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.

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

[22] M. Kang, X. Leng, Z. Lin, and K. Ji, “A modified faster R-CNN based on CFAR algorithm for SAR ship detection,” in Proc. Int. Workshop Remote Sens. Intell. Process., 2017, pp. 1–4.
[47] 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, 2022, Art. no. 5218018.

[48] 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, Jan. 2022.

[49] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-cross attention for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., Feb. 2019, pp. 603–612.

[50] X. Yang, J. Yan, Z. Feng, and T. He, “R3Det: Refined single-stage detector with feature refinement for rotating object,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 3163–3171.

[51] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, “HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation,” IEEE Access, vol. 8, pp. 120234–120254, 2020.

[52] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7794–7803.

[53] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7132–7141.

[54] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 3–19.

Shuang Zhao received the B.S. degree in communication engineering from the Harbin Engineering University, Harbin, China, in 2018. She is currently working toward the Ph.D. degree in computer vision and synthetic aperture radar (SAR) image processing, especially on SAR object detection and recognition.

Qi Liu received the B.S. degree in rail traffic signal and control from the Beijing Jiaotong University, Beijing, China, in 2018. She is currently working toward the Ph.D. degree in computer vision and information system with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, and the University of Chinese Academy of Sciences, Beijing, China.

Her research interests include computer vision and synthetic aperture radar (SAR) image processing, especially on SAR object recognition.

Weidong Yu (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1994 and 1997, respectively.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include spaceborne SAR system design and SAR signal processing.

Jiyu Lv received the B.E. and M.S. degrees in control theory and engineering from the China University of Petroleum (East China), Qingdao, China, in 1997 and 2003, respectively, and the Ph.D. degree in electromagnetic field and microwave technology from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2006.

She is currently a Research Fellow with the Aerospace Information Research Institute, Chinese Academy of Sciences. Her research interests include spaceborne synthetic aperture radar (SAR) system design and SAR image processing.