ROS-Det: Arbitrary-Oriented Ship Detection in High Resolution Optical Remote Sensing Images via Rotated One-Stage Detector

MINGMING ZHU 1, GUOPING HU 2, SHUAI LI 3, HAO ZHOU 2, SHIQIANG WANG 2, YULE ZHANG 1, AND SHIJIE YUE 1

1 Graduate College, Air Force Engineering University, Xi’an 710051, China
2 Air and Missile Defense College, Air Force Engineering University, Xi’an 710051, China
3 Aeronautics Engineering College, Air Force Engineering University, Xi’an 710038, China

Corresponding author: Guoping Hu (hgp_kgd123@163.com)

This work was supported in part by the Natural Science Foundation Research Project of Shaanxi Province, China, under Grant 2020JM-345.

ABSTRACT To address these problems, namely, arbitrary orientations, various sizes, and dense distributions of ship detection, we propose an arbitrary-oriented ship detection method via rotated single-stage detector (ROS-Det), which integrates a feature pyramid network (FPN) based on an improved ResNet50, rotated anchors, a classification network, and a regression network together. Firstly, to improve robustness against various sizes of ships, the FPN is used to fusion multiscale convolutional feature maps. Through several tweaks, the improved ResNet50 can receive more information and reduce the computational cost. Secondly, for the purpose of arbitrary-oriented ship detection, rotated anchors, skew intersection over union (IoU), and skew non-maximum suppression (NMS) are introduced to RetinaNet. Then, on account of the disadvantages that the arbitrary-oriented object detection methods usually cause loss discontinuity problem, we improve the traditional smooth L1 loss function by introducing an IoU constant factor. Finally, based on several techniques such as data augmentation and transfer learning, we achieve ship detection on a public ship dataset HRSC2016. Through comparison experiments, we have analyzed and discussed the validity of our proposed ROS-Det, which achieves the state-of-the-art performance.

INDEX TERMS Ship detection, feature pyramid network, data augmentation, transfer learning.

I. INTRODUCTION

With the development of aerospace technologies and sensors, a number of high-resolution and high-quality remote sensing images have been developed to facilitate object detection [1]–[3]. Ship detection has played a vital role in many fields, such as harbor dynamic surveillance, traffic monitoring, maritime management, and national defense construction.

Traditional ship detection methods mainly consist of feature extraction and classification recognition. For example, voting-based pose estimation methods [4], deep forest ensemble-based methods [5], Haar-like feature-based methods [6] and Fourier HOG-based methods [7]. Most of these methods rely heavily on manual features and sliding windows, which results in causing low accuracy and high time cost.

In 2014, Girshick et al. [8] introduced an iconic model in the object detection field: a region-based convolutional neural network (R-CNN). To further reduce the time cost, an improved R-CNN algorithm, i.e., Fast R-CNN, was proposed [9]. Since the above two models obtain candidate regions through artificial methods, a speed bottleneck exists. Faster R-CNN [10] replaces the selective search (SS) algorithm with a region proposal network (RPN). At the same time, the RPN and the subsequent detection network share the same convolutional layers, which greatly reduces the detection time. To solve the multiscale problem, a feature pyramid network (FPN) [11] was proposed, which can be applied to Faster R-CNN. The above methods are representative two-stage detectors, and there are other methods, e.g., Mask R-CNN [12], region-based fully convolutional network (R-FCN) [13], and Cascade R-CNN [14].
Different from two-stage detectors, one-stage detectors simultaneously complete two tasks, i.e., classification and localization. You Look Only Once (YOLO) [15], Single Shot MultiBox Detector (SSD) [16], and RetinaNet [17] are representative one-stage detectors. Specifically, YOLO simultaneously completes feature extraction, object classification and positioning tasks based on CNNs, thereby greatly improving detection speed. After dividing the image into S×S grids, YOLO directly predicts the location and category for each object. SSD has an improved ability to detect small objects. RetinaNet introduces a new loss function, i.e., the focal loss, thereby significantly improving the detection accuracy.

In recent years, many ship detection algorithms and models based on deep learning have been proposed. For example, You et al. [18] first used the scene mask extraction network for scene segmentation, and then excluded false targets based on the feature map and inferred scene mask. Shao et al. [19] achieved ship detection by using CNNs as well as saliency detection. They employed comprehensive ship features, e.g., deep features, saliency maps, and coastline priors. Wu et al. [20] introduced a coarse-to-fine idea into ship detection.

The above-mentioned methods are known as horizontal region detection methods. However, there are three main disadvantages. 1) Horizontal bounding boxes (HBBs) cannot reflect the real shape of ships with various orientations, as shown in Fig. 1(a). 2) As shown in Fig. 1(b), object and background pixels cannot be effectively separated, especially for ships with a large aspect ratio. 3) As a result of the large overlap between bounding boxes, dense ships may be missed, as shown in Fig. 1(c). Compared with HBBs, rotated bounding boxes (RBBs) can reflect the physical size of ships, contain fewer background pixels, and efficiently separate dense ships, as shown in Fig. 1(d).

Several methods for arbitrary-oriented scene text detection have been proposed. For example, Jiang et al. [21] proposed a text detection method called a rotational region CNN (R2CNN) by introducing angle information. Ma et al. [22] proposed the rotation RPN (RRPN) for scene text detection, which improved upon the RPN and region of interest (RoI) of Faster R-CNN by introducing a rotation factor. Liao et al. [23] proposed a text detector called the rotation-sensitive regression detector (RRD). Inspired by these methods above, some rotated detectors for ship detection have been proposed. Liu et al. [24] introduced a method called RR-CNN for ship detection and a public ship dataset called the 2016 High-Resolution Ship Collection (HRSC2016). Yang et al. [25] integrated a dense FPN, RRPN, and multiscale RoI Align to achieve ship detection. Zhang et al. [26] employed a rotated RPN to generate RBBs with orientation angle information. Tian et al. [27] considered the multiscale problem in ship detection. As result of their complex structures, there exists a speed bottleneck. One-stage rotation detectors can be a good choice if one is willing to sacrifice accuracy for speed. However, the loss discontinuity problem for angle regression still limits the performance of the one-stage rotation detectors.

To address these above-mentioned problems, we present a rotated one-stage detector (ROS-Det) for arbitrary-oriented ship detection based on RetinaNet, which integrates an FPN based on an improved 50-layer residual neural network (ResNet50), rotated anchors, a classification network, and a regression network. Compared with other methods, such as Faster R-CNN [10], Cascade R-CNN [14], SSD [16], RetinaNet [17], R-FCN [13], R2CNN [21], RR-CNN [24], RDPFPN [25], RRPN [22], R2PN [26], Tian et al [27], RRD [23], RoI-Transformer [28] and RSDet [29], our proposed ROS-Det method is able to achieve higher detection performance. The main ideas and contributions are as follows:

1) Different from previous ship detection methods, our proposed ROS-Det method can handle arbitrary orientations, various sizes and dense distributions in ship object detection.

2) To improve robustness against various sizes of ships, we propose an FPN based on an improved ResNet50 that contributes to addressing different sizes of ships. Moreover, the improved ResNet50 is able to receive more information and reduce the computational cost.

3) For the purpose of arbitrary-oriented ship detection, we employ a rotated anchor, the skew intersection-over-union (IoU) and skew non-maximum suppression (NMS) to improve the one-stage detector RetinaNet.

4) To address loss discontinuity problem, we improve the traditional smooth L1 loss function by introducing an IoU constant factor.

II. METHODS

The architecture of the proposed ROS-Det method is shown in Fig. 2, which includes four parts: an FPN based on
improved ResNet50, rotated anchors, a classification network (Cls-Net), and a regression network (Reg-Net). Specifically, the FPN based on an improved ResNet50 is used to extract and fuse multiscale feature maps. Then, we add rotated anchors of multiple aspect ratios, sizes and angles at each FPN level. Finally, Cls-Net and Reg-Net perform classification and regression, respectively.

A. FEATURE PYRAMID NETWORK BASED ON AN IMPROVED ResNet50

A model tweak [30] generally does not increase the time cost, but it may substantially improve performance. Therefore, an improved ResNet50 is used to extract feature maps over an entire input image. First, we briefly introduce ResNet50 [31], which mainly includes five block groups, as shown in Fig. 3.

For a $1 \times 1$ convolutional layer with a stride of 2, only 1/4 of the input information will be obtained during the convolution operation. To ensure that no information is ignored, we switch the stride sizes of the first two convolutional layers in path A. At the same time, in path B, we add a $2 \times 2$ average pooling (AvgPool) layer and change the stride size of the $1 \times 1$ convolutional layer to 1. To simplify the calculation, we replace the $7 \times 7$ convolutional layer in the first block group with three $3 \times 3$ convolutional layers. The above tweaks are illustrated in Fig. 4.

Scale imbalance (a multiscale object or input image in the dataset) plays a non-negligible role in affecting the overall detection accuracy [32]. Fig. 5 illustrates some ship samples from the HRSC2016 dataset, and we observe that there is a large difference in the size of ships in the same image. The common methods to solve the above problem are: (1) feature hierarchy-based methods (e.g., SSD [16], MSCNN [33]); (2) image pyramid-based methods (e.g., SNIP [32]); (3) feature pyramid-based methods (e.g., FPN [11]). Considering that the feature information varies among
different layers of the FPN, feature hierarchy-based methods making predictions separately from different layers by hierarchy-based methods is unreliable. Since an image pyramid-based method needs to make predictions separately from multi-scale images, we use one only as a data augmentation method.

To avoid these issues above, we use an FPN [11] based on the improved ResNet50 mentioned above. For the improved ResNet50, we represent the final output of the five block groups as $C_i$, $i = 1, 2, 3, 4, 5$, which are reduced by $\{2, 22, 23, 24, 25\}$ times compared to the original image. First, the top-down pathway upsamples the feature maps. Then, each lateral connection merges the feature maps in different pathways (i.e., the bottom-up pathway and top-down pathway). Finally, a $3 \times 3$ convolutional layer is added to generate the final feature map $P$ (including $P_3$ through $P_5$). In addition, $P_6$ is obtained via a $3 \times 3$ convolutional layer on $C_5$, and $P_7$ is obtained via rectified linear unit (ReLU) activation and a $3 \times 3$ convolutional layer on $P_6$.

B. ROTATED ANCHORS

To generate an RBB, we use $(x, y, w, h, \theta) \in [-\pi/2, 0)$ to redefine the bounding box, as shown in Fig. 6. Similar to the anchors used in Faster R-CNN [10], we use translation-invariant anchor boxes. For rotation detection and denser scale coverage, we employ multiple aspect ratios $\{1:1, 1:2, 2:1, 1:3, 3:1, 1:5, 5:1, 1:7, 7:1\}$, sizes $\{20, 21/3, 22/3\}$, and angles $\{-15^\circ, -30^\circ, -45^\circ, -60^\circ, -75^\circ, -90^\circ\}$ at each level. The settings of anchors are summarized in Fig. 7. Each spatial position will generate $A=162$ anchors.

Anchors are assigned to ground-truth object boxes if the IoU between the anchor and ground-truth object boxes is at least 0.5; anchors are assigned to the background if the IoU is in $[0, 0.4)$. Anchors that are unassigned do not participate in training. For each anchor, the length 5-vector of box regression represents the offsets between each anchor and its assigned ground-truth.

It is inaccurate if we use the traditional IoU computation method for HBBs to compute the IoU of the RBBs. Therefore, the skew IoU computation is redefined to consider the geometric principles of triangulation. Given two rectangles $R_{ABCD}$ and $R_{EFGH}$, the geometric principles and pseudo code of skew IoU computation are shown in Fig. 8 and Table 1, respectively. Specifically, we generate the point set $PSet$ between $R_{ABCD}$ and $R_{EFGH}$. The $PSet$ includes the intersection points $PSet1$ of two rectangles and the vertices $PSet2$ of one rectangle inside another. Then, we compute the intersection area $PSet$ by triangulation. We generate a convex polygon by sorting the points of $PSet$ into anticlockwise order, and then sum the areas of the triangles. Finally, we compute the skew IoU.

C. CLASSIFICATION AND REGRESSION NETWORKS

1) CLASSIFICATION NETWORK

The Cls-Net, a simple and small FCN, is attached to one of the pyramid levels and shares parameters with each other. Specifically, the Cls-Net consists of four sets of $3 \times 3 \times 256$ convolutional layers followed by ReLU activations, a $3 \times 3 \times A$ convolutional layer, and sigmoid activation, as shown in Fig. 9. Given the $W \times H \times 256$ feature map, binary predictions per spatial location are output.
As mentioned above, five offsets in total are predicted:

\[ b_x = (x - x_a) / w_a, \quad b_y = (y - y_a) / h_a \]  \hspace{1cm} (1)
\[ b_w = \log (w / w_a), \quad b_h = \log (h / h_a), \quad b_\theta = \theta - \theta_a \]  \hspace{1cm} (2)
\[ b_x' = (x' - x_a) / w_a, \quad b_y' = (y' - y_a) / h_a \]  \hspace{1cm} (3)
\[ b_w' = \log (w' / w_a), \quad b_h' = \log (h' / h_a), \quad b_\theta' = \theta' - \theta_a \]  \hspace{1cm} (4)

where \((x, y, w, h, \theta)\) indicate the box’s center coordinates, width, height, and angle, respectively. \(\theta_a, \theta', \theta\) indicate the anchor, predicted bounding box, and ground-truth, respectively (likewise for \(x, y, w, h\)).

For theCls-Net and Reg-Net, a multitask loss function is defined as follows:

\[
L = \frac{1}{N} \sum_{n=1}^{N} L_{cls}(p'_n, p_n) + \frac{1}{N} \sum_{n=1}^{N} p_n \times \sum_{i=[x, y, w, h, \theta]} \left| \frac{L_{reg}(b'_ni, b_n)}{L_{reg}(b'_ni, b_n)} \right| - \log (IoU) \]  \hspace{1cm} (5)

\[
L_{cls}(p'_n, p_n) = FL(p'_n, p_n) = FL(p_n)
\]  \hspace{1cm} (6)
\[
L_{reg}(b'_ni, b_n) = smoothL_{1}(b'_ni - b_n)
\]  \hspace{1cm} (7)

\[
smoothL_{1}(z) = \begin{cases} 
0.5z^2 & |z| < 1 \\
|z| - 0.5 & \text{otherwise}
\end{cases}
\]  \hspace{1cm} (8)
\[
p_n = \begin{cases} 
p_n' & \text{ship} \\
1 - p_n' & \text{background}
\end{cases}
\]  \hspace{1cm} (9)
\[
\alpha_n = \begin{cases} 
\alpha & p_n = 1 \\
1 - \alpha & \text{otherwise}
\end{cases}
\]  \hspace{1cm} (10)
\[
\gamma
\]  \hspace{1cm} (11)

where \(n\) is the index of an anchor and \(N\) indicates the number of anchors. \(p'_n\) is the category probability calculated by the sigmoid function. \(p_n\) represents the ground-truth class, i.e., ships. \(b'_n, b_n(i = [x, y, w, h, \theta])\) represent the offsets of the predicted bounding box and ground-truth, respectively. The classification loss is the focal loss (FL) [17], in which \(\alpha \in [0, 1]\) is a weighting factor, and \(\gamma\) is a focusing parameter. In this paper, we use \(\alpha = 0.25\) and \(\gamma = 2\) in all experiments. The regression loss is the smooth L1 loss [9].

However, there exists a loss discontinuity for the angle parameter in the rotation detection, as shown in Fig. 10. During the regression of an anchor \((0, 0, 60, 30, -\pi/2)\), an optimal form is to rotate the anchor counterclockwise to obtain the predicted box \((0, 0, 60, 30, -\pi/2)\), but the smooth L1 loss between the predicted offsets and target offsets is very large as a result of the angular periodicity and edge exchange. To avoid the loss discontinuity above, the model must be regressed on other relatively complex forms. Specifically, the anchor needs to be rotated clockwise to obtain the yellow box \((0, 0, 60, 30, -\pi/6)\), then the yellow box is scaled to obtain the final predicted box \((0, 0, 30, -\pi/6)\). Although the latter is ideal, it increases the regression difficulty. SCRDet [34] has confirmed the effectiveness of introducing the IoU constant factor in addressing boundary problem. Therefore, the IoU constant factor \(\frac{1 - \log(IoU)}{L_{reg}(b'_ni, b_n)}\) is introduced into traditional smooth L1 loss, as shown below:

\[
L = \frac{1}{N} \sum_{n=1}^{N} L_{cls}(p'_n, p_n) + \frac{1}{N} \sum_{n=1}^{N} p_n \times \sum_{i=[x, y, w, h, \theta]} \left| \frac{L_{reg}(b'_ni, b_n)}{L_{reg}(b'_ni, b_n)} \right| - \log (IoU) \]  \hspace{1cm} (12)
where $\text{IoU}$ indicates the overlap between the predicted bounding box and the ground-truth. The improved regression loss consists of two parts: $L_{\text{reg}}(b'_{ni}, b_{ni})$ and $-\log(\text{IoU})$. When the angular periodicity and edge exchange occur, the regression loss is approximately equal to 0 ($-\log(\text{IoU}) \approx 0$).

4) SKEW NON-MAXIMUM SUPPRESSION

NMS is extensively used as a postprocessing method for object detection to eliminate redundant bounding boxes. In rotation detection, the traditional NMS method is insufficient because of the added angle parameter. Given the IoU threshold $T$ (e.g., 0.7), an anchor (false positive) with a ratio of 1:8 and an angle difference of 15 degrees has an IoU of 0.31, which is less than the threshold $T$. However, it should actually be regarded as a positive sample. Therefore, we modify the skew NMS method by defining two constraints: (a) we keep the bounding boxes with an IoU larger than 0.7; (b) if the IoU is greater than 0.3 and less than 0.7, we retain the bounding boxes with angle differences of less than 15 degrees.

D. DATA AUGMENTATION

The performances of deep learning models are heavily dependent on massive numbers of well-annotated training samples. However, the number of labeled remote sensing images is very limited. Therefore, obtaining a large-scale well-annotated remote sensing dataset is much more laborious and expensive than obtaining a natural image dataset. Compared with object classification, object detection is more difficult to implement data augmentation strategies for object detection. The main reason is that more complexity will be introduced as a result of distorting the image, bounding-box locations and object sizes. In this paper, we adopt random rotation, graying, multiscale training and flipping.

III. EXPERIMENTS AND RESULTS

The experimental environment includes an Intel Core i9-9820X processor and four NVIDIA GeForce RTX 2080Ti GPUs. The deep learning framework is TensorFlow running on Ubuntu16.04.

A. EXPERIMENTAL DATASET

The HRSC2016 [25] dataset contains 1061 images and 2976 ship samples collected from Google Earth. It includes 436 training images, 181 validation images and 444 testing images. Fig. 11 shows examples from the HRSC2016 dataset.

The annotation files in the HRSC2016 are in XML format, where the data annotations include HBB format (boxxmin, boxymin, boxxmax, boxymax) and RBB format (mboxcx, mboxcy, mboxw, mboxh, mboxang), as shown in Fig. 12(a). The annotation files in the Visual Object Classes (VOC) dataset are in XML format, where the data annotations are in HBB format and are determined by the following values: (xmin, ymin, xmax, ymax). The annotation files used by our model are also in XML format, where the data annotations are in RBB format and are determined by the following values: (x0, y0, x1, y1, x2, y2, x3, y3). Therefore, we use the code to convert the annotation files in the HRSC2016 dataset into the above two annotated files, as shown in Fig. 12(b) and Fig. 12(c).

B. EVALUATION INDEX

The evaluation indexes are the precision-recall curve (PRC) and average precision (AP).

The precision and recall are defined as follows:

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

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

where TP, FP, and FN denote the numbers of true positive samples, false positive samples, and false negative samples, respectively. In this paper, the IoU threshold is set to 0.5. If the threshold is exceeded, then the object is detected.

The AP is defined as the area under the PRC, and the definition is as follows:

$$\text{AP} = \int_0^1 P(R)dR \times 100\%$$

C. TRANSFER LEARNING AND NETWORK TRAINING

Generally, the pretrained model can be applied to new object detection tasks through transfer learning, which is efficient and can prevent overfitting. The weights of the model are initialized by the pretrained improved ResNet50 model by default. The initial learning rate is 0.0005, and the maximum number of iterations is 100,000. The learning rate is reduced...
to 0.00005 after 60,000 iterations, and to 0.000005 after 80,000 iterations. In addition, we choose the momentum optimizer during optimization. The learning rate and loss value for the sequence of iterations are shown in Fig. 13.

**FIGURE 13.** The learning rate and loss value with different iterations. (a) Learning rate. (b) Loss value.

### D. PARAMETER ANALYSIS

The settings of the anchors have a key influence on the detection results. To obtain the best detection performance during network training, we investigate the detection results of the proposed ROS-Det using different anchor settings, as shown in Table 2.

As shown in Table 2, the AP gradually increases as the number of anchors increases and plateaus at 88.2%. This demonstrates that using anchors with multiple aspect ratios and angles is an effective means to improve the AP. When using 5 and 9 aspect ratios, the APs of rows 4 and 7 are lower than those of rows 3 and 6, respectively, demonstrating that the AP will drop when the number of angles is increased from 6 to 7. In addition, the AP value for 11 aspect ratios and 6 angles is lower than that for 9 aspect ratios and 6 angles, suggesting that the AP will drop when the number of ratios is increased from 9 to 11. Therefore, in the proposed ROS-Det, we use 9 aspect ratios and 6 angles, i.e., {1:1, 1:2, 2:1, 1:3, 3:1, 1:5, 1:7, 7:1} and {−90, −75, −60, −45, −30, −15}, respectively. In Table 2, we also summarize the training and inference times under different aspect ratios and angles. As the number of anchors increases, the training time and inference time gradually increase. Although the time cost of 9 ratios and 6 angles is not the smallest, it is still efficient.

**FIGURE 14.** How the PRCs of the proposed ROS-Det change using different anchor settings. The higher PRC denotes a larger AP value. It is shown that the PRC for 3 aspect ratios and 4 angles lies at the innermost, demonstrating that its AP value is the worst. The PRC for 9 aspect ratios and 6 angles is the highest, which suggests that this setting is the best.

### E. ABLATION STUDY

To quantitatively illustrate the effectiveness of each part, we perform a comprehensive ablation study.

#### 1) EFFECTIVENESS OF THE DATA AUGMENTATION STRATEGY

Table 3 shows the detailed detection results of different data augmentation schemes. As shown in Table 3, our proposed...
TABLE 2. Detection results of the proposed ROS-Det using different settings of anchors.

| Setting           | Aspect Ratio         | Angle           | AP     | Training time | Inference time |
|-------------------|----------------------|-----------------|--------|---------------|---------------|
| 3 ratios, 4 angles| {1:1, 1:3, 3:1}      | (-90, -60, -45, -15) | 69.9   | 07h 07m 48s  | 0.182s        |
| 5 ratios, 4 angles| {1:1, 1:3, 3:1, 1:5, 5:1} | (-90, -60, -45, -15) | 74.1   | 08h 40m 08s  | 0.314s        |
| 5 ratios, 6 angles| {1:1, 1:3, 3:1, 1:5, 5:1} | (-90, -75, -60, -45, -30, -15) | 78.9   | 11h 01m 19s  | 0.331s        |
| 5 ratios, 7 angles| {1:1, 1:3, 3:1, 1:5, 5:1} | (-90, -75, -60, -45, -30, -15, 0) | 87.8   | 11h 50m 08s  | 0.340s        |
| 7 ratios, 6 angles| {1:1, 1:2, 2:1, 1:3, 3:1, 1:5, 5:1} | (-90, -75, -60, -45, -30, -15) | 87.5   | 13h 46m 18s  | 0.349s        |
| 9 ratios, 6 angles| {1:1, 1:2, 2:1, 1:3, 3:1, 1:5, 5:1, 1:7, 7:1} | (-90, -75, -60, -45, -30, -15) | **88.2** | 15h 23m 06s  | 0.374s        |
| 9 ratios, 7 angles| {1:1, 1:2, 2:1, 1:3, 3:1, 1:5, 5:1, 1:7, 7:1} | (-90, -75, -60, -45, -30, -15, 0) | 88.0   | 19h 03m 20s  | 0.391s        |
| 11 ratios, 6 angles| {1:1, 1:2, 2:1, 1:3, 3:1, 1:5, 5:1, 1:7, 7:1, 1:9, 9:1} | (-90, -75, -60, -45, -30, -15) | 88.1   | 19h 02m 54s  | 0.395s        |

The short names for time are defined as: h-hour, m-minute, and s-second.

ROS-Det achieves the highest AP, i.e., 88.2%. For schemes 1, 2, 3 and 4, the AP value gradually increases, which proves that these data augmentation strategies are helpful for improving the ship detection results. The PRCs of different data augmentation schemes are also shown in Fig. 15. The PRC of ROS-Det has the highest AP value.

2) EFFECTIVENESS OF THE IMPROVED ResNet50

Table 4 shows the results of three pre-trained networks. As shown in Table 4, using the improved ResNet50 achieves the best performance in terms of the AP 88.2%. From TABLE 3. Detection results of different data augmentation schemes.

| Scheme | Data Augmentation Strategy | AP  |
|--------|----------------------------|-----|
| 1      | R, V, G, H, M              | 84.6|
| 2      | R, V, G, H                 | 85.0|
| 3      | R, V, G                    | 85.4|
| 4      | R, V, G, M                 | 86.7|
| ROS-Det|                           | **88.2**|

The short names for data augmentation strategies are defined as: R-Rotation, VF-Vertical flip, G-Graying, HF-Horizontal flip, and MS-Multiscale.

TABLE 4. Detection results of different pre-trained networks.

| Pre-trained Network    | AP  |
|------------------------|-----|
| MobileNet V2           | 67.1|
| ResNet50               | 88.1|
| Improved ResNet50 (ROS-Det) | **88.2** |

MobileNet V2, ResNet50 and the improved ResNet50, the AP gradually increases. Compared with the original ResNet-50, our improved ResNet-50 achieves an improvement of approximately 0.1%. The reason is that the improved ResNet50 is able to receive more information, and replaces the 7 × 7 convolutional layer with three 3 × 3 ones. This comparative experiment illustrates the role of the improved ResNet50.
Fig. 16 illustrates how the PRCs of our method change in different pre-trained networks. The PRC of MobileNet V2 is the lowest, indicating that its AP is the worst. The PRC of ResNet50 lies higher than that of MobileNet V2, proving that the feature representation ability of ResNet50 is better than that of MobileNet V2. The improved ResNet50 (ROS-Det) has the highest PRC, indicating it achieves the highest AP.

3) EFFECTIVENESS OF IMPROVED LOSS FUNCTION

Table 5 shows the detailed detection results with different loss functions. As shown in Table 5, the improved loss function achieves better performance than the traditional smooth L1 loss function. This is because the improved loss function can solve the loss discontinuity problem of arbitrary-oriented object detection by introducing the IoU constant factor into the traditional smooth L1 loss. This comparative experiment illustrates the role of the improved loss function. The PRCs of different loss functions are illustrated in Fig. 17. The PRC of improved smooth L1 loss lies higher than that of traditional smooth L1 loss, showing the effectiveness of the improved loss function. In addition, the visualization results of different loss functions are shown in Fig. 18. It can be seen that the improved smooth L1 loss can significantly improve the accuracy of bounding box regression.

F. COMPARISON OF RESULTS

To verify the superiority of the proposed ROS-Det method, we will compare it with other state-of-the-art methods. The AP of each method is recorded and compared, as shown in Table 6. Since HRSC2016 dataset provides ground truths of both HBBs and RBBs, comparison methods include two categories, i.e., (1) horizontal object detection including Faster R-CNN [10], Cascade R-CNN [14], SSD [16], RetinaNet [17] and R-FCN [13]; (2) arbitrary-oriented object detection including R2CNN [21], RR-CNN [24], RDFPN [25], RRPN [22], R2PN [26], Tian et al [27], RRD [23] Rol-Transformer [28] and RSDet [29].

As can be seen in Table 6, our proposed ROS-Det method achieves the AP of 88.2 % in HRSC2016 dataset, it significantly outperforms other state-of-the-art methods, e.g., improved by R-FCN/RSDet over 1.7 % on AP. These observations verify the superior performance of ROS-Det.

1) Comparison to the horizontal object detection methods. The performance gaps are significant when comparing ROS-Det and the horizontal object detection methods including Faster R-CNN [10], Cascade R-CNN [14], SSD [16], RetinaNet [17] and R-FCN [13]. Specifically, the AP of ROS-Det is higher than Faster R-CNN by 7.5 %, Cascade

---

**TABLE 5. Detection results of different loss functions.**

| Pre-trained Network                  | AP     |
|-------------------------------------|--------|
| Smooth L1 Loss                      | 87.1   |
| Improved ResNet50 (ROS-Det)         | **88.2** |

![FIGURE 17. Precision-recall curves of different loss functions.](image1)

![FIGURE 18. Visualization Results of different loss functions. (a) Smooth L1 loss. (b) Improved Smooth L1 loss.](image2)
FIGURE 19. Some comparison Results of different methods.
FIGURE 20. The visualized detection results of the proposed ROS-Det on HRSC2016 dataset.

TABLE 6. Performance comparison of different methods on HRSC2016.

| No. | Method         | Backbone  | AP   |
|-----|----------------|-----------|------|
| 1   | Faster R-CNN [10] | VGG16     | 80.7 |
| 2   | Cascade R-CNN [14] | ResNet50  | 80.9 |
| 3   | SSD [16]         | VGG16     | 83.9 |
| 4   | RetinaNet [17]   | ResNet50  | 82.6 |
| 5   | R-FCN [13]       | ResNet50  | 86.5 |
| 6   | R2CNN [21]       | ResNet101 | 73.1 |
| 7   | RR-CNN [24]      | VGG16     | 75.7 |
| 8   | RDFPN [25]       | ResNet101 | 76.2 |
| 9   | RRPN [22]        | ResNet101 | 79.1 |
| 10  | R2PN [26]        | VGG16     | 79.6 |
| 11  | Tian et al [27]  | VGG16     | 80.8 |
| 12  | RRD [23]         | VGG16     | 84.3 |
| 13  | RoI-Transformer [28] | ResNet101 | 86.2 |
| 14  | RSDet [29]       | ResNet152 | 86.5 |
| 15  | ROS-Det (ours)   | Improved ResNet50 | **88.2** |

R-CNN by 7.3 %, RetinaNet by 5.6 %, SSD by 4.3 %, and R-FCN by 1.7 %, respectively. The main reason is that ROS-Det achieves the purpose of arbitrary-oriented ship detection, which can better handle arbitrary orientations, various sizes and dense distribution in ship object detection.

2) Comparison to the arbitrary-oriented object detection methods. Our ROS-Det method significantly outperforms the arbitrary-oriented object detection methods including R2CNN [21], RR-CNN [24], RDFPN [25], RRPN [22], R2PN [26], Tian et al [27], RRD [23] RoI-Transformer [28] and RSDet [29]. Specifically, the AP of ROS-Det is higher than R2CNN by 15.1 %, RR-CNN by 12.5 %, RDFPN by 12 %, RRPN by 9.1 %, R2PN by 8.6 %, Tian et al by 7.4 %, RRD by 3.9 %, RoI-Transformer by 2 % and RSDet by 1.7 %, respectively. A key reason is that ROS-Det improves the traditional smooth L1 loss function by introducing an intersection-over-union constant factor, which can better handle the loss discontinuity problem in arbitrary-oriented object detection.

To demonstrate the advantages of our method, we conduct a qualitative visualization comparison of different methods on two scenes, i.e., dense distribution and different sizes of ships. As shown in the first and second columns of Fig. 19, there are some densely distributed ships. It can be observed that R2CNN and our method can detect all ships. However, compared with R2CNN, our method has no false detections and its positioning is more accurate. This indicates that our...
method has better robustness against the impact of dense distribution. As shown in the third and fourth columns of Fig. 19, there are some ships with different sizes. It is obvious that our method can detect all ships more accurately. In other words, our method has better robustness against the impact of different sizes.

G. VISUALIZED RESULTS
To display detection effect of the proposed ROS-Det, Fig. 20 shows the visualized detection results on HRSC2016 dataset. It can be seen that our method can detect ships well in various situations, e.g., arbitrary orientations, various sizes and dense distribution.

IV. CONCLUSION
We have proposed ROS-Det, an arbitrary-oriented ship detection method, which consists of a FPN based on an improved ResNet50, rotated anchors, a classification network, and a regression network. In consideration of various sizes of ships, ROS-Det fuses multi-scale features by using the FPN based on an improved ResNet50. Through several tweaks, the improved ResNet50 is able to receive more information and reduce the computational cost. The rotated anchors, skew IOU and skew NMS are employed to achieve arbitrary-orientation ship detection. Moreover, to offset the loss discontinuity impact of arbitrary-oriented object detection, an IOU constant factor is introduced to the traditional smooth L1 loss function.

Although ROS-Det has reached the optimal detection performance, there are still some missed detections and false positives. In addition, there are some small ships and incomplete ships ignored in HRSC2016 dataset. In future work, we will explore more efficient strategies for extremely small objects. In addition, we will further expand and improve the HRSC2016 dataset.

REFERENCES
[1] M. Zhu, Y. Xu, S. Ma, S. Li, H. Ma, and Y. Han, “Effective airplane detection in remote sensing images based on multilayer feature fusion and improved nonmaximal suppression algorithm,” Remote Sens., vol. 11, no. 9, p. 1062, 2019, doi: 10.3390/rs11091062.
[2] Y. Xu, M. Zhu, P. Xin, S. Li, M. Qi, and S. Ma, “Rapid airplane detection in remote sensing images based on multilayer feature fusion in fully convolutional neural networks,” Sensors, vol. 18, no. 7, p. 2335, Jul. 2018, doi: 10.3390/s18072335.
[3] Y. Xu, M. Zhu, S. Li, H. Feng, S. Ma, and J. Che, “End-to-end airport detection in remote sensing images combining cascade region proposal networks and multi-threshold detection networks,” Remote Sens., vol. 10, no. 10, p. 1516, Sep. 2018, doi: 10.3390/rs10101516.
[4] H. He, Y. Lin, F. Chen, H.-M. Tai, and Z. Yin, “Unshore ship detection in remote sensing images via weighted pose voting,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 6, pp. 3091–3107, Jun. 2017, doi: 10.1109/tgrs.2017.2658950.
[5] F. Yang, Q. Xu, B. Li, and Y. Ji, “Ship detection from thermal remote sensing imagery through region-based deep forest,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 3, pp. 449–453, Mar. 2018, doi: 10.1109/lgrs.2017.2879360.
[6] Y. Yu, H. Ai, X. He, S. Yu, X. Zhong, and M. Lu, “Ship detection in optical satellite images using Haar-like features and periphery-cropped neural networks,” IEEE Access, vol. 6, pp. 71122–71131, 2018, doi: 10.1109/access.2018.2881479.
[7] C. Dong, J. Liu, F. Xu, and C. Liu, “Ship detection from optical remote sensing images using multi-scale analysis and Fourier HOG descriptor,” Remote Sens., vol. 11, no. 13, p. 1529, 2019, doi: 10.3390/rs11131529.
[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.
[9] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.
[10] 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.
[11] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 936–944.
[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.
[13] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object detection via region-based fully convolutional networks,” in Proc. Comput. Vis. Pattern Recognit., 2016, pp. 1–11.
[14] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6154–6162, doi: 10.1109/CVPR.2018.00644.
[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.
[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot MultiBox detector,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.
[17] 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. (iCCV), Oct. 2017, pp. 2999–3007.
[18] Y. You, J. Cao, Y. Zhang, F. Liu, and W. Zhou, “Nearshore ship detection on high-resolution remote sensing image via scene-mask R-CNN,” IEEE Access, vol. 7, pp. 128431–128444, 2019, doi: 10.1109/access.2019.2940102.
[19] Z. Shao, L. Wang, Z. Wang, W. Du, and W. Wu, “Salience-aware convolution neural network for ship detection in surveillance video,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 3, pp. 781–794, Mar. 2020, doi: 10.1109/tcsvt.2019.2897980.
[20] Y. Wu, W. Ma, M. Gong, Z. Bai, W. Zhao, Q. Guo, X. Chen, and Q. Miao, “A coarse-to-fine network for ship detection in optical remote sensing images,” Remote Sens., vol. 12, no. 2, p. 246, Jan. 2020, doi: 10.3390/rs12020246.
[21] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, “R2CNN: Rotational region CNN for orientation robust scene text detection,” 2017, arXiv:1706.09579. [Online]. Available: http://arxiv.org/abs/1706.09579.
[22] L. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, “Arbitrary-oriented scene text detection via rotation proposals,” IEEE Trans. Multimedia, vol. 20, no. 11, pp. 3111–3122, Nov. 2018, doi: 10.1109/tmm.2018.2856921.
[23] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, and X. Bai, “Rotation-sensitive regression for oriented scene text detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5909–5918, doi: 10.1109/CVPR.2018.00619.
[24] Z. Liu, J. Hu, L. Weng, and Y. Yang, “Rotated region based CNN for ship detection,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2017, pp. 900–904.
[25] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, “Automatic ship detection in remote sensing images from Google earth of complex scenes based on multiscale rotation dense feature pyramid networks,” Remote Sens., vol. 10, no. 1, p. 132, Jan. 2018, doi: 10.3390/rs10010132.
[26] Z. Zhang, W. Guo, S. Zhu, and W. Yu, “Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 11, pp. 1745–1749, Nov. 2018, doi: 10.1109/lgrs.2018.2856921.
[27] T. Tian, Z. Pan, X. Tan, and Z. Chu, “Arbitrary-oriented inshore ship detection based on multi-scale feature fusion and contextual pooling on rotation region proposals,” Remote Sens., vol. 12, no. 2, p. 339, Jan. 2020, doi: 10.3390/rs12020339.
[28] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, “Learning RoI transformer for detecting oriented objects in aerial images,” 2018, arXiv:1812.00155. [Online]. Available: http://arxiv.org/abs/1812.00155.
[29] W. Qian, X. Yang, S. Peng, Y. Guo, and J. Yan, “Learning modulated loss for rotated object detection,” 2019, arXiv:1911.08299. [Online]. Available: http://arxiv.org/abs/1911.08299

[30] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, “Bag of tricks for image classification with convolutional neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 558–567, doi: 10.1109/CVPR.2019.00065.

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

[32] B. Singh and L. S. Davis, “An analysis of scale invariance in object detection–SNIP,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3578–3587, doi: 10.1109/CVPR.2018.00377.

[33] Z. Cai, Q. Fan, R. S. Fers, and N. Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 354–370.

[34] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, S. Xian, and K. Fu, “SCRDet: Towards more robust detection for small, cluttered and rotated objects,” 2018, arXiv:1811.07126. [Online]. Available: http://arxiv.org/abs/1811.07126

MINGMING ZHU received the M.S. degree in information and communication engineering from Air Force Engineering University, Xi’an, China, in 2018, where he is currently pursuing the Ph.D. degree in information and communication engineering.

His research interests include radar signal processing, deep learning, image processing, and object detection.

GUOPING HU graduated from the Air and Missile Defense College, Xi’an, China, in 1985. He received the M.S. degree from the Air and Missile Defense College, in 1990, and the Ph.D. degree in electrical engineering from Xidian University, Xi’an, in 2010.

He is currently a Professor and the Director of the Institute of Radar Anti-Stealth Technology, Air Force Engineering University, Xi’an. His research interests include radar signal processing, radar anti-stealth technology, wireless communication technology, and image processing.

SHUAI LI received the M.S. and Ph.D. degrees in information and communication engineering from Aeronautics and Astronautics Engineering College, Air Force Engineering University, China, in 2014 and 2019, respectively.

His research interests include computer vision, intelligent computation, and signal processing.

SHIQIANG WANG received the Ph.D. degree in information and communication engineering from Air Force Engineering University, China, in 2012.

He is currently an Assistant Professor with the Air and Missile Defense College, Air Force Engineering University. His current work includes intelligent information processing, radar signal processing, deep learning, and radar engineering.

YULE ZHANG received the B.S. degree from Air Force Engineering University, Xi’an, China, in 2019, where he is currently pursuing the M.S. degree in information and communication engineering.

His research interests include radar signal processing and image processing.

SHIJIE YUE received the B.S. degree from Air Force Engineering University, Xi’an, China, in 2019, where he is currently pursuing the M.S. degree in information and communication engineering.

His research interests include radar signal processing and image processing.

**Hao Zhou** received the M.S. and Ph.D. degrees in information and communication engineering from Air Force Engineering University, Xi’an, China, in 2015 and 2019, respectively.

His research interests include low altitude target detection, direction finding, and deep learning.

**ShiQiang Wang** received the Ph.D. degree in information and communication engineering from Air Force Engineering University, China, in 2012.

He is currently an Assistant Professor with the Air and Missile Defense College, Air Force Engineering University. His current work includes intelligent information processing, radar signal processing, deep learning, and radar engineering.

**Yule Zhang** received the B.S. degree from Air Force Engineering University, Xi’an, China, in 2019, where he is currently pursuing the M.S. degree in information and communication engineering.

His research interests include radar signal processing and image processing.