SAR Object Detection Encounters Deformed Complex Scenes and Aliased Scattered Power Distribution

Yawei Zhang, Yu Cao, Xubin Feng, Meilin Xie, Xin Li, Yao Xue, and Xueming Qian, *Member, IEEE*

***Abstract*—Synthetic aperture radar (SAR) is widely used in terrain classification, object detection, and other fields. Compared with anchor-based detectors, anchor-free detectors remove the anchor mechanism and implement detection box encoding in a more elegant form. However, anchor-free detectors are limited by complex scenes caused by geometric transformations, such as overlaying, shadow, vertex displacement during SAR imaging. And the scattered power distribution of noise is similar to the edge of the object, making it difficult for the detector to locate the edge of the SAR object accurately. In order to alleviate these problems, we propose a high-speed and high-performance SAR image anchor-free detector. First, we propose a shallow feature refinement (SFR) module to effectively extract and retain the detailed information of objects, while coping with deformed complex scenes. Second, we analyze the optimization focus of the detector at different training iterations and propose iteration-aware loss to guide the detector, making the detector more accurately locate the edge of the object disturbed by the noise scattered power distribution. Third, number estimation helps to detect objects with more flexible criteria in box selection without manual labor. Compared with mainstream optical object detectors and SAR dedicated detectors, our method achieves the best speed-accuracy tradeoff on the SAR-ship dataset, with 96.4% average precision when the value of intersection over union is 50% (AP_{50}) at 64.9 frames per second. The experimental results prove the effectiveness of our method.**

***Index Terms*—Iteration-aware loss, number estimation, synthetic aperture radar (SAR) object detection, scattered power distribution aliasing, shallow feature refinement (SFR).**

I. Introduction

Synthetic aperture radar (SAR) has the ground-penetrating capability in all weather and all day. It has unique advantages in many fields [1]–[6]. The availability of large amounts of data and the increase in computing power promote the rapid development of object detectors based on deep learning [7]. However, due to the perspective contraction, top and bottom stagnation, shadows, and other special characteristics of SAR imaging, the interpretation of SAR images is difficult. SAR object detection in deformed complex scenes with noise interference is challenging.

There are some excellent works in SAR object interpretation [8]–[13]. In response to the deficiency of faster R-CNN [14] using only a single-scale feature map to generate object candidate regions, Zhao *et al.* [12] proposed an SAR object detection network by using multiscale features. For the problem of detection in the large scene, Chen *et al.* [13] first uses a lightweight object prescreening full convolutional network to prescreen possible objects. To detect small objects in large-scale SAR images, Kang *et al.* [15] proposed an SAR detector with a high-resolution region proposal network and object detection containing contextual information.

Even though these methods improve SAR detection performance, most of these methods require a tight arrangement of anchors. The performance has a lot to do with the size of the anchors, aspect ratios, the division ratio of positive and negative samples, etc., that need to be carefully adjusted manually. Furthermore, most of these methods rely on the nonmaximum suppression operation, which requires complex intersection over union (IoU) calculations and slows down the speed of the detector.

The limitations *motivate* us to establish a concise and effective anchor-free detector for SAR object detection in complex scenes with noise interference. Existing anchor-free detectors [16], [17] remove the anchor mechanism to improve the detection speed and simplify the postprocessing process. However, unlike natural images, SAR images have their own characteristics that challenge the design of anchor-free SAR object detectors.

Manuscript received December 9, 2021; revised January 30, 2022; accepted February 18, 2022. Date of publication March 9, 2022; date of current version June 10, 2022. This work was supported in part by NSFC under Grant 61772407, Grant 61732008, and Grant 62103317, and Youth Innovation Promotion Association CAS, and in part by the Science and Technology Program of Xi'an, China under Grant 21RGZN0017. *(Corresponding authors: Xubin Feng; Meilin Xie; Xueming Qian.)*

Yawei Zhang, Xin Li, and Yao Xue are with the School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: zhangyawei26@stu.xjtu.edu.cn; lingfengyueguang@stu.xjtu.edu.cn; xueyao@xjtu.edu.cn).

Yu Cao is with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China, with the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China, with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Key Laboratory of Space Precision Measurement Technology, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: caoyu@opt.ac.cn).

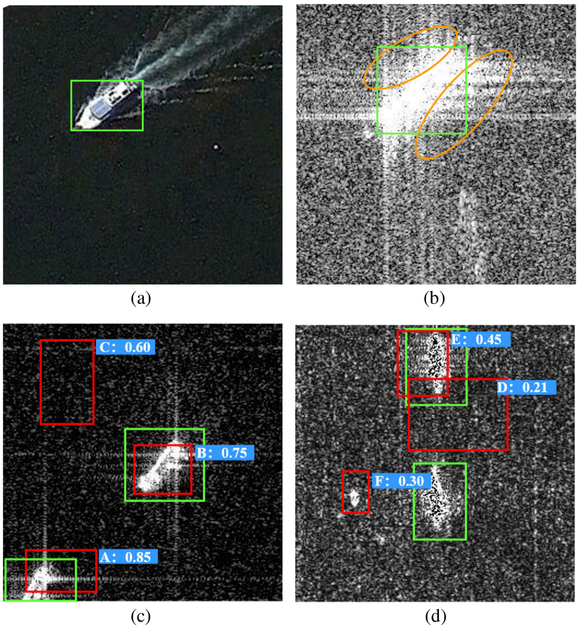
Xubin Feng and Meilin Xie are with the Space Precision Measurement Laboratory, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: fengxubin@opt.ac.cn; xiemeilin@opt.ac.cn).

Xueming Qian is with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, School of Information and Communication Engineering, SMILES LAB, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qianxm@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3157749

First, SAR images often present object overlaps, shadows, top and bottom stagnation, all of geometric deformation submerge useful features and make feature extraction more tricky. In optical image detection, to effectively detect objects from complex scenes, Liu *et al.* [18] added a bottom-up path enhancement branch based on feature pyramid network (FPN) [19], which effectively fuses feature and detect objects in complex scenes. ASFF [20] proposed an adaptive spatial feature fusion method. This method fuses the features of different layers by learning the weight parameters. Generally, these FPN-based methods process and merge the deep features extracted by the backbone (feature maps with dimensions of 1/8, 1/16, and 1/32 of the original image size). The shallow features (feature maps whose size is 1/4 of the image size) containing object information, complex background is directly discarded. The geometric deformation in the SAR imaging process makes detectors unable to easily extract the feature information from the complex geographic background. We propose a shallow feature refinement (SFR) module to extract and refine the shallow feature map that contains object information and scene feature with geometric deformation, and then merges it with the information extracted by FPN. *Through SFR, the detector can effectively extract and retain the detailed information of objects, while coping with geometric deformations, such as overlap, shadows, and top and bottom stagnation.*

As the second character of SAR images, the edge of the object is often aliased with the noise scattered power distribution. Fig. 1(a) and (b) is the optical image and SAR image of a ship, respectively. As shown in the orange circle in (b), the ship edge's power distribution and the noise are seriously aliased. Scattered power distribution aliasing makes the anchor-free detector with a single regression loss function unable to accurately cover the intact object. How to guide the detector to accurately locate the edge of the SAR object during the training iterations is a challenging task. In anchor-based detector [21] training, the approximate positions are first proposed by region proposal network, and then they are finely adjusted. Although the complicated anchor calculation slows down the detection speed, the anchor-based detector can achieve high performance by performing constraint refinement on the preset anchor. The anchor-free detectors [16], [17] remove the anchor with a more elegant detection box encoding to improve the speed of the detector. However, the lack of preset anchors slows down the convergence of the detector. We propose an iteration-aware loss to guide the optimization focus in different training iterations. It can help the detector to quickly locate the approximate position of the SAR object in the early iterations so that the detector can have more training iterations to more accurately locate the edge of the SAR object. *The SAR detector trained with iteration-aware loss can obtain more accurate edge positioning, while coping with aliased scattering power distribution.*



(a) (b)

(c) (d)

Fig. 1. Red boxes and the green ones are the prediction boxes and ground truths, respectively. In many SAR images, power distributions between object edges and noises, as shown in the orange regions of (b), are seriously aliased. Such power distribution aliasing makes the detector unable to accurately cover the object. In (c) and (d), the confidence of the detection results is high and low, respectively. A fixed threshold that is too high (e.g., 0.7) or too low (e.g., 0.4) cannot effectively obtain high-quality boxes. Number estimation works as an adaptive method to assist accurate box selection. (a) Optical object with clear edges. (b) SAR object with noise aliasing. (c) Boxes with high confidence. (d) Boxes with low confidence.

Third, due to the changes in the elevation and azimuth angles of the SAR during imaging, the same background has differences under different imaging conditions. The confidence level of the detection boxes varies widely, a fixed threshold is not reliable for effective and robust detection. Fig. 1(c) and (d) are two examples of the detector's detection of objects in SAR images. The green boxes are the ground-truths containing the objects, and the red ones are the predicted boxes. When the confidence threshold is high (e.g., 0.7), the left image gets correct object boxes, while the right image can not obtain a correct prediction at all. When the confidence threshold is low (e.g., 0.4), there will be serious false detection in the left image, and the detection result is good in the right one. How to better filter the low-quality boxes with high confidence in the detection result is a problem that needs to be considered. We propose number estimation to filter the low-quality detection boxes more effectively without manual labor. Number estimation will filter the low-quality detection boxes according to the prediction of object number in SAR image. *The number estimation can retain the detection boxes in SAR images coping with imaging differences in similar backgrounds adaptively.*

The main contributions of this article are summarized as follows.

1) We propose SFR for refining the shallow features extracted by deformed complex scenes to cope with geometric deformations, such as overlap, shadows, and top and bottom stagnation. After fusion with the feature map extracted by traditional FPN, a more representative feature map containing detailed information can be obtained, laying a better foundation for subsequent regression and classification tasks.

2) We propose an iteration-aware loss that can guide the detector to locate more accurate edges of the object from aliasing scattering power distributions, so as to realize the training optimization focus to determine the position roughly in early training iterations and then refine it in later training iterations.

3) We propose a number estimation that can filter low-quality detection boxes effectively in SAR images with imaging differences. This method can avoid the troublesome manual fixed threshold selection and retain the correct high-quality detection boxes with low confidence.

II. Related Works

A. Optical Object Detection

Deep convolutional neural networks achieve great success in crowd scene analysis [22], medical [23], and other fields in visual task. Common object detectors can be divided into the two-stage detector and one-stage ones according to the stage division. Typical representatives of the two-stage detector are faster R-CNN [14], cascade R-CNN [24], and Libra R-CNN [25]. Typical representatives of one-stage object detectors are SSD [26] and YOLOv3 [27]. Normally, the one-stage object detector is faster, and the two-stage object detector has higher performance. Compared with the two-stage detectors represented by Faster R-CNN [14], the one-stage detectors represented by CenterNet [16] have advantages in speed.

The proposal of CenterNet [16] and FCOS [17] is a landmark work based on anchor-free object detectors, compared to the original anchor-based detector, the anchor-free object detectors achieve faster detection speed and inferior performance. Then anchor-free detector has gradually become a hot spot in academic research [28]. The core idea of CenterNet [16] and FCOS [17] is to predict the center point of the object to be detected and use this as the basis to predict the distance from the center point to the bounding box. The difference is that CenterNet [16] uses heatmap to enhance the center point regression, while FCOS [17] uses centerness to strengthen the center point regression.

In natural scenes, massive amounts of data greatly promote the development of object detection. In SAR image detection, most of the existing methods are anchor-based [29], [30]. Inspired by the idea of anchor-free, this article proposes a fast and effective SAR object detector.

B. SAR Object Detection

At present, deep learning has made remarkable achievements in optical image object detection and classification [6], [31]. But SAR object detection is still a challenging and important task [7]. SAR object detection's purpose is to extract the potential area containing the object. SAR detection method can be divided into two types: 1) Traditional object detection method represented by constant false alarm rate (CFAR) [32], [33]. 2) Deep learning method migrated from optical image detection to SAR image detection [6], [34]–[38]. And in Fig. 2(a) of miniSAR dataset [39] and (b) of SAR-Ship dataset [40], the green boxes are the ground-truths, and the red circles are the interference after imaging in complex scenes. Due to geometric deformation, SAR image objects in complex scenes are often confused with the background. Fig. 2(c) and (d) are the images in MSTAR dataset [41] and SSDD dataset [42]. The speckle noise brought by SAR imaging and the noise power distribution close to the object edge cause problems for the accurate edge positioning of the detector.

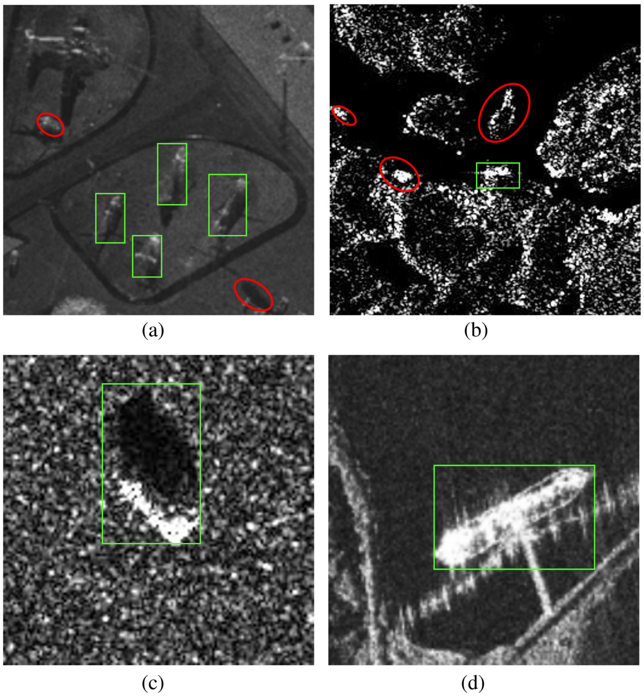


Fig. 2. Examples of SAR datasets, the green boxes are the real objects. The deformed complex background,as shown in the red circles in (a) and (b) can easily be confused with the objects. The noise power distribution in (c) and (d) make it difficult to locate the object edge. (a) Example of miniSAR dataset. (b) Example of SAR-ship dataset. (c) Example of MSTAR dataset. (d) Example of SSDD dataset.

CFAR-based SAR object detection mainly performs statistical modeling on clutter, and readily considers the characteristics of the object. Inspired by human visual attention mechanisms, some algorithms [43]–[45] are proposed to simulate human selective visual attention mechanisms, construct saliency maps for optical images, and extract regions in optical images. However, high-resolution SAR images often have complex backgrounds, and the object area occupies a small proportion of the image. The CFAR-based algorithm needs to traverse the entire image cell by cell, which is inefficient and the representation of manually extracted features is weak.

As for the SAR object detection method based on deep learning, given the problem that convolutional neural networks have too many parameters and require a large amount of training data, Cozzolino *et al.* [30] proposed to use dense sliding windows to obtain image blocks, and then quickly input these images into a fully convolutional neural network to classify the object and clutter. Aiming at the problem of the difficulty of applying deep learning technology due to the lack of a publicly labeled dataset for SAR images, Li *et al.* [42] constructed the first public SAR image ship object dataset SSDD. It contains ship SAR images under different resolutions, sizes, sea conditions, sensor types, etc. There are a total of 2456 ship objects. The author improved the performance of SAR ship detection using strategies including feature fusion, transfer learning, and hard negative mining based on faster-RCNN [14]. To reduce the amount of calculation, Wang *et al.* [46] added the spatial

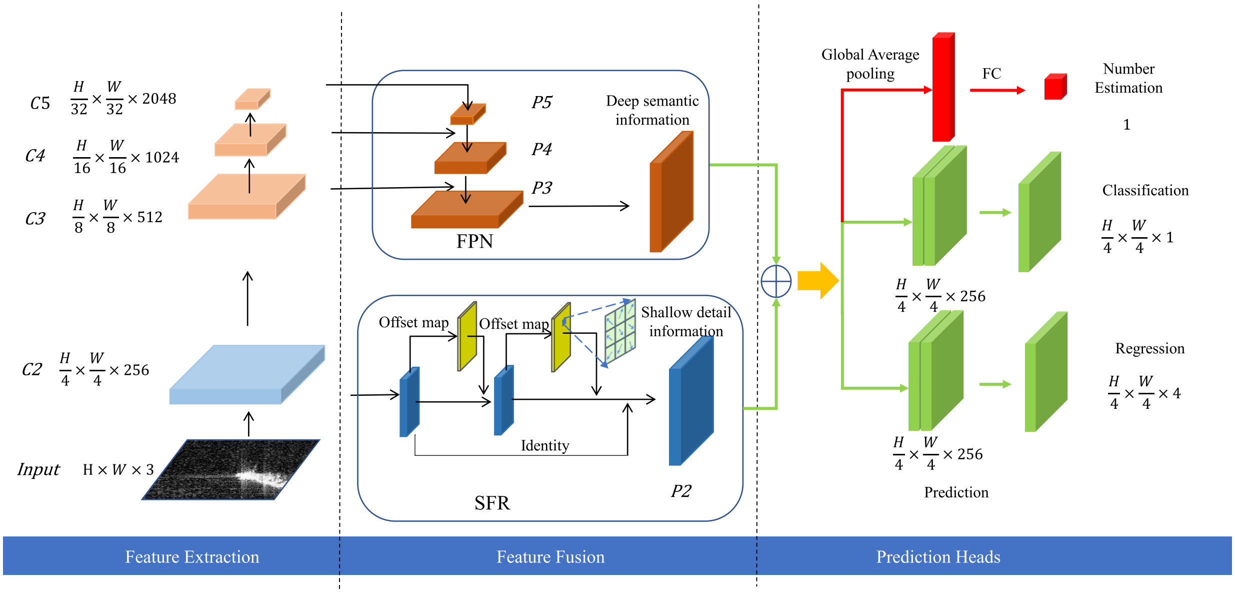

C5
$\frac{H}{32} \times \frac{W}{32} \times 2048$
C4
$\frac{H}{16} \times \frac{W}{16} \times 1024$
C3
$\frac{H}{8} \times \frac{W}{8} \times 512$
C2
$\frac{H}{4} \times \frac{W}{4} \times 256$
Input
H × W × 3
P5
P4
P3
Deep semantic information
FPN
Offset map
Offset map
Shallow detail information
Identity
SFR
P2
Global Average pooling
FC
Number Estimation
1
Classification
$\frac{H}{4} \times \frac{W}{4} \times 1$
$\frac{H}{4} \times \frac{W}{4} \times 256$
Regression
$\frac{H}{4} \times \frac{W}{4} \times 4$
$\frac{H}{4} \times \frac{W}{4} \times 256$
Prediction
Feature Extraction
Feature Fusion
Prediction Heads

Fig. 3. Network architecture. It consists of feature extraction, feature fusion, and prediction heads. The proposed SFR can improve the information expression in $P2$ coping with deformed complex scenes. The number estimation branch can provide adaptive guidance for filtering out low-quality detection boxes in different imaging conditions. Iteration-aware loss designed for regression branch enables the detector to locate the edge of the object more accurately in the aliased noise.

groupwise enhance (SGE) attention module to CenterNet [16], and enhanced the spatial features of each group. Based on the SSDD dataset [42], Jiao *et al.* [29] improved faster-RCNN [14] and proposed an end-to-end detection network for the ship detection problem of multiscale and multiscene SAR images. To obtain better feature maps, Fu *et al.* [47] proposed feature balance and refinement network (FBR-Net). To achieve multiscale and multiscene SAR ship detection, DCMSNN [29] builds a densely connected multiscale neural network.

CNN-based SAR object detectors are mostly adjusted from classic anchor-based detectors, which are slow and contain many manually adjusted hyperparameters. In this article, we combine anchor-free ideas to build a detector dedicated to SAR objects, which can detect SAR objects in complex backgrounds and noise more accurately.

III. Method

A. System Overview

The detector constructed in this article is one-stage and anchor-free. The overall architecture is shown in Fig. 3. It includes three parts: Feature extraction, feature fusion, and prediction heads. ResNet50 is used for feature extraction, and the extracted features are $C2, C3, C4,$ and $C5$. The scales of the four feature maps $C2, C3, C4,$ and $C5$ are 1/4, 1/8, 1/16, and 1/32, respectively, of the original image, corresponding to the information from shallow to deep in the SAR image. Feature extraction can extract the features of the objects hierarchically from the image, while suppressing the interference of background and noise. In feature fusion part, $C3, C4,$ and $C5$ are features fused by the classical FPN structure, and the shallow feature $C2$ are feature refined by SFR. The results of the two parts are summed as the output of the feature fusion part. Feature fusion can further optimize and improve the existing features, and more powerfully obtain the characteristics of the object in the SAR image. Based on feature map from feature fusion part, three prediction heads, number estimation, classification, and regression, are used to predict the location of the object center point, the shape of the detection box, and the number of objects in the SAR image.

In this section, we first introduce the feature extraction and fusion structure. We present the flow of prediction and postprocessing with number estimation in the second part. Iteration-aware loss is given at the end of this section.

B. Feature Extraction and Fusion

ResNet50 [48] is used as the backbone. Suppose the SAR image is expressed as $I \in R^{3 \times W \times H}$. When the resized image is feature extracted through ResNet50 [48], we can obtain the feature map of 1/4, 1/8, 1/16, and 1/32 down-sampled from the original image level by level. These feature maps are called $C_i \in R^{2^{6+i} \times W/2^i \times H/2^i}, i \in [2, 3, 4, 5]$. As the backbone deepens, the size of the feature map gradually decreases, and the interference information is gradually removed. The geometric deformation of the object and background caused by SAR imaging has a stronger influence on feature extraction. FPN and SFR are responsible for fusing the extracted features.

1) Traditional FPN: FPN can fuse the feature map with strong low-resolution semantic information. The structure of FPN is shown as in Fig. 3. On the left part, $C3, C4,$ and $C5$ are the features extracted by the backbone. The size of them is gradually reduced to half of the upper level, and the number of channels is doubled. On the right part, FPN sequentially processes the extracted features. $C5$ layer first is upsampled to change the size and undergoes 3×3 convolution to change channels to match $C4$. Subsequently, the obtained feature map is directly added to the $C4$ layer to obtain $P4$. Do this process again to get $P3$.

2) Shallow Feature Refinement: As shown in the left part of Fig. 4, the natural scene will undergo geometric deformation after SAR imaging, which will cause serious interference to the

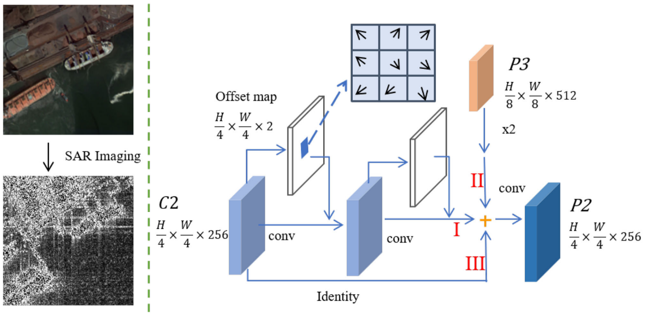

SAR Imaging
Offset map
$\frac{H}{4}\times\frac{W}{4}\times 2$
P3
$\frac{H}{8}\times\frac{W}{8}\times 512$
x2
C2
$\frac{H}{4}\times\frac{W}{4}\times 256$
conv
conv
II
conv
I
III
P2
$\frac{H}{4}\times\frac{W}{4}\times 256$
Identity

Fig. 4. SFR architecture. The deformed complex background in SAR imaging (as shown in left part) challenges the feature extraction and fusion. SFR merges the refined shallow features $C2 \in R^{W/4\times H/4\times 256}$: $I \in R^{W/4\times H/4\times 256}$, the upsampled and downchanneled $P3 \in R^{W/8\times H/8\times 256}$: $II \in R^{W/4\times H/4\times 256}$ and $C2$ itself: $III \in R^{W/4\times H/4\times 256}$ to generate the final feature $P2 \in R^{W/4\times H/4\times 256}$ to overcome the deformed complex background.

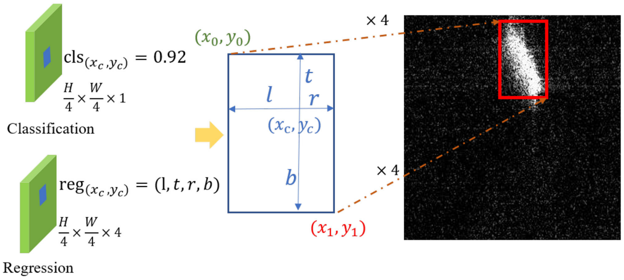

$\text{cls}_{(x_c,y_c)} = 0.92$
$\frac{H}{4}\times\frac{W}{4}\times 1$
Classification
$\text{reg}_{(x_c,y_c)} = (\text{l}, t, r, b)$
$\frac{H}{4}\times\frac{W}{4}\times 4$
Regression
(x_0, y_0)
t
l
r
(x_c, y_c)
b
(x_1, y_1)
× 4
× 4

Fig. 5. Prediction box decoding. The outputs of classification head and regression head are $\text{cls} \in R^{W/4\times H/4\times 1}$ and $\text{reg} \in R^{W/4\times H/4\times 4}$, respectively. Classification head predicts the probability $\text{cls}_{(x_c,y_c)}$ (for example, 0.92) that each pixel is an object center pixel by pixel. Regression head correspondingly predicts the distance $\text{reg}_{(x_c,y_c)}$ from this point to the four sides of the detection boxes. Enlarge the decoded detection boxes four times to get the final prediction results.

extraction of shallow features. We can fuse the deep features $C5$, $C4$, and and $C3$ to obtain a more representative feature map $P3$ through FPN. However, we cannot use the same method to process and merge the shallow feature $C2$ [17], [20], [49]. It contains more scenes interference information that is not filtered by the convolution layers.

Therefore, we design the SFR to cope with the geometric deformation of the complex background in the SAR imaging and refine the object feature, while suppressing the interference of the complex background in shallow $C2 \in R^{W/4\times H/4\times 256}$. The structure of the SFR is shown in Fig. 4. $P2 \in R^{W/4\times H/4\times 256}$ is derived from the addition of three features. In the first part, we predict the offset of each pixel $(\Delta x, \Delta y)$ in $C2$ when it is convolved to obtain offset map $\in R^{W/4\times H/4\times 2}$. When performing 3×3 convolution, the convolved pixels in $C2$ will be shifted accordingly according to Offset map (as shown in the enlarged part of Fig. 4). Repeat twice to get the first part of the feature map of $I \in R^{W/4\times H/4\times 256}$. The refined I feature can effectively remove interference, while retaining detailed features.

The second part is the upsampled $P3 \in R^{W/8\times H/8\times 512}$. The dimensionality of $P3$ is changed to $R^{W/4\times H/4\times 512}$ after upsampling, and then the channels are reduced from 512 to 256 by 3×3 convolution to match the dimensionality of $P2 \in R^{W/4\times H/4\times 256}$. The $II \in R^{W/4\times H/4\times 256}$ obtained after upsampling and downchanneling of $P3$ can enable $P2$ to obtain high-level semantic information, which is conducive to the detection of medium and large objects. The third part III is $C2 \in R^{W/4\times H/4\times 256}$ itself to help $P2 \in R^{W/4\times H/4\times 256}$ maintain the original feature of the SAR image.

Through the refinement of shallow features and the fusion of upper-level semantic features, SFR can effectively deal with the complex imaging background and geometric features of deformation in SAR images, and improve the feature extraction and fusion capabilities of detectors.

C. Prediction Heads

The features fused by FPN and SFR are processed through three detection heads to obtain the output of the network.

1) Classification and Regression Heads: Classification head makes a probability prediction of whether each pixel of $P2 \in R^{W/4\times H/4\times 256}$ contains the object and sets its output as $\text{cls} \in R^{W/4\times H/4\times n}$, where $W/4 \times H/4$ is the size of the feature map, and n is the number of categories in the dataset. The classification score of each pixel is between $[0, 1]$. Regression head assumes that each pixel in $P2$ is the center of the box, and predicts a 4-D vector $\text{reg} = (l, r, t, b)$ for each pixel, where $l, r, t,$ and b represent the distance from the object center to the left, right, top, and bottom of the detection box, respectively. Remember that the output of regression head is $\text{reg} \in R^{W/4\times H/4\times 4}$. Then, we can obtain the category and location of the predicted object, as shown in Fig. 5.

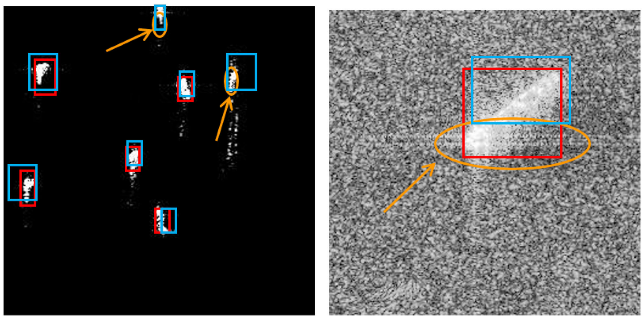
For each pixel, we predict the probability that it is an object center point $\text{cls}_{(x_c,y_c)}$, and the distance between the point and the four sides of the detection box $\text{reg}_{(x_c,y_c)} = (l, r, t, b)$. Combine it into a detection box for each pixel, where (x_0, y_0) are the coordinates of the upper-left corner of the prediction box, and (x_1, y_1) are the coordinates of the lower-right corner. Then enlarge its coordinates four times as the final detection result.

2) Number Estimation: Due to the changes in the elevation and azimuth angles of the SAR during imaging, the same background has differences under different imaging conditions. The confidence level of the detection boxes varies widely. As shown in Fig. 1(d), the confidence level of prediction boxes is from 0.6 to 0.8, and even the low-quality detection box C has a confidence of 0.65. Despite the similar background in Fig. 1(d), due to imaging differences, the confidence level of the boxes is 0.2 to 0.5, and the confidence for high-quality detection box E is only 0.45.

On the one hand, the selection of a fixed threshold is often based on manual experience. On the other hand, a fixed threshold is difficult to effectively filter out low-quality detection boxes in SAR images with imaging differences.

To alleviate the abovementioned problems, we use the number estimation head to predict the number of objects in each SAR image. The output of the number estimation head is num. We decode cls, reg, and num into the commonly used boxes and labels for object detection. The entire decoding process is shown in Algorithm 1.

We combine cls and reg to obtain all boxes and labels. The number estimation head predicts the number of objects num. cls passes through 3×3 max-pooling first. Then select the corresponding number of center points in cls_{max} according to num by $\text{topk}(\text{cls}, \text{num})$. Perform the same operation on reg to



(a) (b)

Fig. 6. Red boxes are the detection results of L1 Loss, and the blue boxes are the detection results of RIoU Loss. As shown in the orange regions, RIoU Loss can determine more objects in the early iterations (a) of training, while L1 Loss can obtain higher quality boxes in the later iterations (b). Iteration-Aware Loss adaptively combines L1 Loss and RIoU Loss according to their characteristics to locate the edge of the object accurately. (a) The result of the 5th epoch. (b) The result of the 65th epoch.

Algorithm 1: Bounding Boxes Decoding Algorithm.

```
Input: cls, reg, num.
Output: boxes, labels
Get $\text{cls}_{\text{max}}$ by performing a 3-by-3 maxpool on cls
$\text{cls} = \text{cls} * (\text{cls}_{\text{max}} == \text{cls})$
Get scores,cat, $ys, xs$ by performing topk(cls, num)
labels = (scores,cat)
Get $\text{reg}_{\text{max}}$ by performing topk(reg, num)
$boxes =$
    $(xs - \text{reg}[0], ys - \text{reg}[1], xs + \text{reg}[2], ys + \text{reg}[3])$
return boxes,labels
```

obtain reg_{max}. Finally, cls_{max} and reg_{max} are decoded to obtain the final detection result $(\text{boxes}, \text{labels})$.

D. Loss Function

1) Iteration-Aware Loss: The noise scattered power distribution in the SAR image is often close to the distribution of the object edge, which causes interference to the fine positioning of the SAR object. For this reason, we design iteration-aware loss to adaptively guide the anchor-free detector to focus on the approximate position of the object in the early training iterations and use more attention to locate the fine edge of the object in the later training iterations.

Fig. 6(a) and (b) shows the detection results of the detector using L1 loss and CIoU loss [50] at the 5 th and 65 th epoch. The red boxes are the prediction using L1 Loss, and the blue ones are the prediction using CIoU Loss [50]. It can be seen that in the early iterations of training, CIoU loss [50] can realize the quick determination of the approximate positions, and can focus the two objects missed by L1 loss [yellow circle in the upper right corner in (a)]. In the later iterations of training, as shown in Fig. 6(b), L1 loss can achieve more accurate detection. Detector tends to quickly locate the approximate positions of the objects in the early training iterations and then refine the existing detection boxes in the later iterations. A single loss function throughout the training iterations cannot adapt to the different optimization focuses of the network in the learning process.

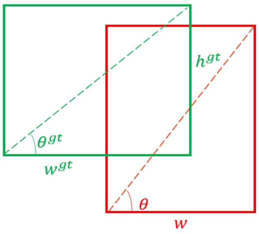


Fig. 7. Illustration of θ^{gt} and θ. The green box and the red one are ground-truth and prediction box, respectively.

To adapt the loss function to the different emphasis of the detector in different training iterations, we propose iteration-aware loss. In the initial stage of training, we design ratio IoU loss (RIoU Loss) to guide the detector to focus the object rough position

$$L_{\text{RIoU}} = 1 - \text{IoU} + \frac{v^2}{1 - \text{IoU} + v} \tag{1}$$

$$v = (\theta^{gt} - \theta)^2 \tag{2}$$

where L_{RIoU} guides the anchor-free detector to quickly determine the rough position of the SAR object in the early stage of training by constraining the degree of overlap between the prediction box and the ground-truth and the aspect ratio. And v measures the consistency of the aspect ratio. As shown in Fig. 7, θ^{gt} and θ are used to measure the aspect ratio of ground-truth and prediction box. $w^{gt}h^{gt}, w, h$ are their width and height. As the later training progresses, L1 loss takes over the position to guide the detector to focus on the refinement of boxes. The definition of iteration-aware loss is as follows:

$$L_{\text{Iteration-Aware}} = (1 - n/t) * L_r + n/t * L_1 \tag{3}$$

where n and t are the number of currently executed epoch and the total number of training epoch, respectively. l_1 and l_r are, respectively, L1 loss and RIoU loss. The optimization focus transition of the two loss functions in the training iterations can effectively adapt to the optimization focus of the detector in different iterations of learning to locate the edges of boxes.

Then we analyze why iteration-aware loss is effective. RIoU loss directly guides the detector to optimize IoU and aspect ratio. It is a measure of the relative error of the coincidence degree and shape similarity between the prediction box and ground-truth. In the early iterations of training, using RIoU loss as the loss function of the detector regression task can effectively guide the detector to predict the rough position of the object. In the later iterations of training, the optimization focus is to refine the predicted existing detection boxes in aliased scattered power distribution. The formula of L1 Loss is

$$L_1 = |l - l_0| + |r - r_0| + |t - t_0| + |b - b_0| \tag{4}$$

where (l, r, t, b) is the detector's prediction of the distance from the center point to the four sides of the box, and (l_0, r_0, t_0, b_0) is its corresponding target value. L1 Loss measures the prediction level by the absolute error between the box prediction value and the target value. As shown in Fig. 8, the large object has a larger L1 Loss than the small one, which is not conducive to the convergence of the model in the early training iterations when

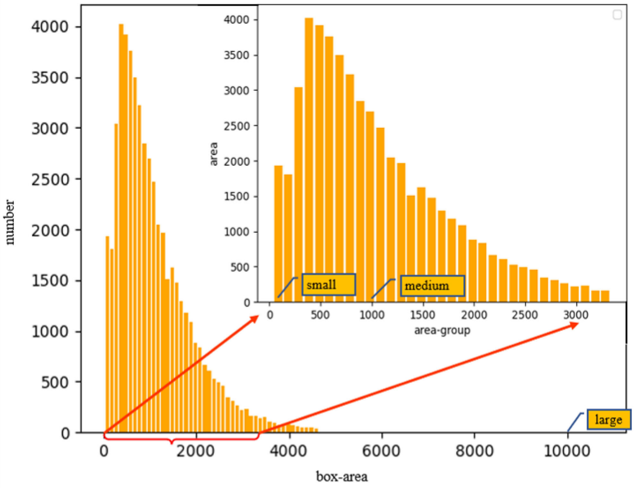


Fig. 8. Statistical analysis of box scale in SAR-ship dataset.

the prediction box accuracy is not high. In the later iterations of training, the L1 Loss can effectively measure the absolute loss of the predicted value and the target value, which is beneficial for the detector to strengthen the localization of the object edge while overcoming the aliased scattered power distribution.

Iteration-aware loss enables the detector to determine the approximate position of the object more quickly in the early iterations of training so that more attention can be used to refine the position of the box in the aliased power distribution in the later iterations of training.

2) Total Loss Function: We define total loss function L as follows:

$$L = w_c * L_{\text{cls}} + w_r * L_{\text{reg}} + w_n * L_{\text{num}} \tag{5}$$

where L_{cls} is focal loss [51], which is used to ease the imbalance of positive and negative samples of the dataset. L_{reg} uses iteration-aware loss, which is used to guide the detector's optimization focus in different iterations. For the number estimation function L_{num}, we use mean square error loss to measure the error in number estimation. For the above three tasks, the weight combination $[w_c, w_r, w_n]$ is [1, 1, 0.1].

IV. Experiment

To show the effectiveness of the proposed SAR object detector, we make comprehensive comparisons between our approach with mainstream optical image detection methods. Then we conduct ablation experiments to explore the effectiveness of SFR and iteration-aware loss for the detector in SAR image object detection. The visual analysis of number estimation is given at the end of this section.

A. Experiment Setup

1) Datasets: The SAR-ship dataset [52] contains more than 40 000 ship slices from the Gaofen No. 3 SAR. The image size is 256 × 256. Some examples are shown in Fig. 11. The green boxes are ground-truths of the objects, and the red ones are the prediction of the positions and categories of the objects, and the corresponding confidence scores are given. The initial data contain 102 Gaofen-3 pictures and 108 Sentinel-1 pictures. For Gaofen-3, the images have resolutions of 3, 5, 8, and 10 m with Ultrafine Strip-Map, Fine Strip-Map 1, Full Polarization 1, Full Polarization 2, and Fine Strip-Map 2 imaging modes, respectively. The data enhancement methods used include random center cropping, mirror inversion, and optical transformation. According to the definition of small object, medium object, and large object in the COCO dataset [53], we conduct statistical analysis on the three scale objects in the SAR-ship dataset [52], and the analysis results are shown in Fig. 9. It can be seen that small and medium objects dominate the SAR-Ship dataset [52], while large objects are rare.

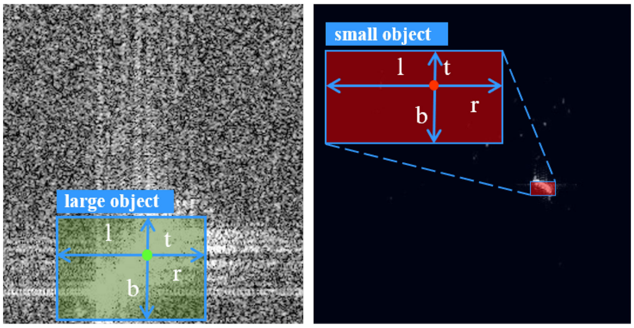


Fig. 9. Large box (left) has a larger L1 loss than the small box (right) when it is the same as the IoU of ground truths. It is not conducive for the detector to locate most of the objects on the SAR-ship dataset in the early iterations of training.

a) MSTAR-D dataset: The multiclass detection datasets of SAR objects are very scarce, in order to verify the performance of our method in multiclass SAR image object detection in more complex scenes, we construct the MSTAR-D dataset by performing manual instance-level annotation on the MSTAR image classification dataset [54]. The MSTAR-D dataset contains 5172 image slices of ten categories of SAR objects: BTR70 (armored transport vehicle), BMP2 (infantry fighting vehicle), T72 (tank), 2S1 (self-propelled howitzer), BRDM2 (armored reconnaissance vehicle), BTR60 (armored transport vehicle), D7 (bulldozer), T62 (tank), ZIL131 (cargo truck), and ZSU234 (self-propelled antiaircraft gun). As shown in Fig. 10, the objects of each category are very similar, and it is very challenging to accurately identify the class of objects and locate them.

2) Training and Evaluation Metrics: The ratio of training set: validation set:test set is 7:2:1 in SAR-Ship dataset [52] and MSTAR-D dataset. We use ResNet50 [48] pretrained on ImageNet as the backbone of the detector for feature extraction. The model uses the Adam optimizer to train with 48 images per batch. The initial learning rate is 1.25×10^{-4}, and a total of 70 epochs are trained. At the 45th, and 60th epoch, the learning rate is attenuated to the original 0.1. All experiments were performed on a device containing two 2080TI and CPU E5-2620 v4 @ 2.10 GHz. Similar to most publications, we use the average precision (AP) to evaluate the performance of our method. AP is the area of the precision/recall curve.

B. Comparison With Mainstream Detectors

We compare the proposed method with mainstream optical image object detectors: 1) Anchor-based one-stage object

TABLE I
PERFORMANCE COMPARISON WITH MAINSTREAM OPTICAL IMAGE DETECTORS AND SAR IMAGE DEDICATED OBJECT DETECTORS

	model	backbone	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
General optical detectors	Faster R-CNN [14]	ResNet50	5.19	53.4	91.9	56.3	50.0	57.6	19.9
	RetinaNet [52]	ResNet50	35.63	53.4	90.5	57.0	45.6	62.5	47.4
	FCOS [17]	ResNet50	12.81	56.7	92.0	63.2	49.3	65.3	58.7
	SSD [26]	ResNet50	60.4	56.7	93.4	62.9	51.1	63.3	58.7
	YOLO V3 [27]	DarkNet53	43.9	57.2	93.9	64.7	52.8	62.7	58.9
	FASF [56]	ResNet50	24.2	57.5	93.6	65.6	51.6	64.1	60.5
	CenterNet [16]	ResNet50	**72.7**	58.5	95.4	66.4	53.2	64.0	68.4
SAR dedicated detectors	CenterNet++ [57]	DLA	30.30	—	95.4	—	—	—	—
	MdrlEcf [58]	VGG16	5.05	—	91.7	—	—	—	—
	2S-Retinanet [59]	ResNet50	19.04	—	92.59	—	—	—	—
	ISASDNet [60]	ResNet50	—	60.1	95.3	65.2	**61.5**	58.2	54.4
	Quad-FPN [61]	ResNet50	11.37	—	94.39	—	—	—	—
	Ours	ResNet50	64.9	**62.0**	**96.4**	**72.0**	55.5	**68.4**	**69.6**

[01] RED/BLUE indicate the best/the second best.
The bold entities represent the maximum value of each column, that is, the best results obtained by each method on the index corresponding to that column.

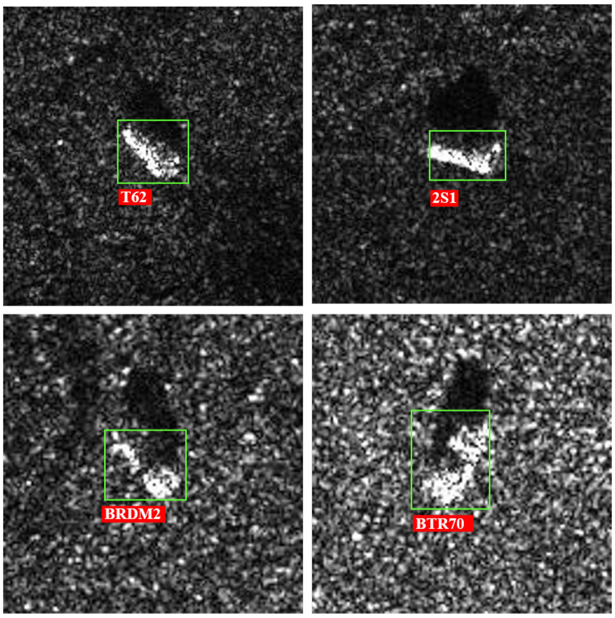


Fig. 10. Samples of MSTAR-D Datset. They are T62 (tank), 2S1 (self-propelled howitzer), BRDM2 (armored reconnaissance vehicle), BTR70 (armored transport vehicle), respectively.

detectors: YOLOv3 [27], FASF [55], SSD [26], and RetinaNet [51]. 2) Anchor-based two-stage object detectors: Faster R-CNN [14]. 3) Anchor-free object detectors: FCOS [17] and CenterNet [16]. The feature extraction network adopts ResNet50 [48] (the feature extraction network of YOLOv3 [27] is darknet53). Also, we compare with some SAR image dedicated object detectors: CenterNet++ [56], MdrlEcf [57], 2S-Retinanet [58], ISASDNet [59], and Quad-FPN [60].

As shown in Table I, anchor-based object detectors like faster R-CNN [14], RetinaNet [51], YOLO v3 [27], and SSD [26] achieve AP_{50} of 91.9%, 90.5%, 93.9%, and 93.4%, respectively. Anchor-free object detectors like FCOS [17] and CenterNet achieve AP_{50} of 92.0% and 95.4%, respectively. Our method can reach 62.0% AP and 96.4% AP_{50} at 64.9 frames per second (FPS). In the small, medium, and large scales, the performance of AP_S, AP_M, and AP_L exceeds CenterNet [16] 2.3%, 4.4%, and 5.2%, respectively. It can achieve the best speed-accuracy tradeoff in object detection tasks with SAR images with complex backgrounds. Compared with a SAR image dedicated object detector, our proposed method also has advantages. Among them, the AP_{50} of CenterNet++ [56], MdrlEcf [57], 2S-Retinanet [58], ISASDNet [59], and Quad-FPN [60] are 95.4%, 91.7%, 92.59%, 95.3%, and 94.39%, respectively. The AP_{50} of our proposed method is 96.5%, which exceeds 1.1%, 4.8%, 3.91%, 1.2%, and 1.11%, respectively.

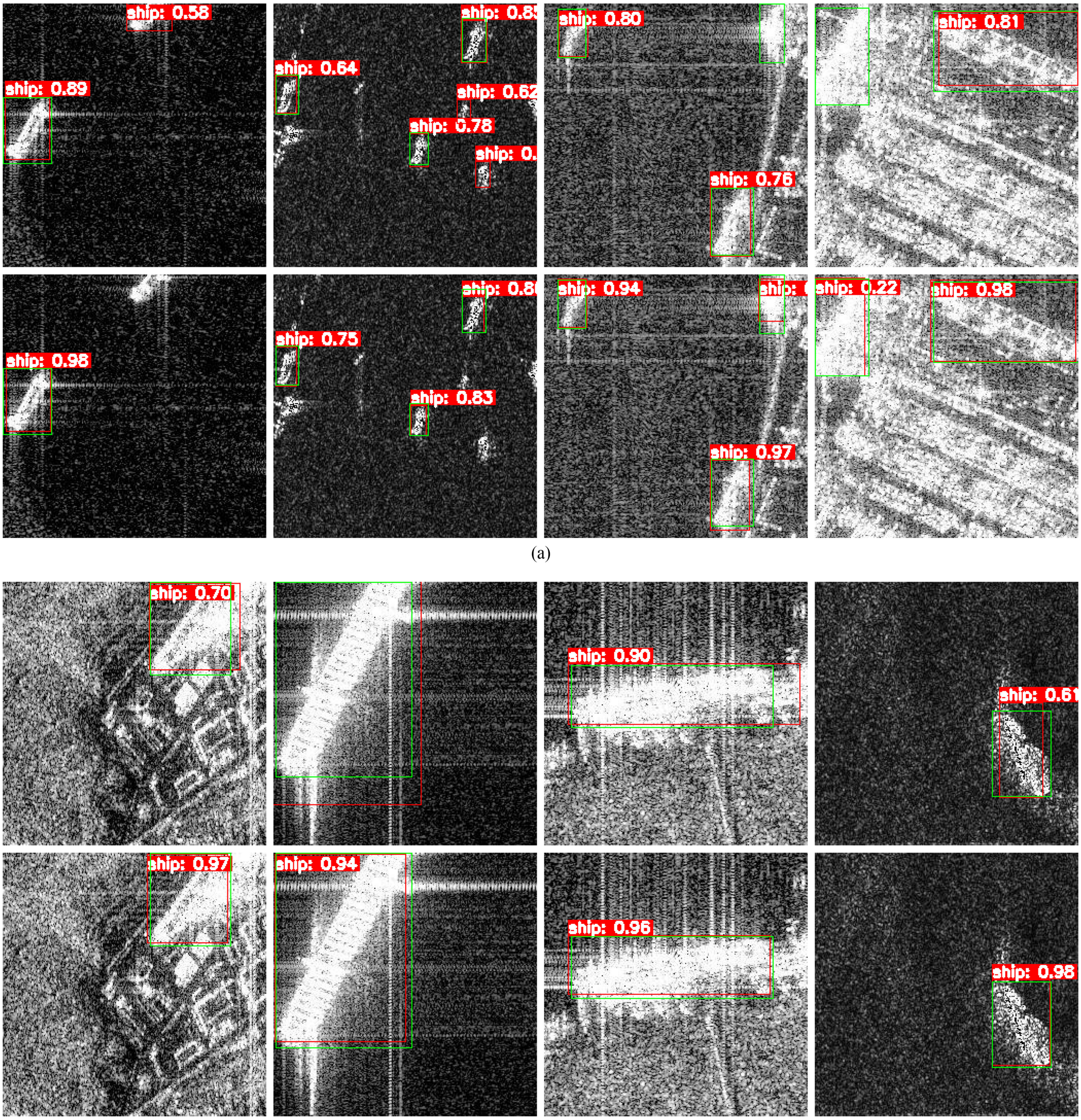
TABLE II
PERFORMANCE ON SINGLE-CLASS AND MULTICLASS SAR DATASETS

Dataset	Classes	Backbone	AP	AP_{50}	AP_{75}
SAR-Ship	1	ResNet50	62.9	96.4	72.0
MSTAR-D	10	ResNet50	67.7	93.4	80.1

In terms of speed, our detector also has a significant advantage. Compared to optical detectors, such as faster R-CNN [14], RetinaNet [51], FCOS [17], SSD [26], YOLO v3 [27], and FASF [55]'s 5.19, 35.63, 12.81, 60.4, 43.9, and 24.2 FPS, respectively, our method leads at 64.9 FPS. The SFR structure and number estimation slightly increase the amount of network parameters. Compared with CenterNet [16], our method sacrifices 8.2FPS speed in exchange for 3.5% AP improvement. Compared with SAR dedicated detectors, our method has an advantage over MdrlEcf [57], and Quad-FPN [60]'s 5.05 and 11.37 FPS at 64.9 FPS. Our method uses Resnet50, which is lighter than DLA [61], and has no offset heads, which is 34.6-FPS faster than CenterNet++ [56]'s 30.30 FPS. The removal of the anchor mechanism makes our method 45.86-FPS faster than anchor-based SAR dedicated detectors 2S-Retinanet [58]'s 19.04 FPS.

The final result is shown in Fig. 11; the green boxes are the ground-truths, and the red ones are the predicted boxes filtered by the number estimation method. In Fig. 11(a), we focus on comparing the results of the proposed method and CenterNet [16] on SAR images containing deformed complex scenes. In Fig. 11(b), we focus on comparing the performance of the proposed method and CenterNet [16] on objects when the edges of objects are aliased with the noise power distribution. Compared with CenterNet [16], our method can effectively detect SAR objects and locate accurate edges.

In order to further verify the advantages of our method in SAR object detection task, we conduct experiments on the constructed MSTAR-D dataset. As shown in Table II, our method



(a)

(b)

Fig. 11. Comparison between the proposed method and the most competitive CenterNet. As shown in (a), SFR can effectively deal with the deformed complex scenes in SAR imaging, identify difficult objects, and suppress false detections. Iteration-aware loss enables the detector to gradually pay attention to the edge information of the object during the training process, suppress the influence of the scattered power of the noise, to achieve more accurate box positioning. (a) The first row is the result of CenterNet, and the second row is the result of the proposed method with SFR. The red boxes and green ones are prediction boxes and ground-truths. SFR can effectively improve the detector's ability to cope with geometric deformations in SAR Imaging. (b) The first row is the result of L1 loss, and the second row is the result of iteration-aware loss. The red boxes and green ones are prediction boxes and ground-truths. Iteration-aware loss can effectively guide the detector to focus on the edges refinement of existing boxes.

achieves 62.9% at AP, 96.4% at AP_{50}, and 72.0% at AP_{75} on SAR-ship dataset. On the MSTAR-D dataset with ten classes of SAR objects, we achieve high performance with 67.7% at AP, 93.4% at AP_{50}, and 80.1% at AP_{75}.

The performance of various objects in the MSTAR-D dataset is shown in Table III. It can be seen that the performance of most classes, such as T62, BTR60, and ZSU234 can reach more than 70%. Some harder-to-identify classes, such as BRDM2, BTR70, and 2S1 achieve 45.7%, 52.7%, and 66.4% performance, respectively. Our method achieves good performance on various objects on the MSTAR-D dataset.

The visualization results are shown in Fig. 12. The green boxes are the ground-truths, and the red ones are the predicted boxes filtered by the number estimation method. It can be seen that the

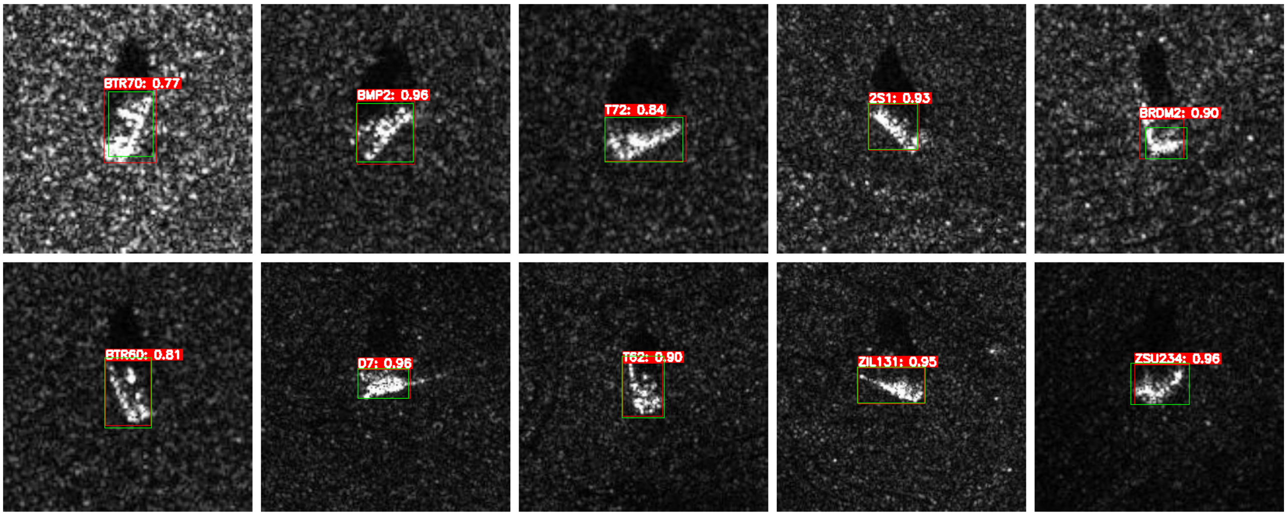


Fig. 12. Visualization results of our method in MSTAR-D. In the difficult case, where the features of different classes are similar, our method achieves accurate location localization and object classification.

TABLE III
PERFORMANCE OF PER-CLASS SAR OBJECTS ON MSTAR-D DATASET

Class	AP	Class	AP	Class	AP
2S1	66.4	BMP2	70.2	D7	81.4
T62	74.3	BRDM2	45.7	BTR70	52.7
BTR60	70.6	T72	71.3	ZIL131	73.6
ZSU234	70.6				

TABLE IV
ABLATION EXPERIMENTS UNDER DIFFERENT CONFIGURATIONS

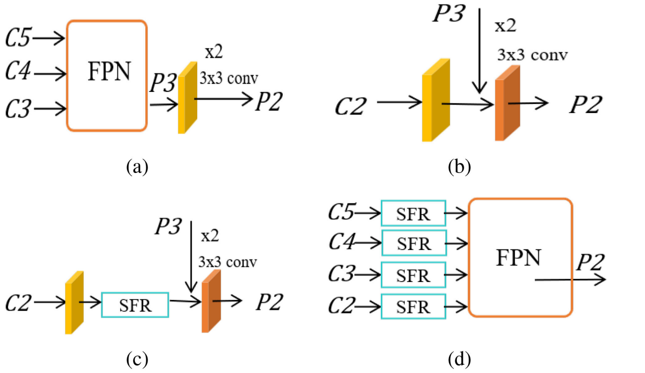
SFR	Iteration-Aware Loss	AP	AP_S	AP_M	AP_L
		58.47	53.20	64.00	68.40
✓		58.80	53.70	64.50	63.00
	✓	61.87	55.33	**68.43**	67.97
✓	✓	**62.01**	**55.56**	68.42	**69.64**

The bold entities represent the maximum value of each column.

features of BTR70, BTR60, and other objects in MSTAR-D are very close, and our method can accurately classify the positions of the objects, while recognizing them.

C. Ablation Experiments

SFR structure and iteration-aware loss make the anchor-free detector more suitable for object detection in complex scenes. To analyze their influence on the detector, we conduct a series of ablation experiments under the same experimental settings. All ablation experiments are briefly reported in Table IV. SFR and iteration-aware loss can increase AP by 0.33% and 3.40%, respectively, based on the baseline. On small objects, the two can increase AP_S by 0.5% and 2.55%, respectively. The improvement of the detector on SAR-ship [52] is mainly due to that SFR refines the shallow features and iteration-aware loss adapts to the optimization focus of the detector in different training iterations. When both SFR and iteration-aware loss are used, the detector achieves the best performance with a AP of 62.01%.



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

Fig. 13. Comparison of four structures for refining feature. (a) Is the original FPN. (b) Is the direct introduction of $C2$, unsuppressed noise corrupts feature. (c) Is the use of SFR to $C2$, it suppresses noise interference while retaining detailed information in $C2$. (d) Is the use of SFR to $C2 \sim C5$, multiple SFR structures can further improve the performance but with an inappropriate computation.

1) Influence of SFR: As shown in Fig. 13, we compare the detection performances of $P2$ feature maps generated in four ways as follows.

a) $P2 \in R^{W/4\times H/4\times 256}$ generated by the basic FPN. $P3 \in R^{W/8\times H/8\times 512}$ is obtained through the classic FPN structure, followed by upsampling to double the size of the feature map and 3×3 convolution to reduce the number of channels to obtain $P2$.

b) $P3$ is performed upsampling and 3×3 convolution as a), and then directly added with original $C2 \in R^{W/4\times H/4\times 256}$ to get $P2$.

c) $P3$ is performed operations as a). Then use the SFR module to refine shallow feature $C2$ and then merge them to generate $P2$.

d) The features $C2{\in}R^{W/4\times H/4\times 256}$, $C3{\in}R^{W/8\times H/8\times 512}$, $C4 \in R^{W/16\times H/16\times 1024}$, and $C5 \in R^{W/32\times H/32\times 2048}$ output by the backbone are all refined using SFR, and then the refined features are fused using the classic FPN structure to obtain the final $P2$.

TABLE V
ABLATION EXPERIMENT OF SFR. DETAILS ABOUT THE FOUR STRUCTURE (A)–(D) ARE GIVEN IN FIG. 13

backbone	Structure	AP	AP_S	AP_M	AP_L
ResNet50	(a)	58.47	53.20	64.00	68.40
ResNet50	(b)	57.78	52.80	63.10	69.90
ResNet50	(c)	58.80	**53.70**	64.50	63.00
ResNet50	(d)	**58.87**	52.86	**65.06**	**71.35**

The bold entities represent the maximum value of each column.

Ablation experiments show that SFR can effectively remove the interference information in the shallow feature map $C2$, and refine the features about small objects. The performance comparison of the four methods is shown in Table V. The direct introduction of $C2$ will damage the final performance of the detector. The performance of AP_S and AP_M are reduced by 0.4% and 0.9%, respectively, and the performance of AP_L is increased by 1.5%. As shown in Fig. 9, the SAR-Ship dataset [52] is mainly based on small and medium objects, so the performance improvement of large objects has little effect on the final performance AP. After using SFR to refine the shallow feature map $C2$, the detector can achieve better performance. Compared with the FPN method, the SFR method exceeds 0.50% on both small and medium objects, and the weakening of the performance on large objects has little effect on the overall performance of the detector. Comparing method (c) and method (d), it can be found that when the deep features $C3$, $C4$, and and $C5$ are refined, the performance is only improved by 0.07% as the calculation increases. Finally, compared to the other three methods, (c) achieves competitive performance with an AP of 58.80%.

Through SFR, the proposed detector can perform better feature extraction and fusion of deformed complex scenes in SAR images. CenterNet [16] has misdetected small objects in the first row in Fig. 11(a). Especially in the first and second columns in Fig. 11(a), after the shallow features are refined by SFR, it can effectively suppress false detection. In addition, as shown in the third and fourth columns, the information of objects is submerged in the deformed scenes, and SFR can effectively identify and retain its characteristics. The SFR module can effectively remove the deformed scenes features in $C2$, while retaining the detailed information of small objects, thereby effectively improving the performance of the detector on SAR objects in complex scenes.

2) Influence of Iteration-Aware Loss: To locate the edge of SAR object with aliased scattered power distribution, we design iteration-aware loss to adapt to optimization focus in different iterations. In the ablation experiment, we compare iteration-aware loss with CIoU loss [50], L1 loss, and the weighted combination of the two (both weight 0.5). The experimental results are shown in Table VI.

It can be seen that L1 loss is significantly better than CIoU loss [50]. The performance of small objects, medium objects, and large objects are, respectively, 1.30%, 3.81%, and 0.91% higher. Using L1 loss as the loss function of regression can build a powerful SAR object detection performance benchmark. The performance of the weighted combination of L1 loss and CIoU loss [50] is between using the two loss functions alone, and what is surprising is that iteration-aware loss is obtained by combining CIoU loss [50] and L1 loss according to the dynamic weight, and has the best performance. iteration-aware loss exceeds L1 Loss by 1.27%, 0.56%, 1.63% in the performance of small objects, medium objects, and large objects, respectively.

TABLE VI
ABLATION EXPERIMENT OF ITERATION-AWARE LOSS

backbone	Loss	AP	AP_S	AP_M	AP_L
ResNet50	CIoU Loss	58.47	53.18	63.97	68.4
ResNet50	IAL	**61.96**	**55.75**	**68.32**	**70.94**
ResNet50	L1 Loss	61.16	54.48	67.76	69.31
ResNet50	CL	61.08	54.57	67.41	69.36

^{01} IAL refer to iteration-aware loss.
^{02} CL refer to the combination of CIoU Loss [50] and L1 Loss.
The bold entities represent the maximum value of each column.

TABLE VII
ABLATION EXPERIMENT OF NUMBER ESTIMATION

method	T or K	AP	AP_{50}
fixed threshold ($T = t$)	0.01	61.6	**95.6**
	0.1	61.3	94.8
	0.2	61.2	94.8
	0.3	60.8	93.9
	0.4	60.5	93.0
	0.5	60.1	92.1
	0.6	59.4	91.1
	0.7	58.1	88.3
	0.8	54.2	81.5
	0.9	37.5	53.2
number estimation ($K = num$)	num	61.2	94.5

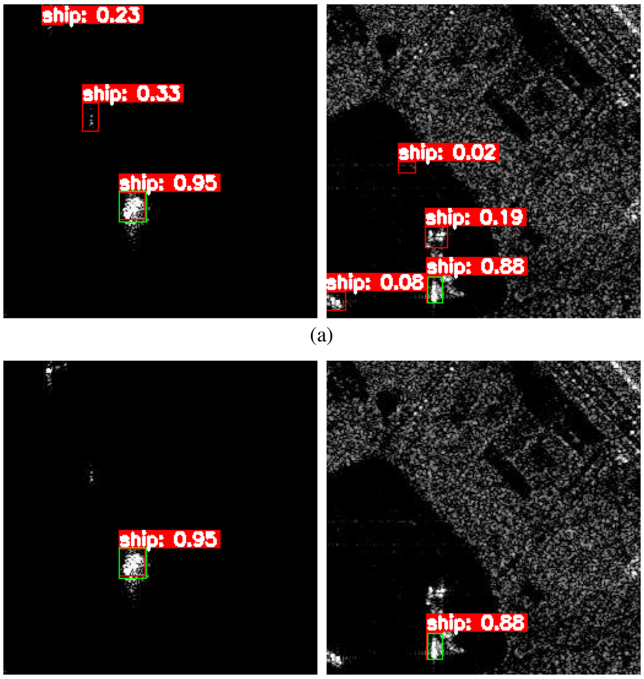
The bold entities represent the maximum value of each column.

Comparing the first row and the second in Fig. 11(b), the method proposed in this article can achieve more accurate boxes positioning on objects. Iteration-aware loss can effectively guide the detector to focus on the refinement of the detection boxes in the later stage of training. Especially on the right side of the first column, the bottom side of the second column, the left side of the third column, and the top side of the last column, the proposed method achieves more accurate box positioning.

D. Visual Analysis for Number Estimation

In practical applications, it is often necessary to select boxes generated by the detector according to a certain standard. But for scenes with similar backgrounds, SAR images sometimes will be different due to different conditions, such as the elevation angle of SAR imaging. We propose the number estimation method to deal with the resulting difference in the confidence level of the prediction boxes in different SAR images.

In practical applications, the confidence threshold is usually not too high or too low. In this part, AP_{50} of boxes selected by the confidence threshold method ($T = t, t = 0.1, 0.2, \ldots, 0.9$) and the method based on number estimation ($K = \text{num}$) is compared. The experimental results are shown in Table VII. In practical applications, the confidence threshold is usually 0.5. The AP_{50} based on number estimation ($K = \text{num}$) is 2.4% higher than that based on confidence threshold ($T = 0.5$). As the threshold decreases, more boxes are involved in the calculation of the AP_{50}. When the threshold is 0.7, 0.6, 0.5, 0.4, and AP_{50}



(a)

(b)

Fig. 14. Comparison of visualization effects when AP_{50} is 96.1% and 92.1%. The red boxes and the green ones are the prediction boxes and ground-truths, respectively. (a) AP_{50} is 95.6% when the threshold is 0.01. (b) AP_{50} is 92.1% when the threshold is 0.50.

is 88.3%, 91.1%, 92.1%, and 93.0% in turn. However, the higher the AP_{50} does not mean the better the effect in actual application.

As shown in the first row of Fig. 14(a), when the threshold is 0.01, many low-confidence prediction boxes fill the entire image, but the AP_{50} is as high as 95.6%. Fig. 14(b) is the result when threshold is 0.5. It can be seen that the low-confidence boxes are filtered out. However, the AP_{50} dropped from 95.6% to 92.1%.

We compare the difference between the two methods based on the confidence threshold method and the number estimation method in the detection of objects in SAR images. In Fig. 15, the green boxes are the ground-truths of the ship objects in the SAR image, and the red ones are the prediction boxes of the detector. It can be seen that in SAR images, using threshold cannot effectively detect the ships and objects in the first column of Fig. 15. In contrast, the number of ships in the second column is 2, 3, 1, and 1. The corresponding number of detection boxes is selected as the final detection result. Thus, low-confidence boxes with confidence of 0.21, 0.48, 0.57, and 0.67 are detected, respectively.

Based on the above analysis, number estimation has a stronger adaptive ability than using a fixed threshold. It does not need to be tried one by one or based on manual experience and can complete the process of the model from training to actual application end to end. In SAR images with different imaging conditions, more difficult and low-confidence objects can be effectively detected.

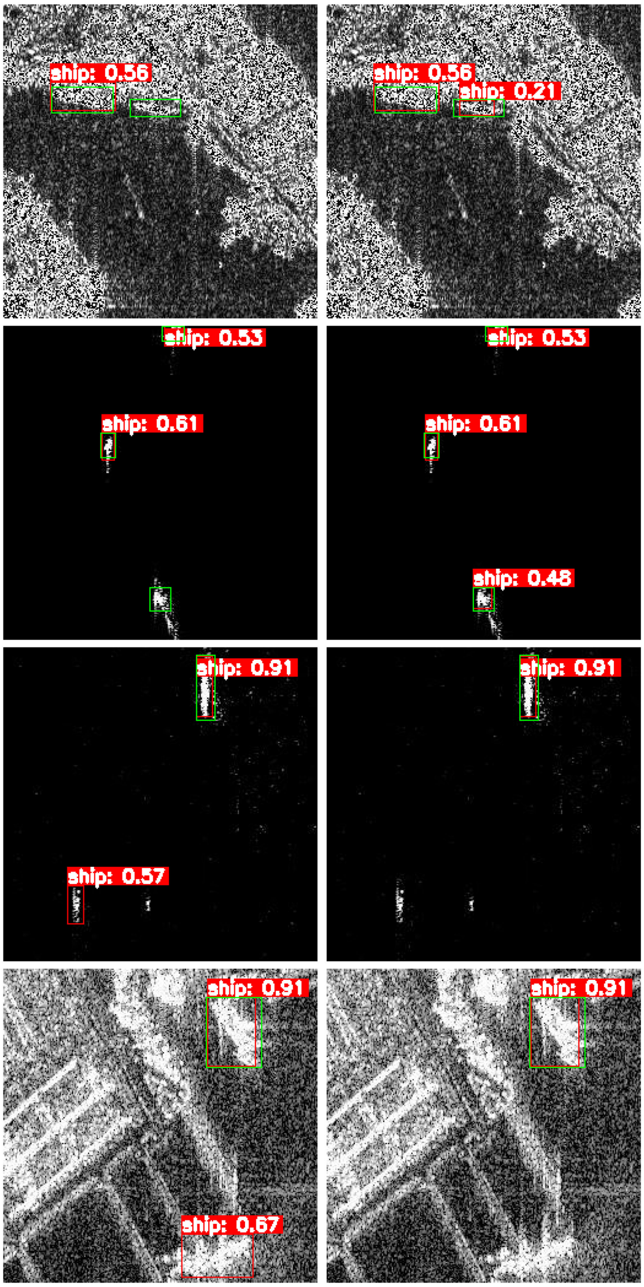


Fig. 15. Comparison of the filtered results based on threshold (column 1) and on number estimation (column 2). The red boxes and the green ones are the prediction boxes and ground-truths, respectively.

V. Conclusion

In this article, we develop the anchor-free method for object detection in SAR images under deformed complex background and aliased noise power distribution. A wide range of experiments show the effectiveness of our proposed method, which achieves 96.4% AP_{50} at 64.9 FPS. The proposed SFR can effectively enhance the detector's ability to extract and fusion shallow features in deformed complex scenes. The iteration-aware loss can effectively adapt to changes in the detector's optimization focus at different training iterations to locate the edge of SAR object more accurately. And number estimation provides a more effective method to filter out low-quality detection boxes in different imaging conditions. We hope that these insights may be useful for other SAR image detection tasks in deformed complex scenes and noise power distribution.

REFERENCES

[1] R. D. West *et al.*, "Polarimetric SAR image terrain classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 11, pp. 4467–4485, Nov. 2019.

[2] M. Ahishali, T. Ince, S. Kiranyaz, and M. Gabbouj, "Performance comparison of learned vs. engineered features for polarimetric SAR terrain classification," in *Proc. Photon. Electromagn. Res. Symp.-Spring*, 2019, pp. 2317–2324.

[3] V.-E. Neagoe, S.-V. Carata, and A.-D. Ciotec, "An advanced neural network-based approach for military ground vehicle recognition in SAR aerial imagery," *Int. Sci. Committee*, vol. 18, pp. 41–48, 2016.

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

[5] C. Chen, C. He, C. Hu, H. Pei, and L. Jiao, "A deep neural network based on an attention mechanism for SAR ship detection in multiscale and complex scenarios," *IEEE Access*, vol. 7, pp. 104848–104863, 2019.

[6] Y. Alebele *et al.*, "Estimation of crop yield from combined optical and SAR imagery using Gaussian kernel regression," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10520–10534, Oct. 2021.

[7] A. V. Etten, D. Lindenbaum, and T. M. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," *CoRR*, vol. abs/1807.01232, 2018.

[8] W. An, M. Lin, and H. Yang, "Stationary marine target detection with time-series SAR imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6406–6413, Jun. 2021.

[9] W. Bao, M. Huang, Y. Zhang, Y. Xu, X. Liu, and X. 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.

[10] H. Bi, J. Deng, T. Yang, J. Wang, and L. Wang, "CNN-based target detection and classification when sparse SAR image dataset is available," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6815–6826, Jun. 2021.

[11] S. Chen, R. Zhan, W. Wang, and J. Zhang, "Learning slimming SAR ship object detector through network pruning and knowledge distillation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1267–1282, Dec. 2021.

[12] J. Zhao, W. Guo, Z. Zhang, and W. Yu, "A coupled convolutional neural network for small and densely clustered ship detection in SAR images," *Sci. China Inf. Sci.*, vol. 62, no. 4, 2019, Art. no. 42301.

[13] H. Chen, Z. Liu, W. Guo, Z. Zhang, and W. Yu, "Fast detection of ship targets for large-scale remote sensing image based on a cascade convolutional neural network," *J. Radars*, vol. 8, no. 3, pp. 413–424, 2019.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[15] 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, 2017, Art. no. 860.

[16] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," *CoRR*, vol. abs/1904.08189, 2019.

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

[18] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[19] T. Y. Lin, P. Dollar, 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.

[20] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," *CoRR*, vol. abs/1911.09516, 2019.

[21] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9756–9765.

[22] Y. Xue, Y. Li, S. Liu, X. Zhang, and X. Qian, "Crowd scene analysis encounters high density and scale variation," *IEEE Trans. Image Process.*, vol. 30, pp. 2745–2757, Jan. 2021.

[23] Y. Xue, Y. Li, S. Liu, P. Wang, and X. Qian, "Oriented localization of surgical tools by location encoding," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 4, pp. 1469–1480, Apr. 2022.

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

[25] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 821–830.

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

[27] A. Farhadi and J. Redmon, "YOLOv3: An incremental improvement," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, 1–6.

[28] X. Li, S. Lai, and X. Qian, "DBCFace: Towards PURE convolutional neural network face detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1792–1804, Apr. 2022.

[29] 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.

[30] D. Cozzolino, G. Di Martino, G. Poggi, and L. Verdoliva, "A fully convolutional neural network for low-complexity single-stage ship detection in sentinel-1 SAR images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 886–889.

[31] X. Wang, S. Lai, Z. Chai, X. Zhang, and X. Qian, "SPGNet: Serial and parallel group network," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2021.3088639.

[32] D. W. Greig and M. Denny, "Knowledge-based methods for small-object detection in SAR images," *Proc. SPIE*, vol. 4883, pp. 121–130, 2003.

[33] C. Tison, J.-M. Nicolas, F. Tupin, and H. Maître, "A new statistical model for Markovian classification of urban areas in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 10, pp. 2046–2057, Oct. 2004.

[34] K. Fu, J. Fu, Z. Wang, and X. Sun, "Scattering-keypoint-guided network for oriented ship detection in high-resolution and large-scale SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11162–11178, Sep. 2021.

[35] Y. He, F. Gao, J. Wang, A. Hussain, E. Yang, and H. Zhou, "Learning polar encodings for arbitrary-oriented ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3846–3859, Mar. 2021.

[36] Z. Hong *et al.*, "Multi-scale ship detection from SAR and optical imagery via a more accurate YOLOv3," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6083–6101, Jun. 2021.

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

[38] J. Tu, F. Gao, J. Sun, A. Hussain, and H. Zhou, "Airport detection in SAR images via salient line segment detector and edge-oriented region growing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 314–326, Nov. 2021.

[39] P. Spudis *et al.*, "Mini-SAR: An imaging radar experiment for the Chandrayaan-1 mission to the Moon," *Curr. Sci.*, vol. 96, no. 3, pp. 533–539, 2009.

[40] 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, no. 6, pp. 852–862, 2019.

[41] Y. Yang, Y. Qiu, and C. Lu, "Automatic target classification: Experiments on the MSTAR SAR images," in *Proc. 6th ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw. Parallel/Distrib. Comput.*, L. Chung and Y. Song, Eds., 2005, pp. 2–7.

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

[43] Y. Yu, B. Wang, and L. Zhang, "Pulse discrete cosine transform for saliency-based visual attention," in *Proc. IEEE 8th Int. Conf. Develop. Learn.*, 2009, pp. 1–6.

[44] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 155–162.

[45] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[46] X. Wang, Z. Cui, Z. Cao, and S. Dang, "Dense docked ship detection via spatial group-wise enhance attention in SAR images," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1244–1247.

[47] 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.

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

[49] T. Lin, P. Goyal, R. B. 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.

[50] Z. Zheng, P. Wang, W. Liu, J. Li, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12993–13000.

[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[52] 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.

[53] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[54] E. R. Keydel, S. W. Lee, and J. T. Moore, "Mstar extended operating conditions: A tutorial," *Proc. SPIE*, vol. 2757, pp. 228–242, 1996.

[55] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 840–849.

[56] H. Guo, X. Yang, N. Wang, and X. Gao, "A centernet model for ship detection in SAR images," *Pattern Recognit.*, vol. 112, 2021, Art. no. 107787.

[57] S. Liu and J. Tang, "Modified deep reinforcement learning with efficient convolution feature for small target detection in VHR remote sensing imagery," *ISPRS Int. J. Geo- Inf.*, vol. 10, no. 3, 2021, Art. no. 170.

[58] F. Gao, W. Shi, J. Wang, E. Yang, and H. Zhou, "Enhanced feature extraction for ship detection from multi-resolution and multi-scene synthetic aperture radar (SAR) images," *Remote Sens.*, vol. 11, no. 22, 2019, Art. no. 2694.

[59] Z. Wu, B. Hou, B. Ren, Z. Ren, and L. Jiao, "A deep detection network based on interaction of instance segmentation and object detection for SAR images," *Remote Sens.*, vol. 13, no. 13, 2021, Art. no. 2582.

[60] T. Zhang, X. Zhang, and X. Ke, "Quad-FPN: A novel quad feature pyramid network for SAR ship detection," *Remote Sens.*, vol. 13, no. 14, 2021, Art. no. 2771.

[61] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2403–2412.

Yawei Zhang received the B.S. degree in information engineering from the Xi'an Jiaotong University of Technology, Xi'an, China, in 2018. He is currently working toward the M.S. degree in information and communication engineering with Xi'an Jiaotong University, Xi'an, China.

His research interests include object detection, radar image interpretation, and image processing.

Yu Cao received the B.S. degree in electrical engineering and automation, and the M.S. degree in power electronics and power transmission from Northwestern Polytechnical University, Xi'an, China, in 2014 and 2017, respectively. He is currently working toward the Ph.D. degree in signal and information processing with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.

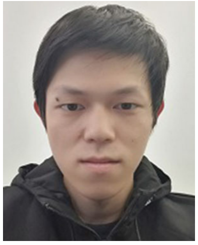
His research interests include photoelectric tracking control and remote sensing image processing.

Xubin Feng received the B.S. degree in computer science and technology, and the M.S. degree in computer architecture from Xidian University, Xi'an, China, in 2010 and 2013, respectively, and the Ph.D. degree in communication and information system from the Xi'an Institute of Optics and Precision Mechanics (XIOPM), Chinese Academy of Sciences (CAS), Xi'an, China, in 2021.

He is currently an Associate Researcher owiht XIOPM, CAS. His research interests include photoelectric tracking control and remote sensing image processing.

Meilin Xie received the B.S. degree in control engineering from the Lanzhou University of Technology, Lanzhou, China, in 2011, the M.S. degree in control engineering from Northwestern Polytechnical University, Xi'an, China, in 2014, and the Ph.D. degree in signal and information processing from the Xi'an Institute of Optics and Precision Mechanics (XIOPM), Chinese Academy of Sciences (CAS), Xi'an, China, in 2020.

He is currently an Associate Researcher and a master Supervisor with XIOPM, CAS. He is also the Deputy Director of the Photoelectric Tracking and Measurement Technology Research Laboratory, XIOPM, CAS. His research interests include photoelectric tracking control and remote sensing image processing.



Xin Li received the B.S. degree in information engineering from the Xi'an Jiaotong University of Technology, Xi'an, China, in 2018. He is currently working toward the M.S. degree in information and communication engineering with Xi'an Jiaotong University, Xi'an, China.

His research interests include object detection, face detection, and image processing.

Yao Xue received the M.Eng. degree from the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2013, and the Ph.D. in computer science from the Department of Computer Science, University of Alberta, Canada, in 2018.

He is currently a Lecturer with the School of Information and Communications Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China. He is also with the Centre for Intelligent Mining Systems Laboratory, University of Alberta, under the supervision of Dr. N. Ray. His research interests include cover computer vision related areas, like image retrieval, visual summary, object detection, and recognition.

Xueming Qian (Member, IEEE) received the B.S. degree in automation and the M.S. degree in image processing and pattern recognition from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree in signal and information processing from the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, in 2008.

Since 2008, he has been teaching with the Department of Information and Communication Engineering, School of Telecommunications, Xi'an Jiaotong University, where he became an Assistant Professor in 2008, an Associate Professor in 2011, and a Full Professor in 2014. He is currently with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, School of Information and Communication Engineering, and SMILES LAB, Xi'an Jiaotong University. His current research interests include social media, big data, mining, and search.