Scattering-Keypoint-Guided Network for Oriented Ship Detection in High-Resolution and Large-Scale SAR Images

Kun Fu, Jiamei Fu, Zhirui Wang, and Xian Sun

Abstract—**Ship detection in synthetic aperture radar (SAR) images is a significant and challenging task. Recently, deep convolutional neural networks have been applied to solve the detection problem and made a great breakthrough. Previous works mostly rely on the manually designed anchor boxes to search for the region of interests, which is less flexible and suffers from a heavy computational load. Moreover, these detectors have limited performance in large-scale and complex scenes due to the strong interference of inshore background and the variability of object imaging characteristics. In this article, a novel ship detection method based on the scattering-keypoint-guided network is proposed to remedy these problems. First, an anchor-free network is built to eliminate the effect of anchor boxes, in which a more robust representation scheme is designed for the arbitrary oriented objects. Second, a context-aware feature selection module is introduced to dynamically learn both local and context features. In this process, the semantic information of objects can be enhanced while suppressing the background interference. Third, according to the SAR imaging mechanism, a set of scattering keypoints is defined to describe the local scattering regions and reflect the discriminative structural characteristics of ships. Based on this conception, a novel feature adaption method is proposed with the purpose of dealing with the imaging variability issue. Furthermore, to demonstrate the effectiveness of the proposed improvements, we build the Gaofen-3 ship detection dataset. Meanwhile, the public SAR ship detection dataset is introduced to verify the robustness and generalization ability of the detector. Experimental results on these two datasets show that the proposed method achieves the state-of-the-art performance.**

Index Terms—**Context-aware feature selection (CFS), convolutional neural network (CNN), oriented ship detection, scattering keypoints, synthetic aperture radar (SAR).**

I. Introduction

As an active microwave imaging sensor, synthetic aperture radar (SAR) has all-day operation capability and is independent of any weather conditions. This makes it a prime imaging modality in various application fields [1]–[4]. In recent years, with the rapid development of SAR imaging techniques, automatic SAR object detection [5], [6] has drawn a lot of attention. Ship detection in high-resolution and large-scale SAR images is one of the major tasks, which plays a critical role in maritime management and surveillance.

Traditional ship detection approaches typically consist of three main stages: preprocessing, prescreening, and discrimination [7]. Preprocessing steps including speckle filtering and sea–land segmentation, are initially performed to suppress the clutter and eliminate the interference of land area. In the prescreening stage, certain algorithms are applied to search for the region of interests, in which the local clutter background may be extracted as the candidate parts. Then, the discrimination phase aims to distinguish the true objects from false alarms. For the second stage, commonly used techniques include contrast-based methods and texture-based methods. Contrast-based methods take advantage of the fact that the radar cross section of an object is averagely higher than that of the sea clutter. Constant false alarm rate (CFAR) [8]–[10] is one of the most representative algorithms and has been extensively studied. Mainly focusing on the modeling of image background, numerous research studies explore different statistical distributions to fit the heterogeneous sea clutter, such as alpha-stable distribution [11], compound Gaussian [12], and generalized Gamma distribution [13]. In addition, many variants of the CFAR algorithm, including cell-averaging CFAR [14], greatest-of CFAR [15], smallest-of CFAR [16], and ordered-statistic CFAR [17], are proposed to deal with the nonuniform clutter and high density of objects. These methods usually perform well in simple scenes while cannot robustly deal with the complex sea states. Texture-based approaches essentially explore the distinct texture properties to distinguish between the objects and the clutter. Tello *et al.* [18] analyze the local region with the discrete wavelet transform and compute the spatial correlation to reduce the background noise. Gierull [19] adopts a kind of textured background model to facilitate the detection of small ships in challenging marine environments. Texture-based detectors have higher robustness, whereas the inshore object detection performance is still inferior. In general, traditional methods usually achieve good results

Manuscript received July 19, 2021; revised August 19, 2021; accepted August 29, 2021. Date of publication September 2, 2021; date of current version November 12, 2021. This work was supported by the National Science Foundation of China under Grant 61725105 and Grant 62076241. *(Corresponding author: Kun Fu.)*

Kun Fu, Jiamei Fu, and Xian Sun are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, with the Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China, with the University of Chinese Academy of Sciences, Beijing 100190, China, and also with School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: fukun@aircas.ac.cn; fujiamei14@mails.ucas.ac.cn; sunxian@aircas.ac.cn).

Zhirui Wang is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhirui1990@126.com).

Digital Object Identifier 10.1109/JSTARS.2021.3109469

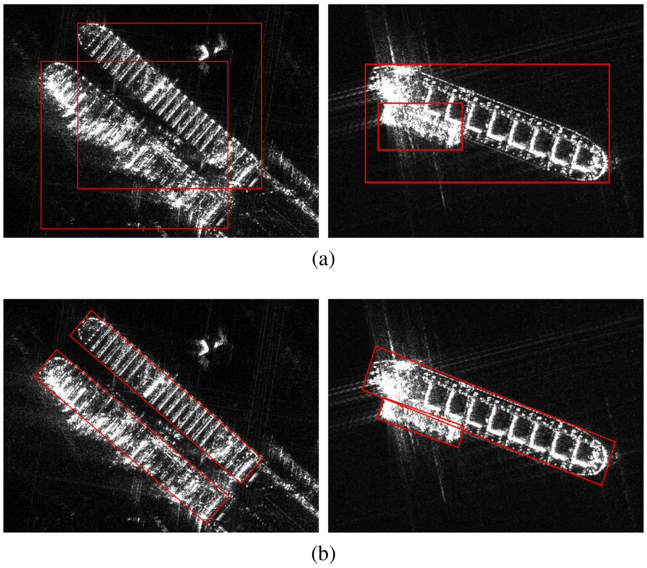


Fig. 1. Two kinds of representations of the bounding box. (a) Horizontal bounding box. (b) Oriented bounding box.

in simple and specific scenes, but they heavily rely on the hand-crafted features and regulations, which limits their generalization ability in complex scenes.

Recently, driven by the success of deep learning methods for general object detection [20]–[22], convolutional neural networks (CNNs) have been introduced to solve the detection problem in remote sensing images [23]–[25]. Benefitting from the powerful feature extraction ability, deep CNN-based methods have made remarkable breakthroughs in SAR object detection and gradually become the substitute of the conventional algorithms in practice. Most of the SAR ship detectors [26]–[28] adopt the same form as that in natural scenes, i.e., the horizontal bounding box (HBB), to describe the object location, as shown in Fig. 1(a). However, this kind of bounding box is not an optimal way to represent the arbitrary oriented ship with large aspect ratio, as it only gives a coarse localization, failing to indicate the accurate orientation and scale information. Furthermore, the HBB representation contains much interference of background or nearby objects. This leads to the misalignment for the subsequent tasks, such as the object recognition.

In order to solve these problems, some studies [29], [30] focus on the oriented bounding box (OBB)-based detector that provides a finer localization for the object in high-resolution SAR images, as shown in Fig. 1(b). Although these approaches have been proved to be effective, there are still several limitations for the application in large-scale and complex scenes. First, the existing detectors rely on the manually designed anchor boxes to provide the references for the bounding box classification and regression. In order to maintain a high recall, anchor boxes with certain angles, scales, and aspect ratios are usually densely sampled over the input image. Nevertheless, the average distribution of the objects in large-scale SAR image is sparse, and redundant anchor boxes cause a heavy computational load. Moreover, these methods suffer from the problem of corner case caused by the discontinuous representation, i.e., the related parameters of the OBB usually have a considerable variation when the box is nearly horizontal, increasing the difficulty of the regression. Second, some look-alikes (e.g., wharf area, reef, ambiguities, and certain inland facilities) have similar visual attributes with ships. These irrelevant objects are likely to be misidentified, affecting the reliability of ship detection and resulting in a considerable high false alarm rate. Considering that the surrounding environment can help to rule out the false alarms, previous works [31]–[33] design a specific module to capture the context information. But these methods merely acquire the surrounding context following a stationary paradigm and fail to fully utilize both the local and context information. Third, the imaging results for one object under various conditions may have quite different characteristics. This will lead to the result that the detector cannot accurately locate the same kind of object when the SAR imaging condition changes. We argue that the visual features learned by the CNNs are incapable of adapting to the significant scattering change. Thus, it is necessary to delve into the scattering mechanism to guide the adaptive feature learning. Some studies [1], [34] analyze the detection problems and object characteristics on the basis of SAR imaging principles. However, the scattering information is only exploited in the preprocessing stage, and it is still a challenge for the network to adaptively learn the discriminative features for the objects in different conditions.

Based on the above considerations, a scattering-keypoint-guided network (SKG-Net) is proposed for the arbitrary oriented ship detection in large-scale SAR images. Specifically, the network is built on an anchor-free detection pipeline. The bounding box is encoded by a set of parameters that depict the boundary shifts from the center point. To deal with the corner case, the detector guides the selection of the OBB or HBB regression result based on an additional rotation factor of the detection box. Then, our method explores the context information and the SAR imaging characteristics in a novel direction, achieving more robust performances in complex scenes. The proposed context-aware feature selection (CFS) module utilizes two kinds of extractors to acquire local features and the surrounding context, respectively, and combine them via a dynamic selection mechanism. In addition, a conception based on the scattering keypoint is introduced. The keypoints obtained by the network are used in conjunction with the convolution serving as the refined sampling points. This operation aims to extract the salient and aligned features for the objects, making the model better adapt to different imaging conditions.

The main contributions in this article can be summarized as follows.

1) Different from most anchor-based methods, we propose an anchor-free detection network, in which the bounding box is described in a robust representation scheme for the arbitrary oriented object. By eliminating the effect of anchor boxes, our method is flexible and computation efficient.
2) The CFS module is introduced to learn both the local and context features in a dynamic fashion. In the process of

multiscale feature fusion, it enhances the semantic information of the objects while suppressing the background interference. The false alarm rate can be significantly reduced.

3) Taking the SAR imaging principles into account, a set of scattering keypoint is defined to depict the local scattering regions with distinct structural characteristics of ships. Based on this conception, a scattering-keypoint-guided feature adaption method is proposed, which increases the model adaptability in different imaging conditions and advances the localization accuracy to a certain degree.
4) The Gaofen-3 ship detection dataset (GF3SDD) is constructed, which provides a strong support for the research of oriented SAR ship detection in large-scale and complex scenes. Several existing methods as well as the proposed network are performed on this dataset, severing as the benchmark for the following work. The GF3SDD will be released to boost the research of SAR ship detector in the future.

Extensive studies on GF3SDD and SAR ship detection dataset (SSDD) are conducted to validate the effectiveness of the proposed improvements. Our method outperforms the recent competitive detectors and shows a desirable performance on oriented SAR ship detection.

The rest of this article is organized as follows. Section II briefly reviews some related works. Section III illustrates the details of the proposed method. In Section IV, the experimental results and analysis are given. Finally, Section V concludes this article.

II. Related Work

A. CNN-Based General Object Detector

Deep CNNs have been successfully applied in the field of computer vision and become the mainstream in general object detection task. Modern CNN-based detectors are typically divided into anchor-based and anchor-free methods.

In anchor-based methods, a set of anchor boxes with different scales and aspect ratios is manually predefined. They serve as reference boxes for region proposals, with the aim to search the possible regions containing objects in a sliding window style. Inspired by the proposal-based detector [35], Faster R-CNN [20] initially introduces the anchor boxes and replaces the traditional selective search method with the region proposal network (RPN), making the whole pipeline end-to-end trainable. Later, on the basis of Faster R-CNN, a variety of improved approaches have been proposed. In order to reduce the computational load, R-FCN [36] adopts a regionwise fully convolutional subnetwork as the substitute of fully connected layers after the region proposal. Cascade R-CNN [37] builds a multistage architecture and refines the prediction results by training a sequence of detectors with certain intersection-over-union (IoU) thresholds. Different from these methods above, some frameworks, including SSD [21], RetinaNet [22], and RefineDet [38], are based on a single feedforward network to predict the bounding boxes without region proposal generation and refinement. These unified detectors can achieve high computation efficiency but at the cost of low detection accuracy.

Compared with anchor-based detectors, anchor-free methods are independent of the predefined anchor boxes. Generally, the extra hyperparameters of anchor settings need to be carefully tuned, as they have a great impact on the detection performance. By eliminating the large set of anchor boxes, anchor-free detectors can avoid related heuristic tuning and complex computation. In early works, DenseBox [39] and UnitBox [40] directly generate the results in a point-to-box prediction manner. However, these methods suffer from a relatively low recall. Recently, motivated by other vision tasks like semantic segmentation and keypoint detection, anchor-free methods address the detection problem with alternate strategies and achieve competitive performance with the anchor-based counterparts. For instance, FCOS [41] and FoveaBox [42] encode the bounding box as an inner point with its normalized distances to four boundaries. CornerNet [43] and CenterNet [44] detect several keypoints of the object and then match these keypoints to form a bounding box. On the whole, anchor-free methods usually have a simpler pipeline and show a better tradeoff between accuracy and efficiency.

B. Deep Learning Methods for SAR Object Detection

Due to the powerful learning ability, deep neural networks are utilized for the automatic feature extraction in SAR object detectors and have achieved superior performance in comparison with the traditional algorithms. In the early stage, CNNs are employed in certain parts of the conventional detectors. For example, the fully convolutional network is used for the sea–land segmentation [45], and the RPN is applied to guide the CFAR algorithm [46]. Later, with the increase of available SAR data, various end-to-end methods based on the CNNs are proposed. In order to deal with the complex scenes, the attention mechanism [28], [47] and saliency information [5] are adopted to highlight the salient regions containing objects while suppressing the background interference. Considering the semantic inconsistency across different feature levels, Jiao *et al.* [26] fuse the multiscale features with dense connections, and Cui *et al.* [33] integrate the convolutional block attention module into the pyramid network, adaptively enhancing the significant features of specific scales. Some studies analyze the detection problems from the perspective of imaging mechanism. Fu *et al.* [34] propose a refinement module to address the feature misalignment caused by the surrounding interference with similar scattering phenomenon. Guo *et al.* [1] design a scattering enhancement strategy for the input image to handle the discreteness and variability of objects.

The methods mentioned above locate objects with HBBs, which fails to realize the accurate localization of the arbitrary oriented objects with large aspect ratios. The bounding boxes may include much background interference or irrelevant regions from nearby objects. It will cause the object misalignment issue for the subsequent stages in practical applications, such as the classification part in SAR automatic target recognition system. To solve this problem, several detectors with OBBs are

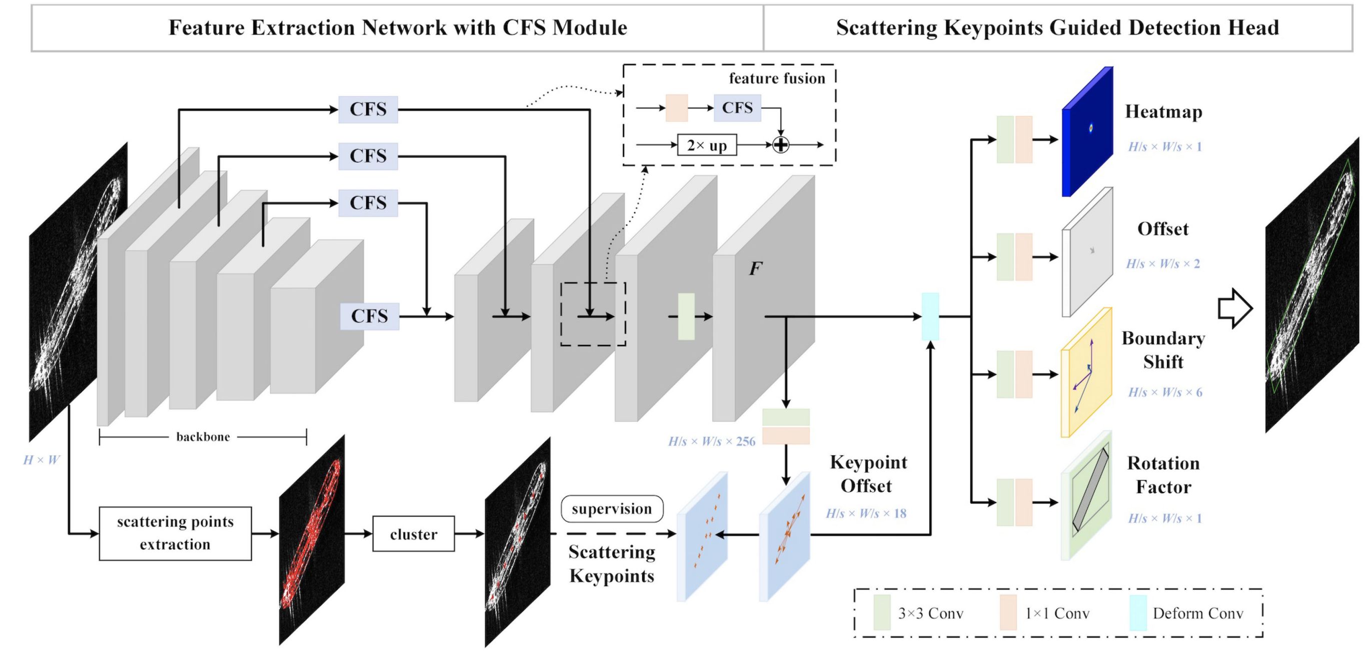


Fig. 2. Overall framework of the proposed detector. It is composed of a feature extraction network with the CFS module and a scattering-keypoint-guided detection head.

built following the idea of general detection models [48], [49]. Compared with traditional methods with separate localization and angle estimation, Wang *et al.* [30] combine these two tasks in an improved SSD detector [21], and Liu *et al.* [29] propose a similar framework using rotatable bounding boxes to detect inshore ships. Based on this architecture, a modified box encoding scheme and an anchor sampling strategy are designed for more accurate orientation prediction in [50]. Following the idea in [37], Pan *et al.* [51] propose a multistage rotational region-based network to optimize the localization results. Yang *et al.* [52] adopt a one-stage model [22] as the basic structure to save the computational cost and introduce a calibration scheme to deal with the feature scale misalignment issue.

Different from the aforementioned detectors, our method adopts the general anchor-free model as the basic framework and extends it to the oriented object detection task. It is more flexible by eliminating the predefined anchor boxes. Furthermore, we exploit the context and scattering information of objects to guide the feature learning in a dynamic fashion, achieving a better detection performance.

III. Proposed Method

The overall framework of the proposed method is shown in Fig. 2. It consists of a feature extraction network with the CFS module and a scattering-keypoint-guided detection head. In this section, we first present the preliminary network structure and explain the representation scheme for the oriented objects. Then, to capture the context information for adaptively learning the discriminative features in the complex scenes, the CFS module is proposed and described in detail. Taking the unique scattering mechanism into account, the conception based on the scattering keypoints is introduced into the detection head with the purpose of guiding the feature adaption, as illustrated in the final part.

A. Preliminary Network Structure

The proposed method is built on a keypoint-based detection pipeline [53] that predicts the center point of the object and directly regresses the size and offset of the bounding box from the extracted features. The feature extraction network adopts a similar encoder–decoder architecture [54]–[56] to generate the feature map for the ensuing prediction part. We use ResNet50 [57] as the backbone. It is divided into five stages with respect to the size of the feature maps in different layers. The outputs of last four stages are denoted as $\{C_2, C_3, C_4, C_5\}$ with the channel dimensions being {256, 512, 1024, 2048}, and their downsampling ratios to the input image are {4, 8, 16, 32}, respectively. In the process of upsampling, the high-level features are fused with the low-level features through skip connections. Specifically, the channel dimension of C_j is first reduced to 256 by applying one 1×1 convolutional layer. In the upsampling pathway, the features with coarser resolution are merged with the corresponding low-level feature maps after being upsampled by a factor of 2 using bilinear interpolation. The merged map with finest resolution is further refined through a 3×3 convolutional layer to obtain the final feature map $F \in \mathbb{R}^{(H/s)\times(W/s)\times 256}$, where $H \times W$ is the size of input image and $s = 4$ is the output stride. Then, four parallel branches are attached to F in the detection head. Each branch has one 3×3 convolutional layer followed by a 1×1 convolutional layer. The detection results can be decoded by the four predicted maps, as shown in Fig. 2.

The predicted heatmap aims to localize the center points of different ships in SAR images. For a ground-truth bounding box B_i, the center point in the input image is defined as

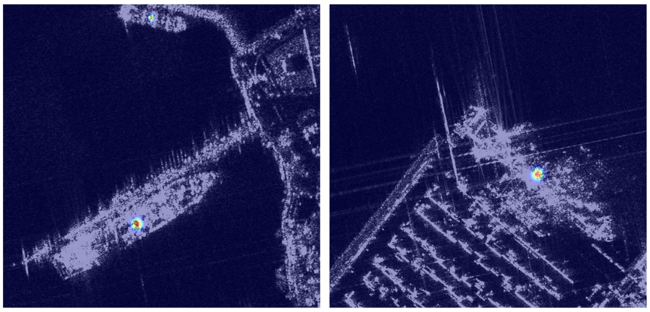


Fig. 3. Visualization of the ground-truth heatmaps.

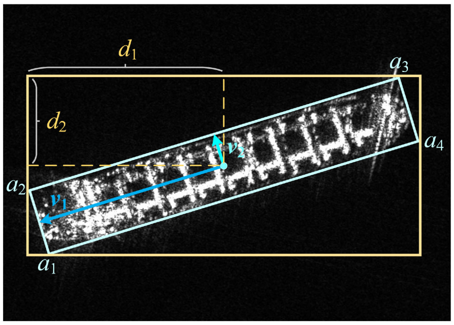


Fig. 4. Representation of the bounding box.

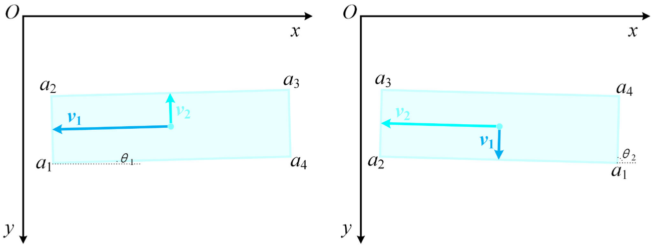


Fig. 5. Illustration of the corner case.

$\tilde{ct}_i = (\tilde{x}_i, \tilde{y}_i)$. It is mapped to the location $ct_i = (x_i, y_i)$ on the heatmap, where $x_i = \lfloor \tilde{x}_i/s \rfloor, y_i = \lfloor \tilde{y}_i/s \rfloor$. The pixel point ct_i is regarded as a positive sample, and the corresponding value on the target heatmap is set to 1. Other points are negative samples with the target values set to 0. Following a similar training strategy as in [43], the penalty for the negative point (x_n, y_n) within a scale-adaptive radius r of the positive location is reduced. The target heatmap values are modified according to a 2-D Gaussian kernel $\exp(-\frac{(x_n-x_i)^2+(y_n-y_i)^2}{2\sigma^2})$, where the kernel center is at the positive point and $\sigma = r/3$, as shown in Fig. 3. Variant focal loss is adopted for the training of the heatmap:

$$\mathcal{L}_{\text{heat}} = -\frac{1}{N}\sum_{x,y} \begin{cases} (1-h_{xy})^{\alpha}\cdot\log(h_{xy}), & \text{if } \hat{h}_{xy}=1\\ (1-\hat{h}_{xy})^{\gamma}\cdot h_{xy}^{\alpha}\cdot\log(1-h_{xy}), & \text{otherwise}\end{cases} \tag{1}$$

where h_{xy} and $\hat{h}_{xy}$ denote the predicted and ground-truth values at location (x, y) on the heatmap, respectively, N is the number of ships, and α and γ are the hyperparameters that adjust the weights of the modulating term ($\alpha = 2$ and $\gamma = 4$ in this article). During the inference, the point on the heatmap cannot be directly mapped back onto the input image because of the precision loss caused by the downsampling stride. To recover this error, the offset map $O \in \mathbb{R}^{(H/s)\times(W/s)\times 2}$ predicts a local offset for each center point. The target offset at ct_i is

$$\hat{\boldsymbol{o}}_i = (\tilde{x}_i/s - x_i, \tilde{y}_i/s - y_i). \tag{2}$$

The training loss is calculated as follows:

$$\mathcal{L}_{\text{off}} = \frac{1}{N}\sum_{i=1}^{N} \text{smooth}_{L_1}(\boldsymbol{o}_i - \hat{\boldsymbol{o}}_i) \tag{3}$$

$$\text{smooth}_{L_1}(t) = \begin{cases} 0.5\cdot t^2, & \text{if } |t| < 1\\ |t| - 0.5, & \text{otherwise}\end{cases}. \tag{4}$$

We introduce a simple representation scheme for the oriented objects. It regresses the box boundary shifts from the center point, as shown in Fig. 4. The OBB and the HBB are predicted simultaneously. For the OBB, the intersection points with the bottom, left, top, and right boundary of the HBB are defined as a_1, a_2, a_3, and a_4, respectively. The OBB can be encoded by two vectors ($\boldsymbol{v}_1$ and $\boldsymbol{v}_2$) depicting the direction and distance of boundary (a_1a_2 and a_2a_3) shifts from the center point. Two additional parameters (d_1 and d_2) are used to denote the size of the HBB. The final representation is

$$\boldsymbol{b} = (\boldsymbol{v}_1, \boldsymbol{v}_2, d_1, d_2) \in \mathbb{R}^6 \tag{5}$$

In practice, it is observed that the problem of corner case occurs when the OBB is approximately horizontal. By the corner case (illustrated in Fig. 5), we mean that the vectors present a considerable variation for the nearly horizontal object and are likely to be in a state of confusion, increasing the difficulty of regression. In this circumstance, the HBB seems to be a better choice for the final output, as it is free from the discontinuous representation. Inspired by Xu *et al.* [58], we predict a rotation factor ρ for each object to select OBB or HBB as the final result, and it is defined as the area ratio between the OBB and the HBB. If ρ is higher than a threshold, then we use the HBB form as the output. The threshold is set as 0.8 in this article. During the training, the predictions of boundary shifts and rotation factor are supervised by the smooth L1 loss (denoted as $\mathcal{L}_{\text{bnd}}$ and $\mathcal{L}_{\text{r}}$) at the center points of ground truth.

B. CFS Module

In SAR images, some irrelevant elements (look-alikes), such as reef and harbor facilities, may present similar visual attributes and scattering mechanism with the ships. They may get incorrectly identified by the detector, affecting the reliability of the detection performance. It is difficult to distinguish these false alarms from the true objects if we only pay attention to the object itself. Generally, the surrounding context can provide more semantic information to understand the object in the complex scene. For instance, the sea clutter has a relatively low scattering intensity, making the ships more salient, while for the inland facilities, the surrounding area usually presents

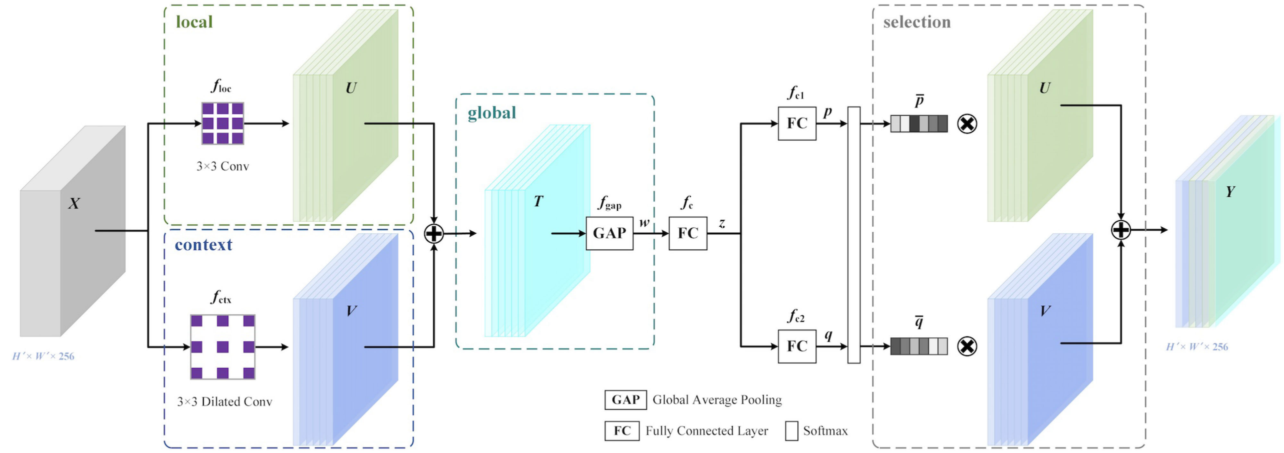


Fig. 6. Structure of the CFS module.

a more complicated scattering phenomenon. One direct way of acquiring the context information is multiscale feature fusion. In deep CNNs, the low-resolution feature maps in high levels have larger receptive fields and more semantic information. By comparison, the low-level features have more spatial but less semantic information. It is necessary to merge the high-level maps with the low-level ones to obtain the high-resolution and semantically strong features. To enrich the context information for the extracted feature F, we improve the feature extraction network with the designed CFS module, as shown in Fig. 2. The CFS module is inspired by the selective kernel mechanism [59] with the aim to adaptively select the discriminative features in the multiscale feature fusion.

The CFS module is depicted in Fig. 6. It contains four main components: the local feature extractor f_{loc}, the surrounding context extractor f_{ctx}, the global information extractor f_{gap}, and the feature selector. To begin with, f_{loc} and f_{ctx} are applied on the input $X \in \mathbb{R}^{H' \times W' \times 256}$ to learn the local feature $U = (U^1, U^2, \ldots, U^{256}) \in \mathbb{R}^{H' \times W' \times 256}$ and the surrounding context $V = (V^1, V^2, \ldots, V^{256}) \in \mathbb{R}^{H' \times W' \times 256}$, respectively. Note that U^j and V^j represents the jth channel dimension of U and V. f_{loc} is composed of a 3×3 convolutional layer, and we adopt one 3×3 dilated convolutional layer with the rate being 3 for f_{ctx}, as it has a larger receptive field and can better capture the context information. Both convolutions are depthwise in order to save the computational cost. Two kinds of features are then fused via the elementwise addition to generate the merged map $T = (T^1, T^2, \ldots, T^{256}) \in \mathbb{R}^{H' \times W' \times 256}$:

$$T = U + V = f_{\text{loc}}(X) + f_{\text{ctx}}(X). \tag{6}$$

Afterward, the global information is extracted from T through the global average pooling f_{gap} along the spatial dimensions. It can be regarded as a channelwise weighted vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_{256}) \in \mathbb{R}^{256}$, and its cth element is computed as follows:

$$w_c = f_{\text{gap}}(T^c) = \frac{1}{H' \cdot W'} \sum_{m=1}^{H'} \sum_{n=1}^{W'} T^c(m, n). \tag{7}$$

This operation is then followed by a variant of multilayer perceptron to guide the adaptive feature selection. Concretely, a fully connected layer f_{c} is initially employed to generate an interfeature $\boldsymbol{z} \in \mathbb{R}^{256/\varepsilon}$, where ε is the reduction ratio of dimension set as 16 by default. Then, two separate fully connected layers, f_{c1} and f_{c2}, are followed to increase the interfeature dimension, and the output vectors are denoted as $\boldsymbol{p} = (p_1, p_2, \ldots, p_{256}) \in \mathbb{R}^{256}$ and $\boldsymbol{q} = (q_1, q_2, \ldots, q_{256}) \in \mathbb{R}^{256}$, respectively. The soft attention vectors ($\bar{\boldsymbol{p}}, \bar{\boldsymbol{q}} \in \mathbb{R}^{256}$) for the feature selection are obtained using softmax operator

$$\bar{p}_c = \frac{e^{p_c}}{e^{p_c} + e^{q_c}}, \quad \bar{q}_c = \frac{e^{q_c}}{e^{p_c} + e^{q_c}} \tag{8}$$

where $\bar{p}_c$ and $\bar{q}_c$ are the scalar weights for the cth channel dimension with respect to the candidate features (U and V). The final output $Y = (Y^1, Y^2, \ldots, Y^{256}) \in \mathbb{R}^{H' \times W' \times 256}$ is generated through a channelwise weighted combination

$$Y^c = \bar{p}_c \cdot U^c + \bar{q}_c \cdot V^c. \tag{9}$$

Here, we also consider some other methods to capture the local and context features, such as using deformable convolutions [60] or extending to more branches. By comparison, the manner shown in Fig. 6 achieves a better balance between accuracy and efficiency.

C. Scattering-Keypoint-Guided Detection Head

According to the SAR imaging principles, one object may present different scattering characteristics under various conditions. Many factors, such as incidence angle, polarization mode, object orientation, and sea condition, have great impact on the imaging results. This brings a huge challenge for the feature extraction in deep-learning-based approaches. It is found that the detector is likely to miss the same object when the imaging condition changes. We argue that the visual features learned by the CNNs are insufficient for the object representation and cannot adaptively express the scattering change. Thus, the information related to the scattering mechanism should be taken full advantage of to guide the adaptive feature learning. However, the scattering information is abstract, and two questions should

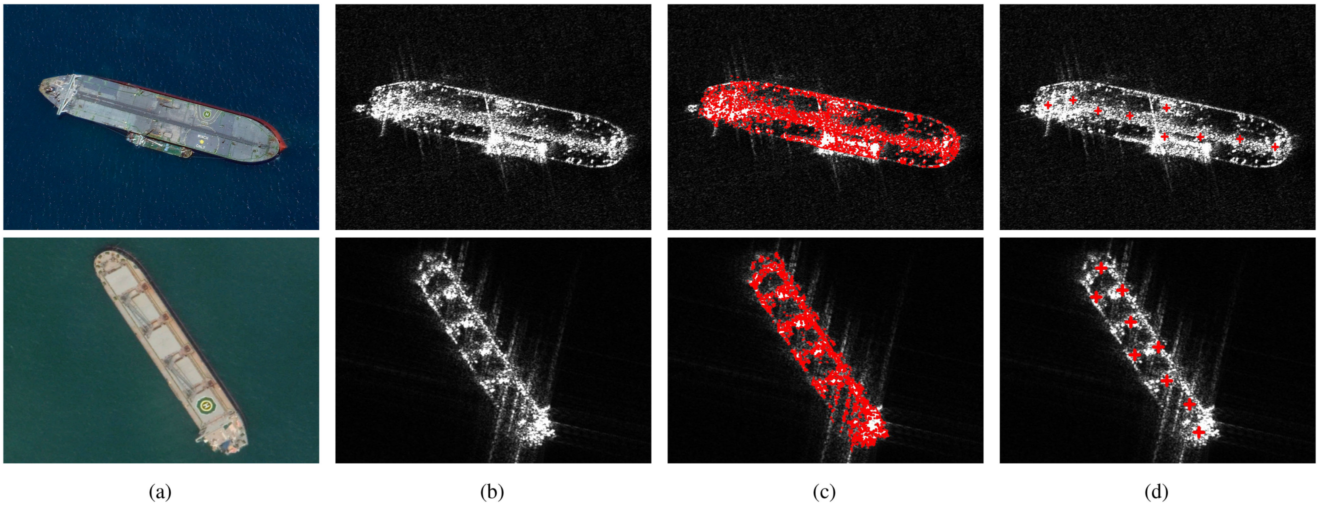


Fig. 7. Illustration of scattering keypoint representation. (a) Optical images of ships. (b) SAR images of ships. (c) Extracted points by Harris corner detector. The red dots represent the extraction results, reflecting the scattering intensity distribution. (d) Cluster results. The red crosses represent the cluster centers, depicting the local scattering region and the structural characteristics of ships.

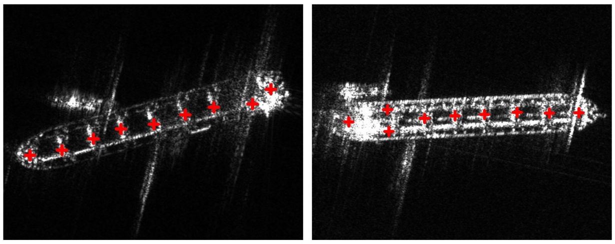


Fig. 8. Extracted scattering keypoints for the same kind of ships with different scattering characteristics.

be considered: 1) How to represent and learn it? 2) How the CNN-based method can deal with it?

In single-polarization SAR images, the distributions of scattering intensity are informative for the object detection. For a given object, we model the local scattering regions as a set of keypoints. Specifically, we extract the points with local peak values through Harris corner detector to reflect its scattering distribution. The extracted points are shown in Fig. 7(c). Considering that these points are redundant and irregular, we partition them into nine clusters (corresponding to the size of 3 × 3 convolution kernel) using K-means. We call the cluster centers shown in Fig. 7(d) as scattering keypoints, which can be treated as representative points, depicting the local scattering regions. It can be seen that these scattering keypoints roughly reflect the structural characteristics and capture the discriminative features of ships. Take the two kinds of ships in Fig. 7 for instance; the salient features for the oil tanker in the first row are the oil pipelines and the superstructure at the rear with strong backscattering, and for the bulk carrier in the second row, the interval between two hatch covers presents a distinct stripe perpendicular to the side of the ship. In addition, as shown in Fig. 8, for the same kind of ships with different imaging results, the scattering keypoints can capture similar structural characteristics. This provides an alternative way for the feature representation, i.e., the scattering feature learning can be formulated as the keypoint localization. Then, the predicted scattering keypoints are used to guide the ensuing feature adaption to deal with the scattering change issue.

Due to the irregular distributions, the scattering keypoints are difficult to handle by the standard convolution. To this end, a more flexible operator is required. Following this intuition, we explore a novel direction of the deformable convolution [60] in combination with the scattering keypoints, with the aim to refine the object features at the center points. The proposed scheme is illustrated in Fig. 2. We use the scattering keypoints to guide the generation of offset field for deformable convolution. Specifically, the keypoint offset map $S \in \mathbb{R}^{(H/s)\times(W/s)\times 18}$ predicts the point offsets from the center point, and it is embedded into the offset field branch of deformable convolution. The predicted keypoints are obtained

$$\phi = \{p^k \mid k = 1, \dots, 9\} = \{p + \Delta p^k \mid k = 1, \dots, 9\} \quad (10)$$

where $\{\Delta p^k \mid k = 1, \dots, 9\}$ refer to the predicted offsets with respect to the location p on S. The target scattering keypoints for the ith object are denoted as $\hat{\phi}_i = \{\hat{p}_i^k \mid k = 1, \dots, 9\}$. Note that $\hat{p}_i^k$ is the point coordinate on input image divided by s. We use the Chamfer loss [61] to supervise the learning of scattering keypoints at the center point of objects

$$\mathcal{L}_{\text{scatter}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{18}\sum_{m=1}^{9} \min_{n} \|p_i^m - \hat{p}_i^n\|_2 + \frac{1}{18}\sum_{n=1}^{9}\min_{m}\|p_i^m - \hat{p}_i^n\|_2\right). \quad (11)$$

Then, the input feature F of detection head is fed into the 3 × 3 deformable convolution $\Phi_{3\times 3}$ with the keypoint offset branch to refine the object features. For the location p on the

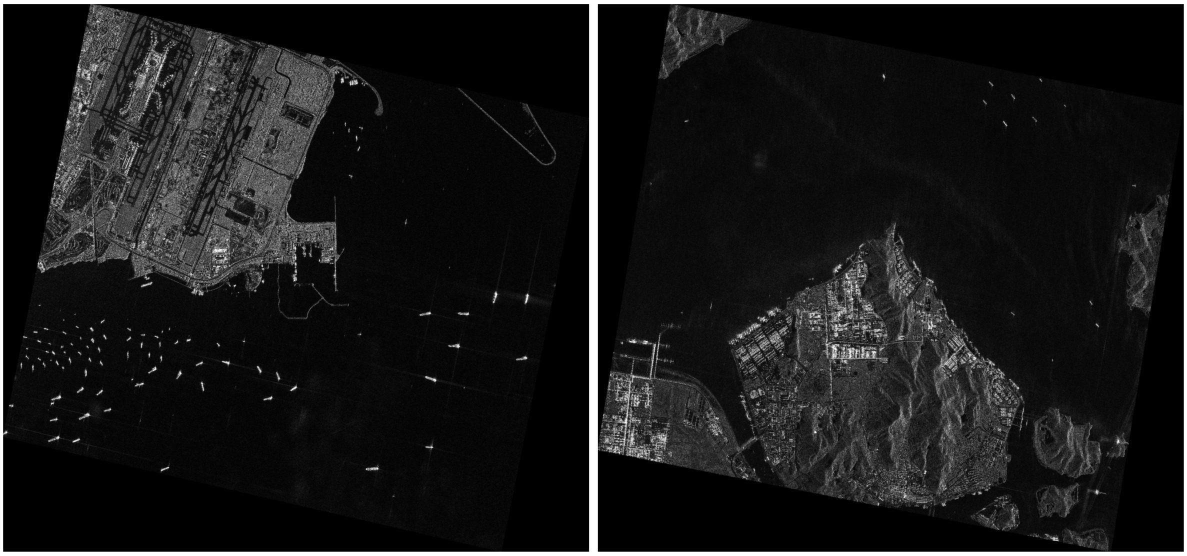


Fig. 9. Image examples in the GF3SDD.

refined feature map $F' \in \mathbb{R}^{(H/s)\times(W/s)\times 256}$:

$$F'(p) = \Phi_{3\times3}(F, \{\Delta p^k\}) = \sum_{k=1}^{9} \omega(\Delta p^k) \cdot F(p + \Delta p^k) \quad (12)$$

where ω is a set of learnable weights, and the feature vector $F(p')$, $p' = p + \Delta p^k = (x_{p'}, y_{p'})$ is computed via the bilinear interpolation

$$F(p') = \sum_{q} \xi(q, p') \cdot F(q) \quad (13)$$

where $\xi(q, p')$ is the bilinear interpolation weight between p' and the integral sampling point $q = (x_q, y_q)$. It is defined as

$$\xi(q, p') = \max(0, 1 - |x_q - x_{p'}|) \cdot \max(0, 1 - |y_q - y_{p'}|). \quad (14)$$

In this process, the object feature is augmented by the information from the scattering keypoints. The output map F' is used for the following bounding box prediction, which is illustrated in Section III-A. Following the multitask learning, the final loss function is

$$\mathcal{L} = \mathcal{L}_{\text{heat}} + \mathcal{L}_{\text{off}} + \mathcal{L}_{\text{bnd}} + \mathcal{L}_{\text{r}} + \lambda\mathcal{L}_{\text{scatter}} \quad (15)$$

where λ is a balancing parameter for $\mathcal{L}_{\text{scatter}}$, and it is set as 1 by default in the experiments.

IV. Experiments

A. Datasets

There are already several datasets available for object detection in SAR images. For example, OpenSARShip [62] is a medium-resolution dataset collected from Sentinel-1. The dataset in [63] provides multiscale ships in small image patches with varying background. AIR-SARShip-1.0 [64] comprises over 30 large-scale images, but the scenarios are relatively simple. Moreover, the datasets mentioned above only provide HBB annotations, and they are less suitable for the task of oriented SAR ship detection in large-scale and complex scenes. To this end, the GF3SDD is constructed to demonstrate the effectiveness of the proposed improvements. In addition, SSDD [65] is utilized to further verify the robustness and generalization ability of the detector.

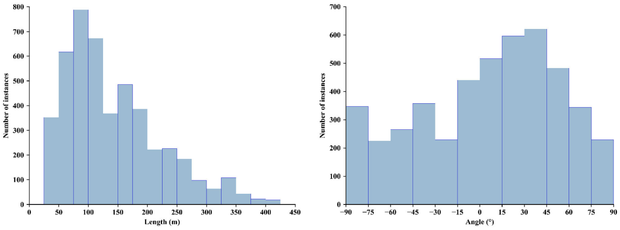


Fig. 10. Distribution of instances with respect to length and orientation.

1) GF3SDD: Gaofen-3 is a C-band civilian spaceborne SAR satellite with the capability of working in multi-imaging and multipolarization modes. In the experiments, a total of 112 single-polarization images with 1-m spatial resolution in spotlight mode are utilized. Some image examples are shown in Fig. 9. Among them, 86 images are selected as the training and validation dataset, and the remaining 26 images are the test dataset. The images cover various scenes (16 ports, multitemporal), with the size ranging from 12 000 × 12 000 pixels to 40 000 × 40 000 pixels. The GF3SDD contains 4653 labeled instances of different scales and orientations, whose distribution is shown in Fig. 10. All the ships are labeled by the professional SAR image interpreters using OBBs, which is considered reliable in this article. This dataset provides a strong support for the research of oriented SAR ship detection in complex scenes.

2) SSDD: The publicly available SSDD dataset consists of 1160 images collected from RadarSat-2, TerraSAR-X and Sentinel-1. The image resolution is range from 1 to 15 m. This dataset includes 2456 labeled ships in various scenes. It is divided into training, validation, and test sets with the proportion of 7:1:2.

B. Implementation Details

The experiments are implemented based on the MMDetection [66] codebase. This article does not apply any conventional preprocessing (e.g., speckle filtering, sea–land segmentation) to the raw images in the GF3SDD. In order to enhance the generalization ability and avoid overfitting, we adopt several augmentation strategies for the training set. Specifically, for each ship, we randomly crop three different sizes of patches (800 × 800, 1024 × 1024, and 1300 × 1300) from the raw images and make sure that this object is included in the patch. In this way, we obtain about 10 200 image patches used to train the detector. During the training, the patches are randomly cropped (the scale is within the range of [0.9, 1.0, 1.1]) for further augmentation and then resized into 608 × 608 as the input of the network. The model is trained on a 16-GB NVIDIA Tesla P100 GPU for 100 epochs with a total 16 images per minibatch. We adopt the Adam optimizer, with the initial learning rate 0.00015, which is divided by a factor of 10 at the 70th epoch.

During the inference, the test images are processed in a sliding window fashion. A series of 1024 × 1024 patches are cropped from the raw images with a stride of 512 and then directly fed into the detector. The outputs of patches are merged to obtain the final detection results in large-scale scenes. In this process, the NMS algorithm with the IoU threshold of 0.2 is applied to filter the overlapped bounding boxes.

C. Evaluation Metrics

In the experiments, the widely used criteria including precision, recall, F1-measure, and average precision (AP) are adopted to evaluate the performance of an object detector.

A detection result is typically labeled as a true positive (TP) if the IoU between the predicted box and the ground truth is higher than a threshold (generally set as 0.5). Otherwise, it is considered as a false positive (FP). If a ground truth has no matched detection box, it is regarded as a false negative (FN). It is worth noting that the IoU is calculated between two OBBs rather than HBBs. The precision measures the correctness of the prediction results, and the recall is the fraction of correctly detected ships among ground truths. They are computed as

$$\text{Precision} = \frac{\text{N}_{\text{TP}}}{\text{N}_{\text{TP}} + \text{N}_{\text{FP}}} \tag{16}$$

$$\text{Recall} = \frac{\text{N}_{\text{TP}}}{\text{N}_{\text{TP}} + \text{N}_{\text{FN}}} \tag{17}$$

where N_{TP}, N_{FP}, and N_{FN} denotes the number of TP, FP, and FN, respectively. The F1-measure is the harmonic mean of the precision and recall

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{18}$$

The AP metric aims to comprehensively reflect the quality of a detector. All the predictions are first ranked in the descending order according to the detection confidence level. The precision and recall are calculated at each unique level to obtain a precision–recall (PR) curve. The curve is expressed as C50 at the IoU threshold of 0.5. The AP summarized the shape of C50, and it is defined as the average of maximum precision for the recall in $\text{S} = \{0, 0.01, \ldots, 1\}$

$$\text{AP} = \frac{1}{101}\sum_{r \in \text{S}} p'(r), \quad p'(r) = \max_{r*:r* \geq r} p(r*) \tag{19}$$

where $p(r*)$ represents the corresponding precision value for the recall $r*$.

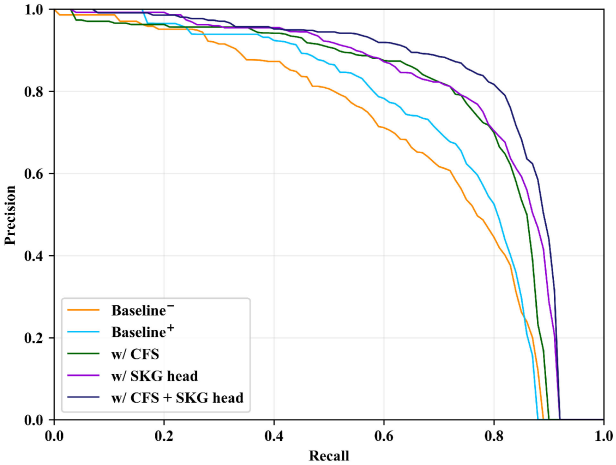


Fig. 11. PR curves of different improvements in the proposed method.

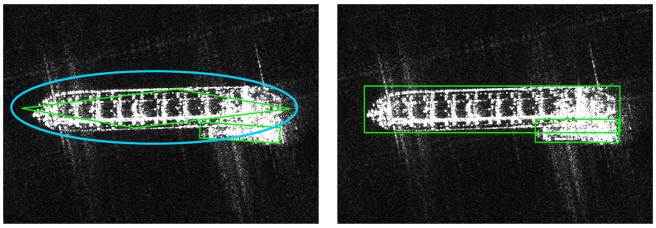
D. Ablation Studies

We perform a series of ablation experiments on the GF3SDD to analyze the contribution of each component in the proposed method. The preliminary network (illustrated in Section III-A) without and with the rotation factor branch is denoted as baseline^{-} and baseline^{+}, respectively. We first evaluate the performance of the rotation factor branch. Then, the CFS module and the scattering keypoint guide (SKG) head based on baseline^{+} are studied in detail. For a fair comparison, all the experiments are conducted using the same settings and training strategies. The overall ablation studies are reported in Table I, and their PR curves are presented in Fig. 11. It is evident that each proposed component brings a significant improvement to the detector. The final network achieves 14.32% higher AP compared with baseline^{-}. Meanwhile, the number of parameters and inference time only have a marginal increase. Note that the inference time here reflects the average runtime speed per patch rather than the large-scale image.

1) Effect of Rotation Factor Branch: In our method, two types of bounding boxes are predicted simultaneous, and the rotation factor branch is introduced to select the HBB or the OBB as the output representation form for a prediction result. The HBB form is adopted if the predicted rotation factor is higher than a predefined threshold. We first investigate the impact of this threshold with varying settings. It is found that the detector achieves the best performance when the threshold

TABLE I
INFLUENCE OF EACH COMPONENT IN THE PROPOSED METHOD

Rotation Factor	CFS Module	SKG Head	Recall	Precision	F1	AP	Params(M)	Time(ms)
×	×	×	0.6843	0.6364	0.6595	0.6831	28.40	**57.4**
✓	×	×	0.7308	0.6726	0.7005	0.7213	28.99	59.7
✓	✓	×	0.7660	0.7434	0.7545	0.7737	29.06	64.1
✓	×	✓	0.7837	0.7512	0.7671	0.7949	31.35	67.5
✓	✓	✓	**0.8189**	**0.7960**	**0.8073**	**0.8263**	31.42	68.9



(a) (b)

Fig. 12. Comparison results of the methods without and with the rotation factor branch. The green boxes are the detection results. The blue circles represent the object with inaccurate localization. (a) baseline$^-$. (b) baseline$^+$.

TABLE II
INFLUENCE OF DIFFERENT LOSS FUNCTIONS FOR THE ROTATION FACTOR BRANCH

Loss	Recall	Precision	F1	AP
BCE	0.7163	**0.6824**	0.6989	0.7159
Smooth L1	**0.7308**	0.6726	**0.7005**	**0.7213**

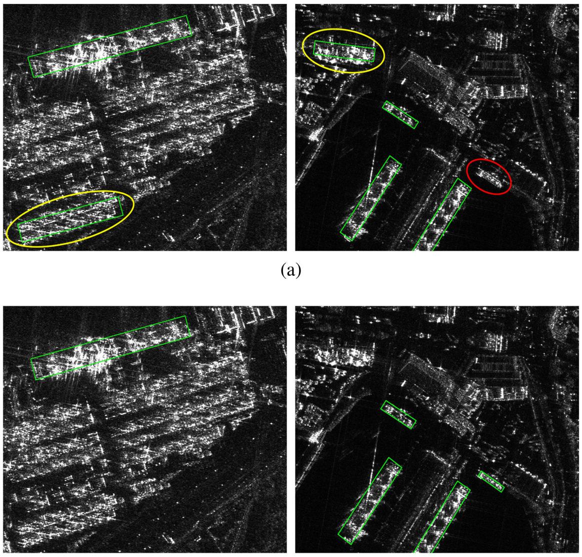
is set to 0.8. Table I shows that the network with rotation factor branch respectively increases the F1 and AP by 4.1% and 3.82% over baseline$^-$. Furthermore, the detection results in Fig. 12 demonstrate its positive effect on improving the localization accuracy. As mentioned above, directly regressing the oriented box boundaries encounters the problem of corner case. It is difficult for the detector to differentiate the boundary vectors of nearly horizontal objects. This results in the failure case, as shown in Fig. 12(a). In contrast, the method with rotation factor branch accurately locates the ships shown in Fig. 12(b) by transforming this issue into the horizontal object detection that can be handled more easily.

During the training, the rotation factor branch is supervised by smooth L1 loss. Considering that the OBB/HBB selection can also be regarded as a binary classification problem, we further verify the influence of the method with binary cross entropy (BCE). The target class label is set to 1 if the area ratio between two kinds of boxes is lower than 0.9. Otherwise, it is set to 0. Table II shows the effect of different loss functions. It turns out that using smooth L1 loss yields a slightly better performance than BCE.

2) Effect of the CFS Module: Two operators in the front end of the CFS module are utilized to extract the local and context features, respectively. We first remove one of these two feature extractors to analyze its impact. In the meantime, the feature selection is modified by the excitation operation, i.e., generate one set of channel weights to recalibrate the extracted feature.

TABLE III
ABLATION STUDIES OF THE CFS MODULE

Ctx	Loc	Select / Excit		Recall	Precision	F1	AP
×	×	×	×	0.7308	0.6726	0.7005	0.7213
✓	×	×	✓	0.7516	0.7328	0.7421	0.7602
×	✓	×	✓	0.7564	0.7217	0.7386	0.7518
✓	✓	×	✓	**0.7684**	0.7332	0.7504	0.7648
✓	✓	✓	×	0.7660	**0.7434**	**0.7545**	**0.7737**



(a)

(b)

Fig. 13. Comparison results of the methods without and with the CFS module. The green boxes are the detection results. The yellow and red circles represent the false alarms and missing ships, respectively. (a) baseline$^+$. (b) baseline$^+$ with CFS.

Then, we verify the importance of the feature selector. It is noteworthy that the following experiments are conducted based on baseline$^+$ instead of baseline$^-$. The ablation studies of CFS are reported in Table III. Ctx and Loc indicate the context and local feature extractor, respectively. Select and Excit indicate the feature selection and excitation operation, respectively. It can be seen that both local and context feature extraction play a positive role. Specifically, the method with context feature learning largely boosts the precision metric. In other words, the false alarm rate is significantly reduced. Then, adding the local feature extractor in the module gains a consistent improvement. We also make a comparison between feature selection and excitation. The better overall performance manifests the effectiveness of the selection operation. This is because the feature selector can

TABLE IV
INFLUENCE OF DIFFERENT SUPERVISION STRATEGIES IN THE PROPOSED FEATURE ADAPTION METHOD AND THE COMPARISON WITH THE ORIGINAL DEFORMABLE CONVOLUTION

Feature Adaption	Supervision	Recall	Precision	F1	AP
×	×	0.7308	0.6726	0.7005	0.7213
Deform conv	×	0.7628	0.7368	0.7496	0.7664
SKG deform conv	Single keypoint	0.7468	**0.7516**	0.7492	0.7710
	Keypoint set	**0.7837**	0.7512	**0.7671**	**0.7949**

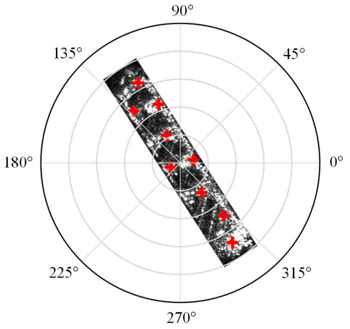


Fig. 14. Scattering keypoints in the polar coordinate.

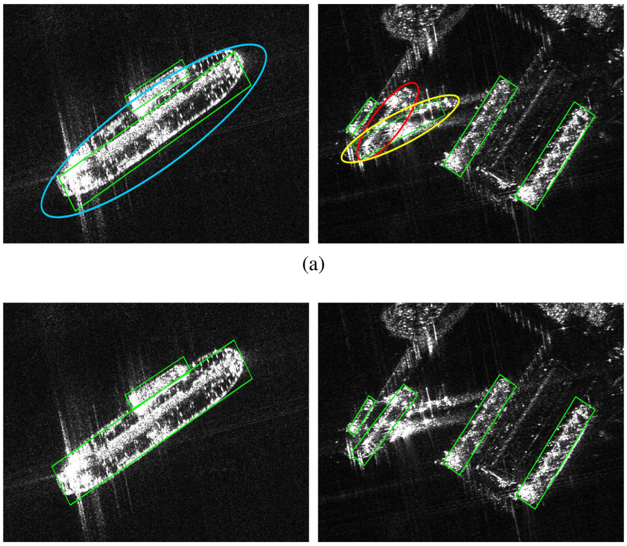


Fig. 15. Comparison results of the original deformable convolution and the SKG deformable convolution. The green boxes are the detection results. The yellow, red, and blue circles represent the false alarms, missing ships, and the objects with inaccurate localization, respectively. (a) Deformable convolution. (b) SKG deformable convolution.

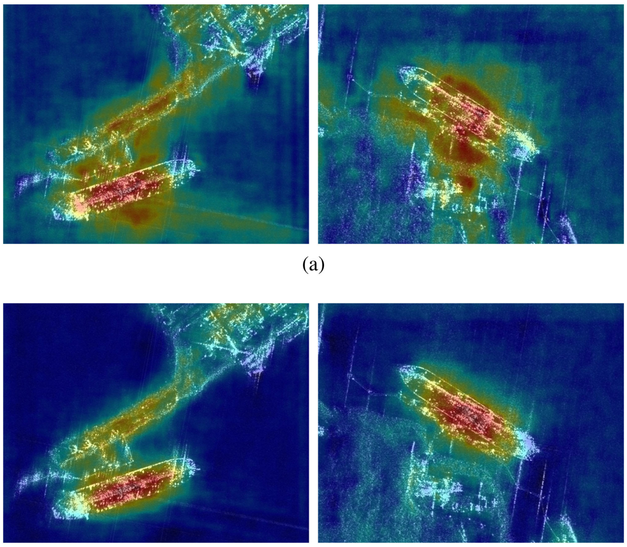


Fig. 16. Visualization of the features extracted by different methods. (a) Method without CFS. (b) Method with CFS.

make good use of both local and context information, processing different kinds of features in a more comprehensive manner. Compared with the baseline, the CFS module finally leads to a considerable increase of 5.24% in terms of the AP. Furthermore, some detection results are demonstrated in Fig. 13. We can observe that a stack of containers in the yard and the wharf facilities are incorrectly detected as the true objects in Fig. 13(a). The baseline method also tends to ignore the small-scale ships. In contrast, the method with CFS can better deal with the complex scenes in SAR images.

3) Effect of SKG Head: In the proposed detection head, the scattering keypoints are introduced to guide the input feature adaption via the deformable convolution. During the experiments, the set of scattering keypoints are sorted according to a certain rule. As shown in Fig. 14, the keypoints can be described in a polar coordinate. The object center is set as the origin of the polar system. We represent the keypoint sequence according to the ascending order of the polar angle and radius. The detector predicts the keypoint offsets with respect to the center of each instance, regarding the localization task as a regression problem. In most cases, smooth L1 loss is used for the supervision on each individual point. Our method adopts an alternative strategy that measures the discrepancy between the entire keypoint set and the ground truth using the Chamfer distance. A comparison is made for these two strategies. From Table IV, we can see that the supervision on point set achieves 2.39% higher AP than the counterpart on single point. This happens because the smooth L1 loss ignores the correlation of different keypoints and is restrictive for the point prediction, which results in an inferior performance. The localization loss of scattering keypoints in (15) is weighted by a balancing parameter λ. We reveal the influence of λ in Table V. The detection result is insensitive to this parameter and is impacted marginally by about 1%. It

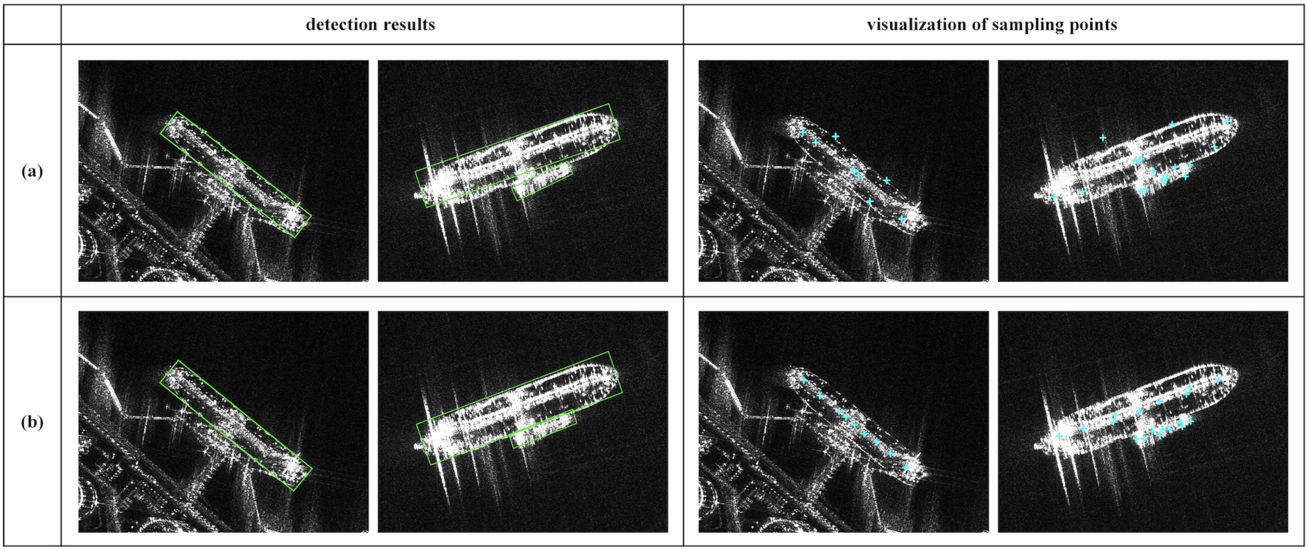


Fig. 17. Comparison between different feature adaption methods. The blue crosses represent the sampling points. The green boxes are the detection results. (a) Deformable convolution. (b) SKG deformable convolution.

TABLE V
RESULTS OF DIFFERENT BALANCING PARAMETER λ

λ	0.3	0.5	0.7	1	2
AP	0.7802	0.7903	0.7853	**0.7949**	0.7821

reaches the best when λ is set to 1, which makes the different terms in (15) achieve a good balance.

We make a further comparison between the original deformable convolution and the SKG method. The main difference between these two methods is the learning of sampling locations. Results in Table IV show that our method outperforms the original deformable convolution and improves the performance by a considerable margin over the baseline. This demonstrates the effectiveness of the guidance provided by the scattering keypoints. The comparison results are visualized in Fig. 15. It can be seen that the original deformable convolution may yield some inaccurate bounding boxes, leading to a relatively low detection rate and localization accuracy.

E. Discussion

1) Analysis of Context Information: In order to better show the effect of the CFS module, some qualitative results are presented in Fig. 16. For the extracted feature F, we compute the sum of the elements of each row in the channel dimension to generate the visualized activation heatmap. Note that the feature F has 1/4 resolution of the input image. Both the heatmap and the input are resized to a unified size for better visualization. As shown in Fig. 16(a), some parts of the harbor have high activation scores, and the context information of objects is obscure. It is difficult for the detector to focus on true objects under the strong interference. By introducing the CFS module in the process of multiscale feature fusion, objects in Fig. 16(b) become more discriminative. Meanwhile, the background clutter can be suppressed to a great extent. We argue that the improvement brought by CFS may be attributed to two aspects. First, benefitting from the long-range dependencies captured by the dilated convolution with large receptive field, the surrounding context is informative for the feature representation in SAR images. Second, context features are selected in a dynamic fashion, taking fully advantage of both spatial and semantic information from multiscale feature levels.

2) Analysis of Scattering Keypoints: The scattering keypoints are defined in Section III-C with the purpose of depicting the distinct and representative locations. Based on this conception, we formulate a feature adaption method with the SKG strategy and the deformable convolution. It can be seen from Figs. 15 and 17 that our method is superior to the original deformable convolution and is more robust to the varying imaging results. There are three possible reasons. First, the scattering keypoints can provide coarse-grained structural information, reflecting the geometric characteristics, e.g., the object scales and orientations, which is beneficial for the adjustment of the receptive field. Second, the proposed SKG scheme empowers the detector with stronger feature extraction ability. As shown in Fig. 17, our method tends to sample the significant features of the object (e.g., the oil pipeline for the tanker). In other words, the scattering keypoints provide an additional guidance for the convolution to search for the meaningful sampling locations. The extracted features can better adapt to the different imaging conditions for the same kind of objects in SAR images. Third, for the original deformable convolution, some sampling points are located outside the boundaries. The interference of nearby objects and the surrounding background with strong scattering intensity may result in the feature misalignment for the instance. This has a direct impact on the performance of the object center localization and the bounding box prediction.

TABLE VI
COMPARISON WITH DIFFERENT CNN-BASED METHODS ON THE GF3SDD

Method	Anchor-free	Recall	Precision	F1	AP	Time(ms)
RRPN	×	0.7228	0.6731	0.6971	0.6948	287.3
R^2CNN	×	0.7692	0.6630	0.7122	0.7297	110.8
RetinaNet OBB	×	0.7324	0.6277	0.6760	0.6466	92.0
Proposed method	✓	**0.8189**	**0.7960**	**0.8073**	**0.8263**	**68.9**

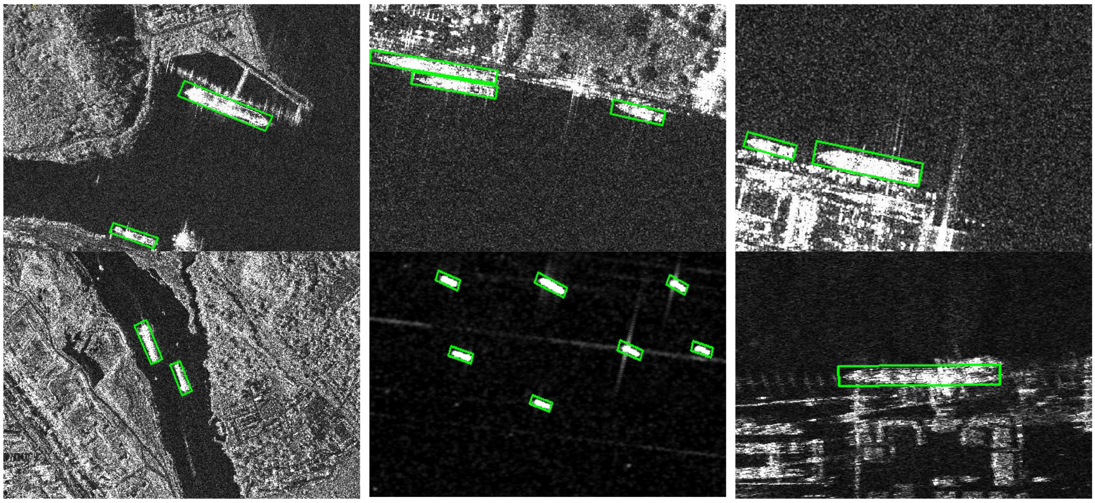
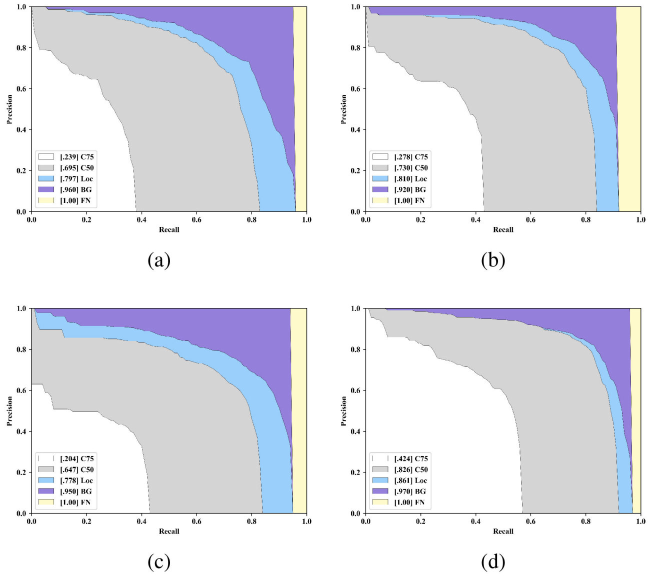


Fig. 18. Detection results on the SSDD.

TABLE VII
COMPARISON WITH DIFFERENT CNN-BASED METHODS ON THE SSDD

Method	Recall	Precision	F1	AP
RRPN	0.9097	0.8695	0.8891	0.8611
R^2CNN	0.9307	0.9115	0.9210	0.9289
RetinaNet OBB	0.8971	0.8389	0.8670	0.8741
Proposed method	**0.9433**	**0.9258**	**0.9345**	**0.9502**



(a) (b)

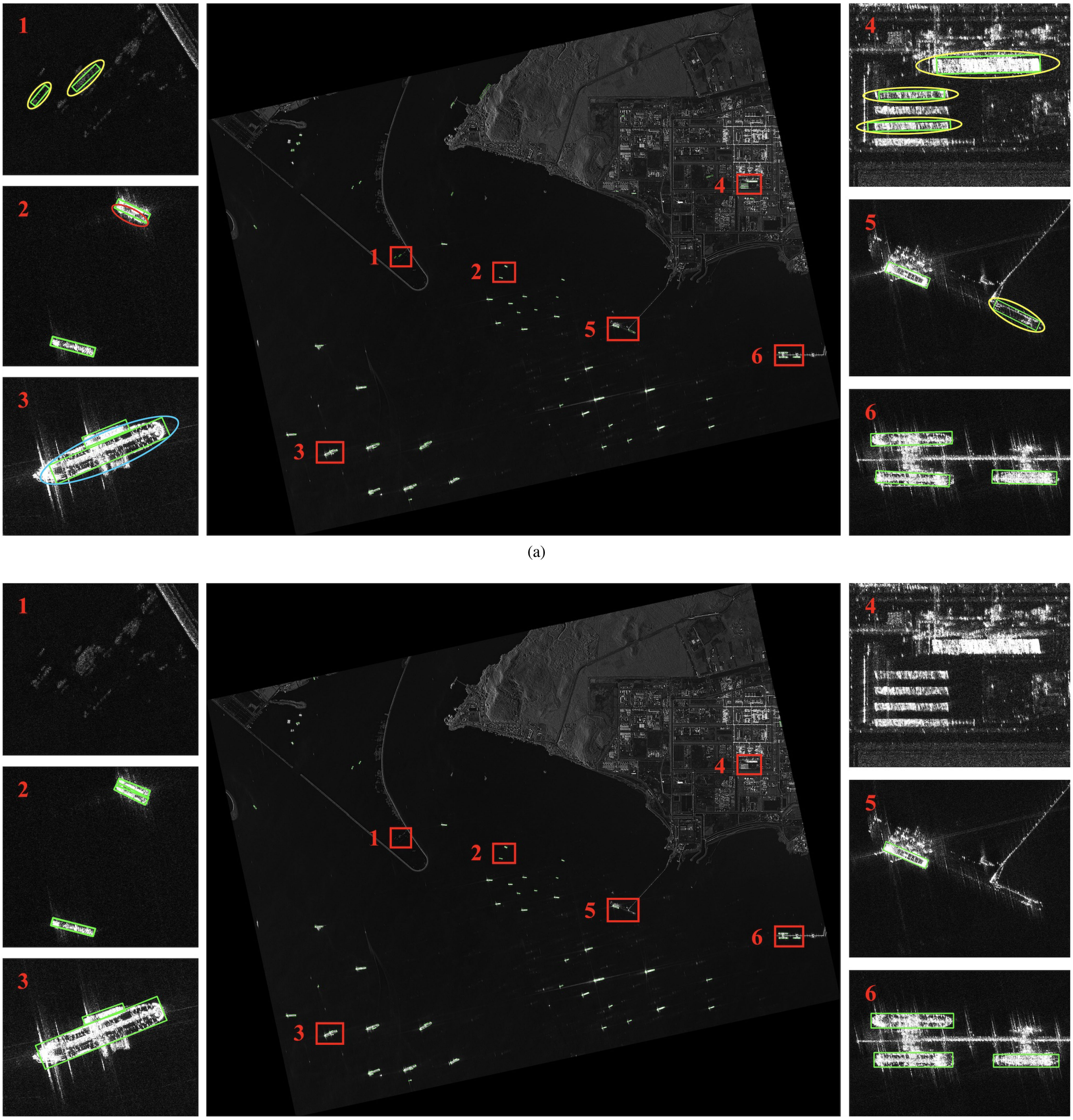
(c) (d)

Fig. 19. Detailed analysis of different method performance. (a) RRPN. (b) R ^{2} CNN. (c) RetinaNet OBB. (d) Proposed method.

F. Comparison With State-of-the-Art Methods

We compare the proposed detector with some state-of-the-art CNN-based methods in Table VI. For a fair comparison, the results of single model with single-scale testing as well as the corresponding inference time are reported. It can be seen that our method achieves an AP of 82.63% on the GF3SDD test dataset and outperforms other competitive detectors in terms of both the accuracy and the speed, showing a great superiority. In order to further verify the generalization ability of the proposed method, all the detectors are experimented on the SSDD. Table VII illustrates that our method still achieves a superior performance. Some detection results are shown in Fig. 18. It is also worth noting that our anchor-free method has simpler architecture and higher computational efficiency. Given the merits of the anchor-free framework and the considerable improvement brought by the effective exploration of SAR imaging mechanism, the proposed method proves to be a promising direction for the oriented SAR ship detection in large-scale and complex scenes.

Moreover, in order to further show the localization capability of different methods, we conduct a detailed analysis on the output results of the GF3SDD to investigate the influence on the performance of each type of detection error. In addition to the evaluation methods mentioned in Section IV-C, we introduce several additional PR curves in Fig. 19. The corresponding metrics are calculated by (19) and shown in brackets in the legend. C75 denotes the PR at the stricter IoU threshold of 0.75, which can better evaluate the quality of the predicted detection



(a)

(b)

Fig. 20. Comparison results of R^{2} CNN and our method on the GF3SDD. The green boxes are the detection results. The yellow, red, and blue circles represent the false alarms, missing ships, and the objects with inaccurate localization, respectively. (a) R^{2} CNN. (b) Proposed method.

boxes. Loc is drawn at the IoU threshold of 0.1 to measure the impact of localization error. This kind of error happens when an object is detected by a misaligned box that has a minor overlap ($0.1 \leq \text{IoU} < 0.5$) with the ground truth. Its influence on the detector performance can be reflected by the blue area in each plot. For the remaining FP boxes with the overlap less than 0.1, they are regarded as the confusion with the background. BG is drawn after removing all the FPs. The yellow area in each plot reflects the error based on the FNs. The comparison in Fig. 19 shows that our method enjoys several advantages. First, with a stricter IoU threshold of 0.75, our method yields a significant improvement over other three detectors, which suggests that it has a better box regressor to generate more high-quality detection results. Second, the proposed detector is least affected by the localization error (about 3.5%, while more than 8.0% for other methods). The objects can be located with well-aligned bounding boxes. Third, the number of FNs is reduced under a very low confidence threshold, while other methods miss more

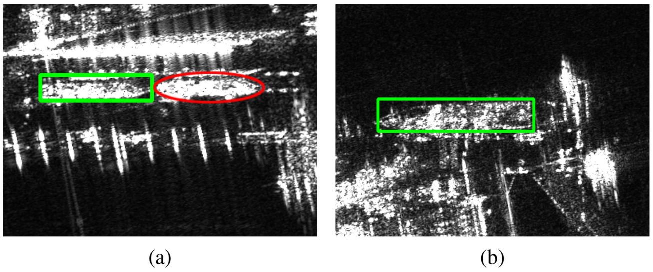


Fig. 21. (a) and (b) Limitations for the proposed method. The red circle represents the missing ship.

objects. The effectiveness of the proposed method can also be verified in Fig. 20. From Fig. 20(a), we can observe that some reef and inshore facilities are incorrectly detected, and R ^{2} CNN cannot well handle the two ships laying alongside of each other. In contrast, our method has a better localization capability and achieves a superior performance in different scenes.

Despite the great improvement in our research, there are still a few limitations. Fig. 21 presents two typical failure cases. In Fig. 21(a), one object in the dockyard is detected, while the other one is ignored. We argue that the surrounding context for the ships located in the inland area is complex and ambiguous, which disrupts the object feature representations. Under this circumstance, the dependencies with nearby objects should be considered. For the object in Fig. 21(b), its rear deck is not included by the misaligned bounding box. The detector fails to capture the sufficient clues to describe this local area with the indistinct scattering characteristics. Although the ship boundary presents the high scattering intensity and seems to be informative for the localization, the detector cannot exploit this kind of implicit and low-level information. Motivated by some studies in the task of segmentation [67]–[69], we will consider a relevant model with the self-attention mechanism and focus on a more effective representation for the scattering characteristics to address the above issues in the future work.

V. Conclusion

In this article, a novel method based on the SKG-Net is proposed for the oriented ship detection in large-scale SAR images. The anchor-free architecture and the proposed representation form for the OBBs make the detector more flexible and efficient. The designed CFS module selects the discriminative features in a dynamic fashion, with the aim to enhance both the semantic and spatial information of objects while suppressing the background interference. Furthermore, considering the unique scattering mechanism, we introduce a novel conception based on the scattering keypoints to guide the adaptive feature leaning in order to deal with the variability issue of object imaging characteristics. Ablation studies show that the context information and scattering keypoints play a vital role in SAR object detection. The comparison with other CNN-based detectors on GF3SDD and SSDD demonstrates the superiority of our method.

References

[1] Q. Guo, H. Wang, and F. Xu, “Scattering enhanced attention pyramid network for aircraft detection in SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7570–7587, Sep. 2021.

[2] B. Liu, Z. Zhang, X. Liu, and W. Yu, “Representation and spatially adaptive segmentation for PolSAR images based on Wedgelet analysis,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 4797–4809, Mar. 2015.

[3] M. Zhang, W. Li, R. Tao, and S. Wang, “Transfer learning for optical and SAR data correspondence identification with limited training labels,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1545–1557, Dec. 2020.

[4] G. Yang, H. Li, W. Yang, K. Fu, Y. Sun, and W. J. Emery, “Unsupervised change detection of SAR images based on variational multivariate Gaussian mixture model and Shannon entropy,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 826–830, May 2019.

[5] L. Du, L. Li, D. Wei, and J. Mao, “Saliency-guided single shot multibox detector for target detection in SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3366–3376, May 2020.

[6] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, and X. X. Zhu, “Buildings detection in VHR SAR images using fully convolution neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1100–1116, Feb. 2019.

[7] D. J. Crisp, “The state-of-the-art in ship detection in synthetic aperture radar imagery,” DSTO Inf. Sci. Lab., Edinburgh, SA, Australia, Tech. Rep. DSTO-RR-0272, 2004.

[8] G. B. Goldstein, “False-alarm regulation in log-normal and Weibull clutter,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES- 9, no. 1, pp. 84–92, Jan. 1973.

[9] G. Gao, L. Liu, L. Zhao, G. Shi, and G. Kuang, “An adaptive and fast CFAR algorithm based on automatic censoring for target detection in high-resolution SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 6, pp. 1685–1697, Jun. 2009.

[10] 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 Observ. Remote Sens.*, vol. 8, no. 7, pp. 3329–3337, Jul. 2015.

[11] C. Wang, M. Liao, and X. Li, “Ship detection in SAR image based on the alpha-stable distribution,” *Sensors*, vol. 8, no. 8, pp. 4948–4960, Aug. 2008.

[12] A. Balleri, A. Nehorai, and J. Wang, “Maximum likelihood estimation for compound-Gaussian clutter with inverse Gamma texture,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 2, pp. 775–779, Apr. 2007.

[13] H. Li, W. Hong, Y. Wu, and P. Fan, “On the empirical-statistical modeling of SAR images with generalized Gamma distribution,” *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 386–397, Jun. 2011.

[14] M. Weiss, “Analysis of some modified cell-averaging CFAR processors in multiple-target situations,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES- 18, no. 1, pp. 102–114, Jan. 1982.

[15] V. G. Hansen, “Constant false alarm rate processing in search radars,” in *Proc. IEE Conf. Publ. 105, Radar-Present Future*, 1973, pp. 325–332.

[16] G. V. Trunk, “Range resolution of targets using automatic detectors,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES- 14, no. 5, pp. 750–755, Sep. 1978.

[17] S. Blake, “OS-CFAR theory for multiple targets and nonuniform clutter,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, no. 6, pp. 785–790, Nov. 1988.

[18] M. Tello, C. Lopez-Martinez, and J. J. Mallorqui, “A novel algorithm for ship detection in SAR imagery based on the wavelet transform,” *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 201–205, Apr. 2005.

[19] C. H. Gierull, “Demystifying the capability of sublook correlation techniques for vessel detection in SAR imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2031–2042, Apr. 2019.

[20] 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.

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

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

[23] P. Wang, X. Sun, W. Diao, and K. Fu, “FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, Dec. 2020.

[24] X. Sun, Y. Liu, Z. Yan, P. Wang, W. Diao, and K. Fu, "SRAF-Net: Shape robust anchor-free network for garbage dumps in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6154–6168, Jul. 2021.

[25] X. Sun, P. Wang, C. Wang, Y. Liu, and K. Fu, "PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 50–65, Mar. 2021.

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

[27] J. Zhao, Z. Zhang, W. Yu, and T. Truong, "A cascade coupled convolutional neural network guided visual attention method for ship detection from SAR images," *IEEE Access*, vol. 6, pp. 50693–50708, 2018.

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

[29] L. Liu, G. Chen, Z. Pan, B. Lei, and Q. An, "Inshore ship detection in SAR images based on deep neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 25–28.

[30] J. Wang, C. Lu, and W. Jiang, "Simultaneous ship detection and orientation estimation in SAR images based on attention module and angle regression," *Sensors*, vol. 18, no. 9, Aug. 2018, Art. no. 2851.

[31] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, Aug. 2017, Art. no. 860.

[32] C. Chen, C. He, C. Hu, H. Pei, and L. Jiao, "MSARN: A deep neural network based on an adaptive recalibration mechanism for multiscale and arbitrary-oriented SAR ship detection," *IEEE Access*, vol. 7, pp. 159262–159283, 2019.

[33] 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.

[34] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1331–1344, Feb. 2021.

[35] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[36] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Dec. 2016, pp. 379–387.

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

[38] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.

[39] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "DenseBox: Unifying landmark localization with end to end object detection," 2015, *arXiv:1509.04874*.

[40] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2016, pp. 516–520.

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

[42] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "FoveaBox: Beyond anchor-based object detector," 2019, *arXiv:1904.03797*.

[43] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 765–781.

[44] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6568–6577.

[45] Q. An, Z. Pan, and H. You, "Ship detection in Gaofen-3 SAR images based on sea clutter distribution analysis and deep convolutional neural network," *Sensors*, vol. 18, no. 2, Feb. 2018, Art. no. 334.

[46] 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.*, May 2017, pp. 1–4.

[47] Y. Zhao, L. Zhao, C. Li, and G. Kuang, "Pyramid attention dilated network for aircraft detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 662–666, Apr. 2021.

[48] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[49] Y. Jiang *et al.*, "R ^{2} CNN: Rotational region CNN for arbitrarily-oriented scene text detection," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2018, pp. 3610–3615.

[50] Q. An, Z. Pan, L. Liu, and H. You, "DRBox-v2: An improved detector with rotatable boxes for target detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8333–8349, Nov. 2019.

[51] Z. Pan, R. Yang, and Z. Zhang, "MSR2N: Multi-stage rotational region based network for arbitrary-oriented ship detection in SAR images," *Sensors*, vol. 20, no. 8, Apr. 2020, Art. no. 2340.

[52] R. Yang, Z. Pan, X. Jia, L. Zhang, and Y. Deng, "A novel CNN-based detector for ship detection based on rotatable bounding box in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1938–1958, Jan. 2021.

[53] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.

[54] T. 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.

[55] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 483–499.

[56] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

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

[58] Y. Xu *et al.*, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[59] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[60] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 764–773.

[61] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2463–2471.

[62] L. Huang *et al.*, "OpenSARShip: A dataset dedicated to Sentinel-1 ship interpretation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 195–208, Jan. 2018.

[63] 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.

[64] X. Sun, Z. Wang, Y. Sun, W. Diao, Y. Zhang, and K. Fu, "AIR-SARShip-1.0: High-resolution SAR ship detection dataset," *J. Radars*, vol. 8, pp. 852–862, 2019.

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

[66] K. Chen *et al.*, "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155.*

[67] X. Sun, A. Shi, H. Huang, and H. Mayer, "BAS ^{4} Net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5398–5413, Sep. 2020.

[68] 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.*, Oct. 2019, pp. 603–612.

[69] S. Peng, W. Jiang, H. Pi, X. Li, H. Bao, and X. Zhou, "Deep snake for real-time instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2020, pp. 8530–8539.

Kun Fu received the B.Sc., M.Sc., and Ph.D. degrees in electronic information engineering from the National University of Defense Technology, Changsha, China, in 1995, 1999, and 2002, respectively.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, remote sensing image understanding, geospatial data mining, and visualization.

Jiamei Fu received the B.Sc. degree in communication engineering from the Beijing Institute of Technology, Beijing, China, in 2018. She is currently working toward the Ph.D. degree in signal and information processing with the University of Chinese Academy of Sciences, Beijing, and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing.

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

Zhirui Wang received the B.Sc. degree in electronic information engineering from the Harbin Institute of Technology, Harbin, China, in 2013, and the Ph.D. degree in electronic information engineering from Tsinghua University, Beijing, China, in 2018.

He is currently an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include synthetic aperture radar (SAR) terrain classification, SAR target detection, and recognition.

Xian Sun received the B.Sc. degree in electronic information engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees in electronic information engineering from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2009.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.