IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received May 31, 2019, accepted June 18, 2019, date of publication June 24, 2019, date of current version July 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2924643

Single Shot Anchor Refinement Network for Oriented Object Detection in Optical Remote Sensing Imagery

SONGZE BAO[1,2], XING ZHONG[1,2,3], RUIFEI ZHU[1,2,3], XIAONAN ZHANG[1,2], ZHUQIANG LI[3], AND MENGYANG LI[1,2]

[1]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Key Laboratory of Satellite Remote Sensing Application Technology of Jilin Province, Chang Guang Satellite Technology Company Ltd., Changchun 130000, China

Corresponding author: Xing Zhong (ciomper@163.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0502600, in part by the Youth Innovation Promotion Association of Chinese Academy of Science, in part by the National Science Foundation for Young Scholars of China under Grant 61505203, in part by the Key Scientific and Technological Research and Development Projects of Jilin under Grant 20180201109GX, and in part by the Scientific and Technological Plan of Changchun under Grant 18SS012.

ABSTRACT Object detection is a challenging task in the field of remote sensing applications due to the complex backgrounds and uncertain orientation of targets. Compared with the horizontal bounding box, the oriented bounding box can provide orientation information while retaining the true size. Most existing oriented object detection methods are based on Faster-RCNN and the other one-stage methods that can achieve real-time speed but have shortcomings in localization and detection accuracy. To further enhance the performance of one-stage methods, we propose an oriented object detection framework that is based on the single shot detector, namely, single shot anchor refinement network (S^{2}ARN). The S^{2}ARN obtains the accurate detection results by performing two consecutive regressions. More precisely, the multilevel features of the backbone are used to regress the coordinate offsets between the predefined rotated anchors and the ground-truth boxes to generate the refined anchors. The classification and regression subnetworks assigned to the output features are used to perform the second regression to determine the class labels and further adjust the location of the refined anchors. In addition, receptive field amplification modules (RFAMs) are inserted to enlarge the receptive field and extract more discriminative features. Furthermore, in the anchor matching step, angle-related Intersection over Union (ArIoU) is used to calculate the Intersection over Union (IoU) score instead of the traditional method. Benefiting from the multiple regressions and the insensitivity of the ArIoU score to the angle deviation, the angle sampling interval of the rotated anchor can be reduced. The experimental results for the two public datasets, HRSC2016 and UCAS-AOD, demonstrate the effectiveness of the proposed network.

INDEX TERMS Convolutional neural network (CNN), remote sensing, oriented object detection, anchor refinement.

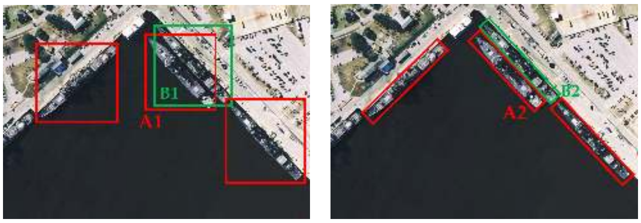
I. INTRODUCTION

In recent years, we have witnessed the remarkable progress of convolutional neural networks (CNNs) in many computer vision tasks such as image classification [1], [2], object detection [3]–[5], image segmentation [6] and medical image processing [7], [8]. Existing generic object detection methods are primarily divided into two-stage region-proposal-based methods [3], [9]–[12] and one-stage regression-based methods [4], [5], [13], [14]. Region-proposal-based methods, such as Faster-RCNN [3] and Mask-RCNN [12], generate a series of proposals by learning a Region Proposal Network (RPN). A region wise classifier is used to determine the object class label and fine-tune the location of detection bounding box. Regression-based methods extract high-level semantic features that are directly applied to bounding box regression

The associate editor coordinating the review of this manuscript and approving it for publication was Xian Sun.

and class label determination. The Single Shot multibox Detector (SSD) [4] and YOLOv2 [14] algorithms utilize the anchor mechanism in RPN which predefines a series of prior boxes (anchors) with different scales and aspect ratios at each spatial location of the feature maps. By calculating the Intersection over Union (IoU) scores between these anchors and the ground-truth boxes, positive and negative samples are separated to train the model. Due to the full convolutional structure, the region of interest (ROI) features do not have to be separately discriminated. Thus, the regression-based methods have a high detection efficiency. However, the detection accuracy is usually lower than that of the two-stage approaches.

In the field of remote sensing, many researchers have applied the generic object detection methods to remote sensing image object detection [15]–[18]. These methods use the same horizontal bounding boxes to detect targets. However, unlike natural images, remote sensing images are always taken in top views, which implies that the objects in remote sensing images are arbitrarily oriented. This orientation causes a misalignment between the objects and detection bounding boxes. In Fig. 1, for slender objects, such as ships, the bounding box of an incline ship contains redundant backgrounds. The ship occupies only a small part of the bounding box area, and a large overlap area between the bounding boxes of adjacent ships exists. Thus, the ship with lower confidence score will be suppressed during the non-maximum suppression (NMS) procedure, which causes a missed detection. In addition, a horizontal bounding box loses the shape information of the target, whereas an oriented bounding box can wrap the target in a tighter way, and the real size of the target can be preserved.



(a) (b)

FIGURE 1. Comparison of horizontal bounding boxes and rotated bounding boxes. (a) Slender targets are detected using horizontal bounding boxes. For two adjacent ships, the horizontal bounding boxes, A1 and B1, have a large overlap area, and the box with a lower confidence score will be suppressed during the NMS procedure. (b) The oriented bounding boxes A2 and B2 wrap targets more tightly.

To overcome the drawbacks of horizontal bounding boxes, many researchers have proposed methods that use oriented bounding boxes for object detection in remote sensing images [19]–[28], most methods are based on Faster R-CNN. Jiang *et al*. proposed R^{2}CNN [19] for text detection, which combines multisize pooled features and has been reimplemented for object detection in remote sensing images by a third-party research group; Liu *et al*. [20] and Zhang *et al*. [28] introduced multiangle anchors in RPN, and extracted Rotated ROI (RROI) features by Rotate ROI pooling. Koo *et al*. [23] extracted Diagonal ROI (DROI) and connected it to the RROI feature, which introduced contextual information and improved the robustness of the algorithm. Azimi *et al.* [21] and Yang *et al.* [25] used more complex backbones to improve the accuracy; however, it reduced detection efficiency. Ding *et al.* [22] employed a subnetwork with a fully connected layer to regress the transformation parameters from horizontal ROIs to rotated ROIs, which reduces the number of anchors and further improves the efficiency by using a light-head R-CNN. These methods, which are based on Faster-RCNN, inherit its defects in computation speed and storage space.

One-stage detection methods are also applied to oriented object detection [29]–[32]. DRBox [29], which is a variant of SSD, sets multiangle anchors to better match ground-truth boxes. In the training phase, Angle-related IoU (ArIoU) was utilized to calculate the IoU between rotated boxes to accurately guide the network in regressing angle deviation. DRBox achieved a detection speed of nearly 60 fps on an input size of 300 × 300 pixels. However, since only the feature map of a single layer was employed, the detection accuracy is limited by the feature representation and the small receptive field. In addition, the performance of the SSD-based methods is susceptible to the threshold settings, thus, acquiring high recall and precision simultaneously is difficult. Liu *et al.* [30] implemented an arbitrary-oriented ship detection method that is based on YOLOv2. Multiple feature maps with different resolutions were reorganized and concatenated, which introduced fine-grained features and improved the recall of small targets. However, the detection performance was limited by the lower angle regression accuracy. Although these one-stage methods can detect targets at a high speed, improvements in the detection and localization accuracy are still needed.

To solve these problems, we propose a high-accuracy one-stage oriented object detection framework named Single Shot Anchor Refinement Network (S^{2}ARN) while maintaining a real-time speed. The entire network is based on a Feature Pyramid Network (FPN) [33] structure with a ResNet [2] backbone. Through three efforts, we improve the localization and detection accuracy. First, Anchor Refinement Branches (ARBs) are introduced to provide high quality refined anchors for Object Detection Branches (ODBs) which further adjust the coordinates of the refined anchors for more accurate bounding boxes. The increased thresholds of two consecutive regressions alleviate the sensitivity of the threshold setting of SSD-based method, and better balance precision and recall. Second, considering that objects in remote sensing images have a variety of scales, a multibranch convolutional structure, namely, Receptive Field Amplification Modules (RFAMs) are designed to expand the effective receptive field of the detection layer and extract more discriminative features. Last, a rotated anchor matching strategy is carefully designed, thus, the targets with various aspect

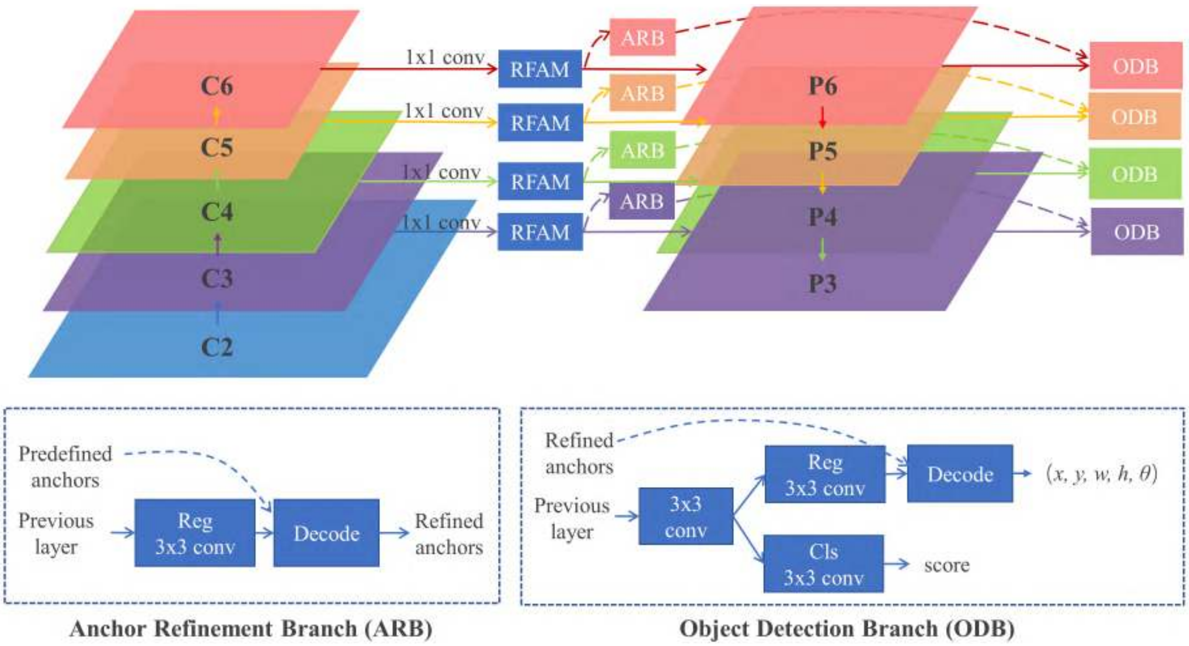


FIGURE 2. Overall structure of the single shot anchor refinement network (S^{2}ARN).

ratios can match a sufficient number of anchors to ensure the recall.

The experimental results based on two public datasets, the HRSC2016 and UCAS-AOD datasets, show the effectiveness of the proposed method. The remainder of this paper is organized as follows. Section II details the proposed method. Section III presents the datasets and evaluation indicators. Section IV presents comparative experiments to verify the validity of the proposed method. Section V concludes this paper.

II. PROPOSED METHOD

In this section, we detail the proposed S^{2}ARN. Fig. 2 depicts the total structure of the network. S^{2}ARN is designed based on an FPN architecture. A 3x3 dilated conv layer with a dilation rate of 2 is appended to C5 to produce C6 which has the same resolution as C5 but larger receptive field size. C6 has rich deep semantic information and is adopted for large object detection. To begin with, the number of channels for the multilevel feature maps {C6, C5, C4, C3} is compressed to 256 by 1×1 conv layers. The output feature maps are input into four RFAMs to further expand the effective receptive field to extract more discriminative features. In this study, we refer to {C6, C5, C4, C3} as "refiners", which are utilized to regress the offsets between the ground-truth boxes and the original predefined anchors. This process is performed by an additional 3×3 conv-layer named the ARB. Then, we decode the offset with the original anchor to obtain the refined anchors. {C6, C5, C4, C3} have strides of {32, 32, 16, 8}, respectively, and a dense to 8 pixels spatial sampling interval ensures that small targets can match enough anchors. The final feature pyramid {P6, P5, P4, P3}, namely, "predictors", are obtained by a top-down pathway and lateral connections. Similar to the SSD, a detection head referred to as the ODB is assigned to each predictor for classification and bounding box regression. The regression subnet of ODB further adjusts the locations of the refined anchors generated by ARB to better fit the ground-truth boxes. At last, confidence threshold screening and NMS are used to eliminate background and redundant detection boxes to obtain the detection results.

A. ARB

The SSD divides the positive and negative anchors based on the IoU scores between ground-truth boxes and anchors, which causes the selection of the IoU threshold to have a significant impact on the performance of the detector. A loose IoU threshold encourages more anchors to be classified into the foreground, which introduces more close false positives and lower precision, whereas, a tight IoU threshold substantially reduces the number of positive anchors, and the training process is overwhelmed by the negative anchors. Although the Focal Loss [34] can alleviate the problem of foreground-background class imbalance, an insufficient number of positive samples can easily cause overfitting. Therefore, obtaining accurate detection results by a single regression is difficult.

Kong *et al.* [35] observed that the misalignment between the optimization target and the inference configuration is an important factor that hinders the performance improvement of the SSD-based methods. In Fig. 3, as in [9] and [35], we plot the IoU values of the ground-truth boxes with their nearby anchors before and after regression to study the regression performance of the SSD-based algorithm. The SSD with ResNet50 backbone is adopted. We use {C5, C4, C3} as

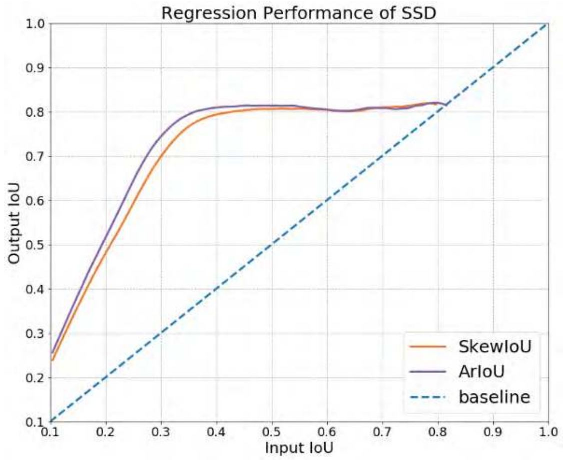


FIGURE 3. Regression performance of the SSD with ResNet50 backbone.

predictors and train on the HRSC2016 dataset for oriented object detection. The IoU scores between ground-truth boxes and their nearby anchors are calculated as input IoU scores. The Output IoU scores are calculated from the predicted boxes and the ground-truth boxes. We apply two kinds of IoU metrics to measure the overlap between two rotated bounding boxes, namely, the SkewIoU [30] metric and the ArIoU [29] metric, which will be described in Section 2.3. We can observe that the IoU value between the ground-truth box and the refined anchor has considerably improved after the regression regardless of which IoU metric is applied. Some anchors that are assigned as negatives may also match ground-truth boxes after the regression. In the training phase, the classification subnetwork classifies the predefined anchor into one of M object categories, if the IoU score related to any ground-truth box is greater than the threshold. During the inference phase, the predicted probability is assigned to the corresponding refined anchor which has a distinctly higher IoU score than the predefined anchor. As a result, the localization performance of the refined anchor does not match the classification score.

To solve these problems, S^{2}ARN uses two consecutive regressions to improve the detection accuracy. For the first regression, the positive and negative anchors are divided by a lower IoU threshold (0.4) to ensure the recall rate, and the offsets between the positive predefined anchors and ground-truth boxes are regressed by ARB. This process focuses only on the coordinate regression, and does not involve the object category determination, therefore classification loss is not calculated. In the second regression, a strict threshold (0.75) is employed as the criteria for selecting the positives, and the offsets between the refined anchors and the ground-truth boxes are further regressed. This process is beneficial for improving the precision rate and localization accuracy. Similar to the SSD, multitask learning is utilized to determine the bounding box coordinates and class label. A higher threshold encourages the refined anchors with a high localization accuracy to be predicted as foreground categories, which renders the localization ability and confidence score of the box more consistent and alleviates the misalignment between the optimization target and the inference configuration. As shown in Fig. 3, some predefined anchors have low IoU scores with ground-truth boxes. After the first regression, the scores have substantially improved. Therefore, when using a tight threshold such as 0.75, a large number of positive refined anchors still exist, which will not cause overfitting problems. Unlike RetinaNet [34], the detection head of each predictor in S^{2}ARN does not share parameters. Because each predictor has different scale features, separate use of the parameters facilitates the full use of these features.

B. RFAM

The Receptive Field (RF) in CNNs is the region of the input space that affects a particular output unit of the network. As pointed out in [36], the pixels in RF do not equally contributes to the final output; only a fraction of the area has an effective influence on the output unit. These pixels constitute the Effective Receptive Field (ERF), which linearly increases with respect to $1/\sqrt{N}$, where N is the number of convolution layers. Using a dilated conv layer or a conv layer with a large stride can efficiently increase the ERF size instead of expanding the network.

In object detection task, the anchor size should match the ERF size of the unit on the predictor. For remote sensing images, targets are often confused in complex backgrounds. Increasing the ERF size can provide more contextual information for the classification subnetwork, which can render a more robust and accurate classification [23]. Liu *et al.* [37] proposed the Receptive Field Block (RFB) based on the structure of the RFs in human visual systems. RFBs were not only assigned to the light weight backbone as extra layers to expand the entire network, but also cascaded after the shallow predictors to increase the ERF size of the shallow feature maps. The RFB is a multibranch convolutional block, and the last layer of each branch is a dilated conv layer with different dilation rates; thus, the previous layer has a variable sampling center. The RFB expands the ERF size of the output unit, and flexibly controls the eccentricity of the equivalent RF of the entire block.

Inspired by this study, considering the shapes of the objects in remote sensing images, the RFAM was designed, as shown in Fig. 4. The RFAM consists of a multibranch structure and a shortcut path structure. In each branch, we use a 1 × 1 convolution layer to compress the number of feature map channels. The sampling center of the intermediate convolution is determined by the last dilated convolution, and its dilation rate can adjust the RF eccentricity of the entire module. The 3 × 3 conv layer in branch1 concentrates on the most important central area. In contrast to the RFB, considering that targets such as ships and vehicles in remote sensing images always have long rectangle shapes, branch2 and branch3 use a 1 × 3 conv layer and 3 × 1 conv layer as the last layer, respectively, each with a dilation rate of 3 to render the RF suitable for these objects. To retain the diagonal information,

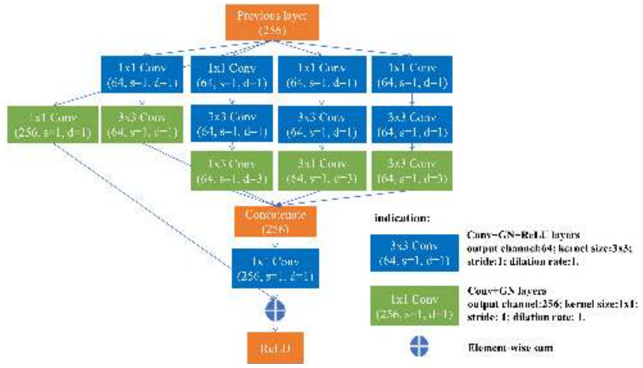
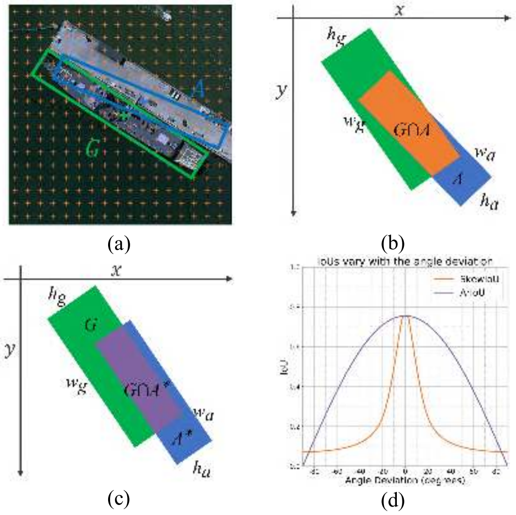


FIGURE 4. Architecture of Receptive Field Amplification Module (RFAM). RFAM uses dilated convolution to control the size and shape of the receptive field.



(a)

(b)

(c)

(d)

FIGURE 5. Comparison of calculation approaches of SkewIoU metric and angle-related IoU (ArIoU) metric. (a) Slender ground-truth *G* and a nearby rotated anchor *A*. The orange crosses represent the centers of the predefined anchors. (b) SkewIoU between *G* and *A*. (c) IoU of *G* and *A**. (d) The SkewIoU and ArIoU scores of *A* and *G* vary with the angle deviation.

branch4 uses a 3 × 3 conv layer with the same dilation rate as the last layer. RFAMs are appended after the feature maps {C6, C5, C4, C3} to extend the ERF of the ARB and ODB.

C. ARIOU AND ANCHOR SETTINGS

1) ARIOU

There are two cases that the IoU calculation is needed in SSD: the first lies in the anchor matching step to distinguish positive and negative anchors; the second is in the NMS procedure to filter out redundant detection boxes. For S^{2}ARN, another IoU calculation is added in the anchor refinement step to match ground-truth boxes with the predefined anchors. The calculation method and the threshold setting of IoU are crucial for SSD-based algorithms.

In the anchor matching step, in many of the oriented object detection methods [22], [23], [26], [28], [30], [31], the convex polygon overlapping area of two rotated boxes is calculated to obtain the IoU, which is known as the SkewIoU metric [30], as shown in Fig. 5(b). When the SkewIoU is applied to a ground-truth box with a high aspect ratio, such as that in Fig. 5(a), the SkewIoU score is sensitive to the change in angle, and a slight angle shift causes a rapid decrease in the IoU score, as shown in Fig. 5(d). Matching slender targets with a sufficient number of anchors is difficult when selecting positive anchors with a conventional positive threshold (such as 0.5), which will decrease the recall rate. One method for alleviating this problem is to reduce the pos-threshold, which will decrease the precision rate of the detector. Another solution is to increase the sampling density of the anchor angle; however, this approach increases the number of anchors and increases the computational burden.

To solve these problems, we applied the angle-related IoU (ArIoU) metric of [29] to calculate the IoU between the oriented ground-truth box *G* and the rotated anchor box *A*, instead of applying the SkewIoU metric. The rotated bounding box is defined by the 5-tuple coordinate (x, y, w, h, θ), where (x, y) represents the geometric center coordinate of the box; w and h are the lengths of the long side of the box and the short side of the box, respectively. The orientation parameter, θ, determining the rotation angle of the bounding box, is defined as the angle between w and the positive x-axis and ranges from 0 to π. The calculation method of the ArIoU is expressed as follows:

$$ArIoU(G, A) = \frac{area(G \cap A^*)}{area(G \cup A^*)} |\cos(\theta_G - \theta_A)| \tag{1}$$

where, G $(x_g, y_g, w_g, h_g, \theta_g)$ is an oriented ground-truth box and A $(x_a, y_a, w_a, h_a, \theta_a)$ is a nearby rotated anchor. A^* is the rotated box which keeps the same parameters as A, with the exception that the angle parameter is θ_g, and is not θ_a. The ArIoU(G, A) monotonically decreases to 0 while the angle deviation increases from 0 degrees to 90 degrees, which forces the anchor with a similar orientation to match the ground-truth box. Compared with the SkewIoU score, the ArIoU score gradually changes with the angle offset. For instance, in Fig. 5(d), with a positive threshold of 0.5, to match A to G, the angle offset of A and G should be less than 10 degrees using the SkewIoU metric. When the ArIoU metric is adopted, the value can be relaxed to 50 degrees, which enables the ground-truth boxes to be matched with more anchors and helps to improve the recall. The ArIoU metric is more robust to a small angle deviation. Thus, we can reduce the sampling interval of the anchor angle to improve computational efficiency while ensuring that each ground-truth box matches adequate anchors. In contrast to the anchor interval of 30 degrees or 60 degrees in [21], [23], [28], we set a rotated anchor every 90 degrees in the ARB.

The ArIoU and SkewIoU metrics are employed in different situations. For the anchor matching step in the training phase, the ArIoU metric is utilized. In the NMS step, SkewIoU scores are calculated to eliminate redundant detection boxes.

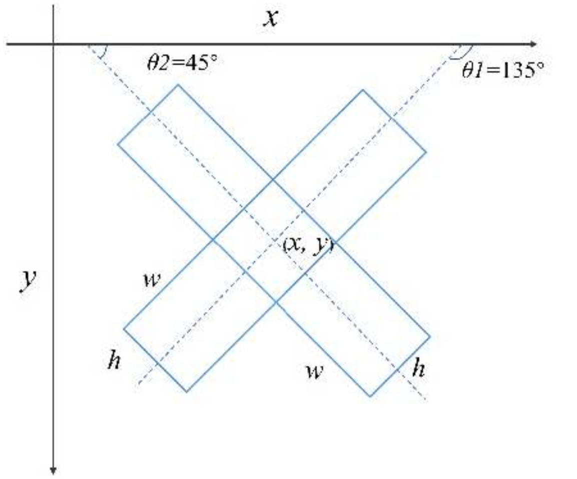


FIGURE 6. Angle settings of the rotated anchors.

2) ANCHOR SETTINGS

In the ARB, we use three parameters, scale, aspect ratio and angle, to generate regular rotated anchors and effectively cover the oriented ground-truth boxes of different shapes. For each refiner {C6, C5, C4, C3}, we define the anchors to have scales of {256, 128, 64, 32} pixels, respectively. Benefiting from the insensitivity of the ArIoU score to small angle offsets, we can set a sparse angle sampling interval for the rotated anchors. We apply two angles {45°, 135°} to control the orientation as shown in Fig. 6. The aspect ratios of the anchors are determined by the shape of the detected target. For ships in the HRSC2016 dataset, multiple aspect ratios of {1:3, 1:5, 1:7} are adopted. For the UCAS-AOD dataset which consists of aircrafts and vehicles, we set the aspect ratios of the anchors to {1:1, 1:2}. For the HRSC2016 dataset, each unit of the refiner has 6 anchors (1 × 2 × 3). For the UCAS-AOD dataset, each unit has 4 anchors (1 × 2 × 2). Although the multiangle anchor is set, the number of anchors on each output unit increases only by one more than that in RPN which has multiple aspect ratios of {2:1, 1:1, 1:2}.

D. ANCHOR MATCHING POLICY AND LOSS FUNCTION

To train the model, we need to distinguish between the positive samples and negative samples from all anchors. The positive anchor needs to satisfy the following conditions: (a) The ArIoU score between the anchor and any ground-truth box is greater than the pos-threshold, simultaneously, the absolute value of the angle deviation should be less than the angle threshold. (b) The anchor has the highest ArIoU score with any ground-truth box. An anchor is assigned a negative label when (a) the ArIoU score is lower than the neg-threshold for all ground-truth boxes or (b) the ArIoU score is greater than the pos-threshold, but the angle deviation is larger than the angle threshold. Unlike RPN, when the ground-truth boxes are associated with anchors, in addition to the IoU constraint, we limited the angle deviation of the matched ground truth and anchor, which enables the anchor with the smallest angle offset to predict corresponding ground-truth. In the ARB, pos-threshold = 0.4, neg-threshold = 0.3 and angle threshold = $\pi/4$ are adopted. For the ODB, pos-threshold = 0.75, neg-threshold = 0.5 and ang threshold = $\pi/8$ are adopted. The threshold setting in the ODB is stricter than that in the ARB. After the first regression, the IoU score between the refined anchor and ground-truth box is relatively high, and the higher threshold encourages the refined anchor with a high localization accuracy to participate in the training process to fit the ground-truth, which is beneficial for suppressing the close false positives and improving the precision rate. Tiny targets are usually harder to match a sufficient number of anchors, leading to low recall. Inspired by the scale compensation anchor matching strategy in [38], for objects whose equivalent scale is less than 40 pixels, we set smaller values for the pos-threshold and neg-threshold. For simplification, all IoU thresholds are reduced by 0.2 for tiny objects.

We use the multitask loss to minimize the objective function, which is defined as (2). Since the ARB is only used to adjust the predefined anchors, and the object category is determined by the ODB, therefore, the classification loss of the ARB is not adopted.

$$\begin{aligned} L &= L_{reg}^{ar} + L_{cls}^{od} + L_{reg}^{od} \\ &= \frac{1}{N_{reg}^{ar}} \sum_{i \in pos} L_{reg}(t_i^{ar}, v_i^*) \\ &\quad + \frac{1}{N_{cls}^{od}} \sum_{j} L_{cls}(c_j^{od}, p_j^\dagger) \\ &\quad + \frac{1}{N_{reg}^{od}} \sum_{j \in pos} L_{reg}(t_j^{od}, v_j^\dagger) \end{aligned} \tag{2}$$

$$L_{cls}(c, p) = \sum_k -p_k \log(c_k) \tag{3}$$

$$L_{reg}(t, v) = \sum_{m \in \{x,y,w,h,\theta\}} smooth_{L1}(t_m - v_m) \tag{4}$$

$$smooth_{L1}(x) = \left\{ \begin{array}{ll} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & otherwise \end{array} \right\} \tag{5}$$

where i and j are the indexes of a predefined anchor and a refined anchor in the ARB and ODB, respectively, and k is the category index for the background class and all objects categories. c_j^{od} represents the predicted probability distribution calculated by the softmax function for the refined anchor j, and $p_j^\dagger$ is the class label of the ground-truth that match with j. The predicted five-tuple parameterized offsets $(t_x, t_y, t_w, t_h, t_\theta)$ of anchor i and refined anchor j are defined as t_i^{ar} and t_j^{od}. The ground truth coordinate offsets v_i^* and $v_j^\dagger$ are encoded by the matched anchors i and j, respectively. The classification loss L_{cls} and the regression loss L_{reg} are defined by (3) and (4). After the anchor matching step, most of anchors are negative, which will overwhelm the training process. We apply hard negative mining to reduce the number of negative samples, which is similar to SSD. The ratio between the negatives and positives is 3:1. The regression loss

L_{reg} is calculated on all positive samples, whereas the classification loss L_{cls} is calculated on the positive samples and the selected negative samples. N_{reg}^{ar} and N_{reg}^{od} represent the number of positive anchors in ARB and ODB, respectively, and N_{cls}^{od} is the sum of the positive anchors and the selected negative anchors in ODB. These parameters are used to normalize the corresponding term in the loss function. The hyperparameter λ controls the balance between the classification task and the regression task and is set to 3. In addition, the ground-truth offset (v_x, v_y, v_w, v_h, v_θ) is encoded by (6):

$$\begin{aligned} &v_x = (x - x_a)/w_a, \quad v_y = (y - y_a) \,/\, h_a, \\ &v_w = \log(w/w_a), \quad v_h = \log(h/h_a), \; v_\theta = \tan(\theta - \theta_a) \end{aligned} \tag{6}$$

The coordinate representation of the ground-truth box, (x, y, h, w, θ), denotes the center coordinates, the width, the height and the angle between the width and positive x-axis, respectively. Similarly, (x_a, y_a, w_a, h_a, θ_a) denotes the parameterized coordinates for a matched rotated anchor or a refined anchor.

III. DATASETS AND EVALUATION INDICATORS

A. DATASETS

We conducted comparative experiments on two public datasets with oriented bounding box annotations, known as the UCAS-AOD [39] and HRSC2016 [40] datasets.

UCAS-AOD. The UCAS-AOD dataset consists of two categories of aircraft and vehicles, each with 1000 and 610 images. These images have two sizes: 1280 pixels × 659 pixels and 1714 pixels × 1176 pixels. All images are collected from Google Earth. The split ratios of the training dataset, validation dataset and test datasets were 50%, 25% and 25%, respectively. The original images were cropped into squares according to the length of the short side with a 50% overlap and resized to 600 pixels × 600 pixels to conserve memory. In addition, we randomly applied the following data augmentation methods during the training phase: horizontal and vertical flipping, random rotation in (0, 90, 180 and 270 degrees) and random translation (within 32 pixels).

HRSC2016. The HRSC2016 dataset is a challenging dataset for ship detection. All images were collected from Google Earth. HRSC2016 contains 1061 labeled images. The image sizes range from 300 pixels × 300 pixels to 1500 pixels × 900 pixels, and most of the sizes are larger than 1000 pixels × 600 pixels. The training dataset, validation dataset and test datasets include 436 images, 181 images and 444images, respectively. We also cropped the images into squares based on the length of the short side and resized them to 600 pixels × 600 pixels. The same data augmentation operations were applied.

B. EVALUATION INDICATORS

To quantitatively evaluate the performance of various object detectors, we utilized the evaluation indicators of recall, precision, and average precision (AP) as well as the precision-recall curve (PRC). To further evaluate the positioning accuracy, we calculated the average IoU (AIoU) scores between the true positive predictions and the matched ground-truth boxes.

1) PRC

The PRC reflects the detection accuracy of the detector at different recall rates. Recall and precision are calculated from the true positive (TP), false positive (FP) and false negative (FN). In object detection task, if a predicted bounding box has an IoU score greater than the threshold (here we chose 0.5) with a ground-truth box of the same category, then it is classified as a TP; otherwise, it is considered to be an FP. Additionally, the redundant predicted boxes that match the same ground-truth box also belong to FP. The ground-truth boxes with no matched predicted boxes constitute FNs. Based on these three components, recall and precision are defined as follows:

$$recall = TP/(TP + FN) \tag{7}$$

$$precision = TP/(TP + FP) \tag{8}$$

2) AP

The AP metric is an evaluation metric that combines recall and precision, which reflects the global performance. AP is the integral of the area under the PRC and the mean average precision (mAP) is the mean of APs across all object classes.

3) AIOU

The Average IoU (AIoU) calculated across all positive predicted bounding boxes and matched ground-truth boxes reflects the localization performance of the detector. We employed the SkewIoU metric to calculate the AIoU score.

IV. EXPERIMENT AND RESULTS

A. IMPLEMENTATION DETAILS

The proposed S^{2}ARN was implemented using the deep learning framework Pytorch 1.0.0 on an Ubuntu 16.04 computer with an Intel® Core$^{\text{TM}}$ i7-6850K CPU and a Nvidia GeForce Titan XP GPU with 12 GB of memory.

We utilized ResNet50 [2] as the backbone. Since the object detection task requires a considerable amount of memory, in our experiment, the GPU holds 16 training images, as indicated in [41], the performance of the batch normalize (BN) layer is influenced by the size of mini-batch. Therefore, we replaced all BN layers in the backbone with group normalize (GN) layers, which behaved more stably. The pretraining weights are provided on the GitHub page of Detectron [42]. The Xavier [43] initialization method was used to initialize the other extra layers. For both datasets, we trained the proposed network for a total of 50k iterations, with a learning rate of 0.001 for the first 30k iterations which decayed to 2e-4 and 4e-5 at 40k iterations and 45k iterations, respectively. The chosen optimizer was the Adam optimizer [44]

TABLE 1. Comparison of the performance of the multiple oriented target detection method for the UCAS-AOD and HRSC2016 validation and test datasets. The bold numbers indicate the highest indicator values for all methods.

Method	Recall (%)			Precision (%)			AP (%)			mAP
	Ship	Plane	Vehicle	Ship	Plane	Vehicle	Ship	Plane	Vehicle	(%)
Baseline	87.0	97.0	91.7	86.5	97.7	89.5	83.6	96.4	87.5	89.2
Baseline + RFAM	89.4	96.9	92.5	86.2	98.0	89.5	85.6	96.5	89.8	90.6
Baseline + ARB	88.4	97.7	92.9	89.7	**98.3**	92.5	87.0	97.2	91.0	91.7
Baseline + RFAM + ARB	**89.5**	**98.0**	**94.3**	88.2	97.5	91.1	**88.1**	**97.6**	**92.2**	**92.6**
DRBox	84.8	95.3	87.1	79.3	97.3	91.1	81.4	94.9	85.0	87.1
YOLOv2-based method	83.3	97.4	87.1	71.7	88.5	70.3	75.6	96.6	79.2	83.8
R-DFPN	87.5	96.2	84.6	**92.1**	98.1	**93.9**	85.0	95.9	82.5	87.8

with a momentum of 0.9, and the batch size was 16 during the training phase.

We performed a series of experiments using the validation and test datasets of the UCAS-AOD and HRSC2016 datasets. The confidence score threshold was set to 0.4 to filter out the background predictions. The SkewIoU metric was used to calculate the IoU score in the NMS procedure and performance evaluation, and the chosen IoU thresholds were 0.2 and 0.5, respectively, due to the small overlap area between rotated bounding boxes. The ResNet-FPN-based SSD without the ARB and RFAM was used as the baseline method, and the "Baseline + ARB + RFEB" architecture represents the proposed S^2ARN. For methods without ARBs, the anchor matching thresholds were pos-threshold $= 0.5$, neg-threshold $= 0.3$ and angle threshold $= \pi/4$, and the remaining settings retained the same as those of the S^2ARN.

Two one-stage methods DRBox [29] and the YOLOv2-based method [30], and a two-stage detector Rotation Dense Feature Pyramid Network (R-DFPN) [26] were adopted for comparative experiments. To ensure the fairness of the experiments, the training parameter settings and the dataset of all methods were consistent. For convenient observation, we combine three categories of objects from the two datasets.

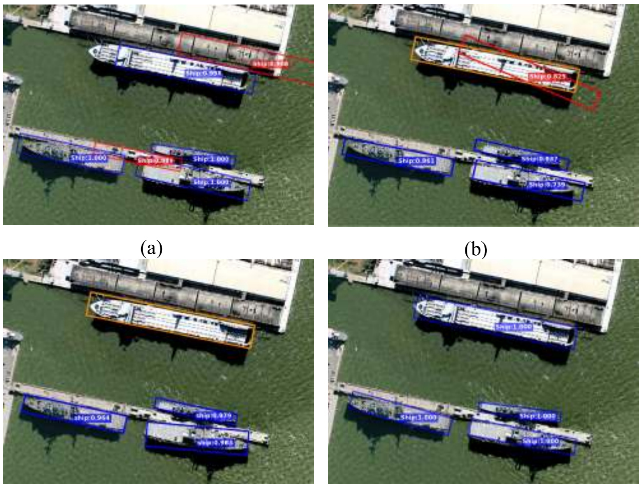

(a)
(b)
(c)
(d)

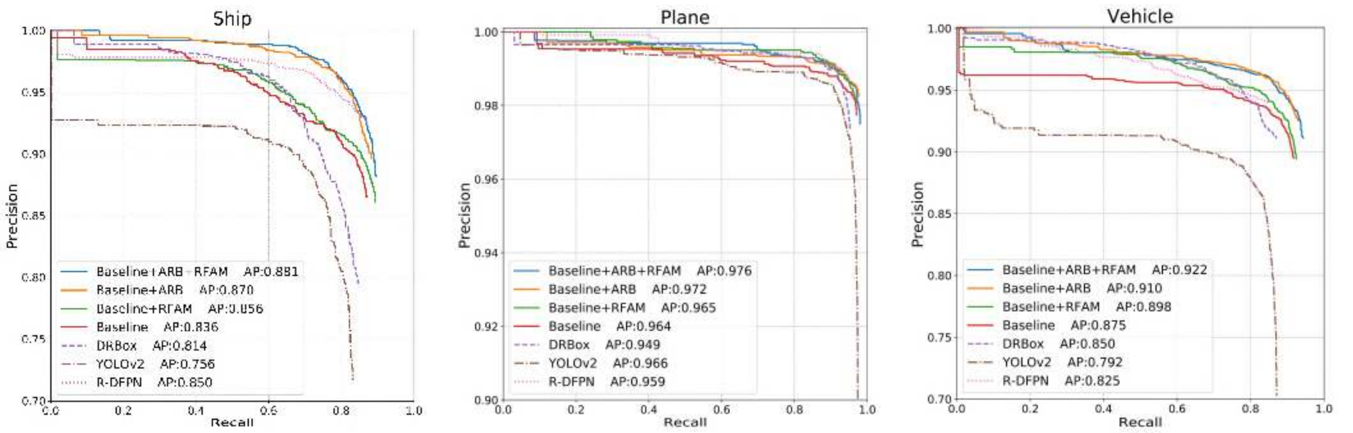
FIGURE 7. Comparison of the detection results of different detectors; blue bounding boxes, red bounding boxes and orange bounding boxes represent true positives, false positives and false negatives, respectively. (a) Detector using rotatable bounding box (DRBox). (b) YOLOv2-based method. (c) Rotation Dense Feature Pyramid Network (R-DFPN). (d) Proposed method. DRBox identifies the shore buildings as ships. The YOLOv2-based method simultaneously generates a false positive and missed detection due to an inaccurate angle regression.

B. RESULTS

As shown in Table 1, our method achieved the best AP performance: 88.1%, 97.6% and 92.2% for the three categories of ship, plane, and vehicle, respectively, while a real-time processing speed was achieved.

After adding RFAMs, mAP increased by 1.4%, which primarily derived from the improved recall from the ship and vehicle categories. Ships docked at ports are easily confused with containers, buildings and wharfs, etc. Similarly, distinguishing cars that are parked on the side of road from shadows and roof vents is difficult. RFAMs increase the effective receptive field of the detection heads of the network, and provide more comprehensive contextual information for the classification subnetwork; thus, the foreground objects are better differentiated, which leads to improved APs.

The "Baseline + ARB" architecture is designed to evaluate the effect of the ARB which utilizes the anchor refinement strategy by adding a 3 × 3 conv layer. The addition of ARBs produced a 2.5% performance improvement, and the AP improved in all categories. The ARB provides the ODB with high-quality refined anchors, which renders the localization performance of the refined anchors consistent with the classification score and alleviates the misalignment between training target and the inference configuration. Two consecutive regressions and the dual threshold setting enable the detector to improve the precision while prevent the recall from decreasing. Due to the prior adjustment of the predefined anchors, the "Baseline + ARB" architecture improves the localization accuracy, as shown in Table 2. The detection bounding boxes that deviate from the ground-truth boxes due to the inaccurate angles are substantially reduced. Compared with the "Baseline" method, the additional computational burden is only derived from the 3 × 3 conv layer in each ARB, which is negligible, and significant performance improvements have been achieved.



(a) (b) (c)

FIGURE 8. Precision-recall curves of different methods for each category. (a) Ship. (b) Plane. (c) Vehicle.

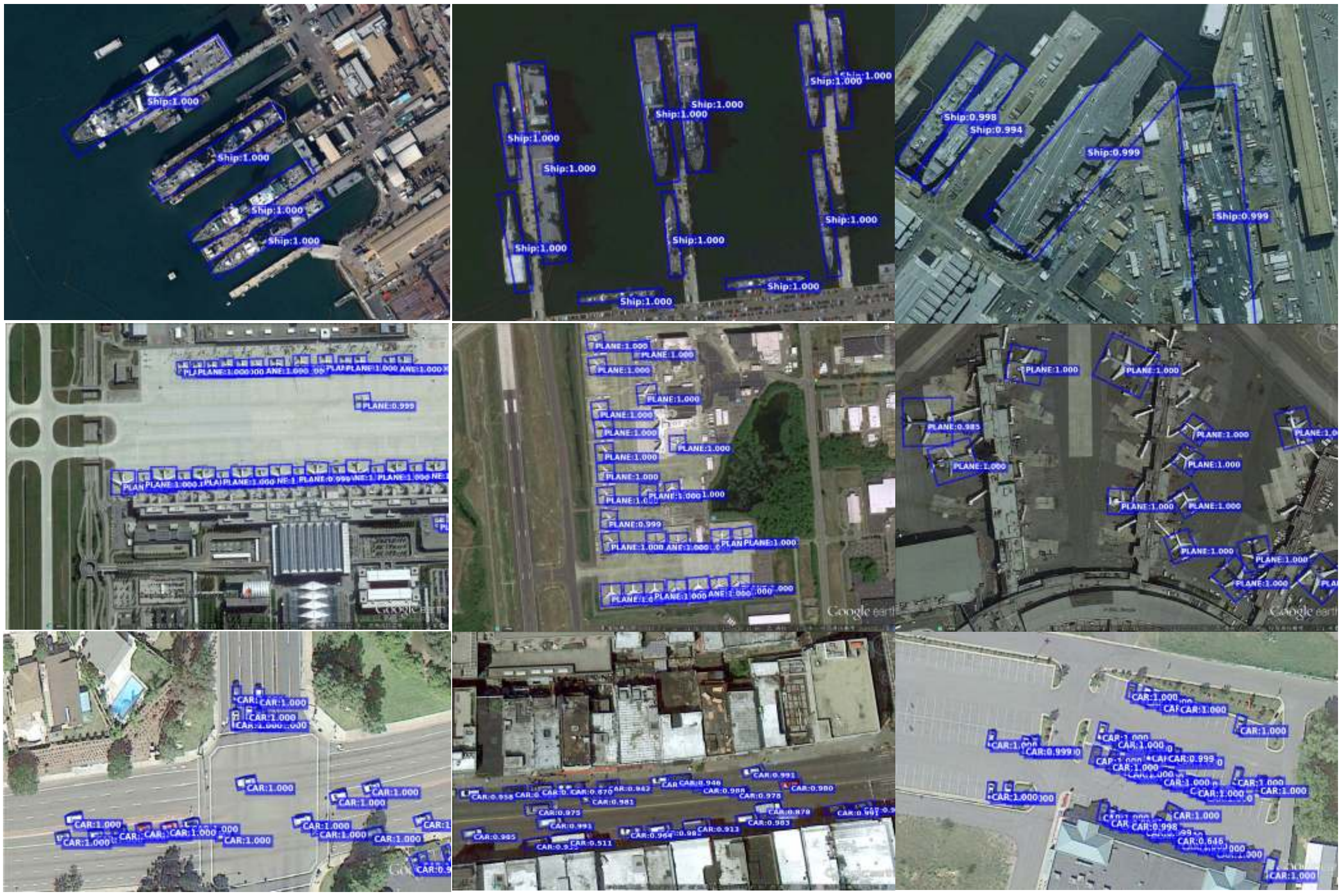


FIGURE 9. Visualization of the detection results of proposed S^{2}ARN in HRSC2016 and UCAS-AOD datasets.

The DRBox is also an SSD-based method, as it uses a VGG16 [45] backbone truncated to the conv4_3 layer, and predicts with a single feature map, leading to a very high detection speed. Due to the small size of the vehicle and the absence of the scale compensation strategy, DRBox has a lower recall rate in the vehicle category. In HRSC2016 dataset, DRBox has a recall of 84.8%. As a result of the limited receptive field, it is difficult to extract sufficient features to effectively distinguish between ships and buildings (as shown in Fig. 7(a)), which causes inferior precision.

The YOLOv2-based method does not perform well on HRSC2016 dataset. The main reason is that the method directly regresses the angle by a sigmoid function, and an

TABLE 2. Comparison of localization performance of the detectors. The Average IoU (AIoU) scores between true positives and the matched ground-truth boxes were used to measure the localization accuracy.

Method		Baseline	Baseline + RFAM	Baseline + ARB	Baseline + RFAM + ARB	DRBox	YoloV2	R-DFPN
AIoU	Ship	0.761	0.774	0.811	**0.821**	0.727	0.730	0.799
	Plane	0.794	0.805	0.823	**0.825**	0.812	0.787	0.815
	vehicle	0.736	0.727	0.744	**0.752**	0.741	0.698	0.746

TABLE 3. Inference time for each method tested on a Nvidia GeForce Titan Xp GPU with a batch size of 1. The size of the input images is 600 × 600 pixels.

Method	Baseline	Baseline + RFAM	Baseline + ARB	Baseline + RFAM + ARB	DRBox	YoloV2	R-DFPN
Inference time (ms)	31.1	39.1	32.4	41.0	**14.4**	18.1	92.8

inaccurate angle regression causes the predicted bounding boxes to deviate from the correct direction, which generates a low IoU score, as shown in Fig. 7(b). These predicitons missed the ground-truth boxes and were determined to be false positives, which further reduces the accuracy.

The R-FFPN is an improved version of Faster-RCNN for rotated object detection. As a two-stage detector, R-FFPN has a high precision, and the total performance is slightly lower than that of S^{2}ARN, which is primarily caused by a lower recall. R-DFPN used multiple feature maps of the dense feature pyramid network to predict rotated proposals. The highest resolution feature map has a stride of 4 pixels, and the angle interval of the rotated anchors is 15 degrees, which cover from -90 to 0 degrees. The high-resolution feature maps and densely sampled anchors increase the computational time.

In Fig. 8, we plot the PRC of each method. S^{2}ARN has superior performance and achieves the best balance between precision and recall, at the same time, S^{2}ARN has the highest localization accuracy and provides more exact bounding boxes as shown in Table 3. Part of the detection results are shown in Fig. 9.

V. CONCLUTION

In this study, we proposed an SSD-based detection method for oriented objects detection in remote sensing images, which is dedicated to improving the detection and localization accuracy with less extra computational cost. We improve the performance of the detector through three efforts. First, since the performance of the original SSD algorithm is affected by the IoU threshold setting, which hinders the ability to achieve a balance between the recall and precision by a single regression, a two-step regression strategy with increased IoU thresholds is proposed to preadjust the coordinates of the predefined anchors to improve the detection and localization accuracy. Second, considering the diversity of the object scales in remote sensing images, the RFAM is introduced to extract more discriminative features for large objects. Last, we solve the problem that slender targets cannot easily match a sufficient number of anchors by deploying ArIoU metric in the anchor matching step, which has high tolerance to the angle deviation and helps to reduce the angle sampling density of the rotated anchors. The experimental results demonstrate the superior performance of the proposed framework for oriented object detection in complex scenes.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

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

[3] 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.*, vol. 2015, pp. 91–99.

[4] 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.

[5] 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.*, Jun. 2016, pp. 779–788.

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

[7] I. Suleymanova, T. Balassa, S. Tripathi, C. Molnar, M. Saarma, and Y. Sidorova, "A deep convolutional neural network approach for astrocyte detection," *Sci. Rep.*, vol. 8, Jan. 2018, Art. no. 12878.

[8] R. Yao, M. Ochoa, P. Yan, and X. Intes, "Net-FLICS: Fast quantitative wide-field fluorescence lifetime imaging with compressed sensing—A deep learning approach," *Light, Sci. Appl.*, vol. 8, p. 26, Mar. 2019.

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

[10] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 784–799.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[13] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*. [Online]. Available: https://arxiv.org/abs/1701.06659

[14] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7263–7271.

[15] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.

[16] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[17] J. Yan, H. Wang, M. Yan, D. Wenhui, X. Sun, and H. Li, "IoU-adaptive deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery," *Remote Sens.*, vol. 11, no. 3, p. 286, Feb. 2019.

[18] S. Chen, R. Zhan, and J. Zhang, "Geospatial object detection in remote sensing imagery based on multiscale single-shot detector with activated semantics," *Remote Sens.*, vol. 10, no. 6, p. 820, 2018.

[19] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, and H. Wang, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*. [Online]. Available: https://arxiv.org/abs/1706.09579

[20] 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.

[21] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," 2018, *arXiv:1807.02700*. [Online]. Available: https://arxiv.org/abs/1807.02700

[22] 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: https://arxiv.org/abs/1812.00155

[23] J. Koo, J. Seo, S. Jeon, J. Choe, and T. Jeon, "RBox-CNN: Rotated bounding box based CNN for ship detection in remote sensing image," in *Proc. 26th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2018, pp. 420–423.

[24] J. O. D. Terrail and F. Jurie, "Faster RER-CNN: Application to the detection of vehicles in aerial images," *arXiv:1809.07628*. [Online]. Available: https://arxiv.org/abs/1809.07628

[25] X. Yang, K. Fu, H. Sun, J. Yang, Z. Guo, and M. Yan, "R2CNN++: Multi-dimensional attention based rotation invariant detector with robust anchor strategy," 2018, *arXiv:1811.07126*. [Online]. Available: https://arxiv.org/abs/1811.07126

[26] 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, 2018.

[27] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *arXiv:1806.04828*. [Online]. Available: https://arxiv.org/abs/1806.04828

[28] 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.

[29] L. Liu, Z. Pan, and B. Lei, "Learning a rotation invariant detector with rotatable bounding box," 2017, *arXiv:1711.09405*. [Online]. Available: https://arxiv.org/abs/1711.09405

[30] W. Liu, L. Ma, and H. Chen, "Arbitrary-oriented ship detection framework in optical remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 6, pp. 937–941, Jun. 2018.

[31] T. Tang, S. Zhou, Z. Deng, L. Lei, and H. Zou, "Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks," *Remote Sens.*, vol. 9, no. 11, p. 1170, 2017.

[32] S. Li, Z. Zhang, B. Li, and C. Li, "Multiscale rotated bounding box-based deep learning method for detecting ship targets in remote sensing images," *Sensors*, vol. 18, no. 8, p. 2702, 2018.

[33] 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.*, Jul. 2017, pp. 2117–2125.

[34] 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.*, Oct. 2017, pp. 2980–2988.

[35] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "Consistent optimization for single-shot object detection," 2019, *arXiv:1901.06563*. [Online]. Available: https://arxiv.org/abs/1901.06563

[36] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4898–4906.

[37] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 404–419.

[38] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S^{3}FD: Single shot scale-invariant face detector," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 192–201.

[39] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3735–3739.

[40] Z. Liu, H. Wang, H. Weng, and L. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1074–1078, Aug. 2016.

[41] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[42] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He, "Detectron," Facebook, Menlo Park, CA, USA, 2018. [Online]. Available: https://github.com/facebookresearch/Detectron

[43] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arxiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

SONGZE BAO was born in Jiamusi, Heilongjiang, China, in 1992. He received the B.S. degree from Jilin University, Changchun, China, in 2015. He is currently pursuing the Ph.D. degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun.

His research interests include computer vision, object detection, and remote sensing image processing.

XING ZHONG received the B.E. degree from Jilin University, Changchun, China, in 2004, and the Ph.D. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, in 2009.

He is currently the Vice President and the Chief Engineer with Chang Guang Satellite Technology Company Ltd. (CGSTL), and also a Full Professor with the University of Chinese Academy of Sciences. His research interest includes satellite's overall design, especially the payload and platform integration technologies.

RUIFEI ZHU received the B.E. degree from Jilin University, Changchun, China, in 2009, and the Ph.D. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, in 2014.

He is currently with Chang Guang Satellite Technology Company Ltd. (CGSTL). His research interests include remote sensing image processing and mining.

XIAONAN ZHANG was born in Yuncheng, Shanxi, China, in 1993. He received the B.E. degree from Harbin Engineering University, Harbin, China, in 2016. He is currently pursuing the master's degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China.

His research interests include object detection and remote sensing image interpretation.

ZHUQIANG LI received the B.S. degree from the China University of Geosciences, Beijing, in 2014, and the master's degree from the School of Geography, Beijing Normal University, Beijing, in 2017.

He is currently with Chang Guang Satellite Technology Company Ltd. (CGSTL). His research interests include remote sensing image classification based on deep learning and 3D urban modeling

MENGYANG LI was born in Heihe, Heilongjiang, China, in 1992. He received the B.S. degree in physics from Jilin University, Changchun, China, in 2015. He is currently pursuing the Ph.D. degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun.

His research interests include image processing, weak target detection, and computer vision.