Hindawi
Journal of Sensors
Volume 2022, Article ID 8515510, 11 pages
https://doi.org/10.1155/2022/8515510

Research Article

OAB-YOLOv5: One-Anchor-Based YOLOv5 for Rotated Object Detection in Remote Sensing Images

Jie Liu, Wensheng Qiao, and Zhaolong Xiong

CETC Key Laboratory of Avionic Information System Technology, Southwest China Institute of Electronic Technology, Chengdu 610036, China

Correspondence should be addressed to Jie Liu; liujie0826@126.com

Received 5 May 2022; Revised 14 November 2022; Accepted 19 November 2022; Published 13 December 2022

Academic Editor: Sushank Chaudhary

Remote sensing images are widely distributed, small in object size, and complex in background, resulting in low accuracy and slow speed of remote sensing image detection. Existing remote sensing object detection is generally based on the detector with anchors. With the proposal of a feature pyramid network (FPN) and focal loss, an anchorless detector emerges, however, the accuracy of anchorless detection is often low. First, this study analyzes the differences and characteristics of the intersection of union (IoU) and shape matchings based on anchors in mainstream algorithms and indicates that in dense or complex scenes, some labels are not easily assigned to positive samples, which leads to detection failure. Subsequently, we proposean one-anchor-based (OAB) object detection algorithm based on the idea of central point sampling in the anchor-free detector. The positive samples and negative samples are defined according to the central point sampling and distance constraint, and an anchor box is preset for each positive sample to accelerate its convergence. It reduces the complexity of the anchor-based detector, improves the inference speed, and reduces the setting of hyperparameters in the traditional matching strategy, rendering the model more flexible. Finally, in order to suppress background noise in remote sensing images, the vision transformer (ViT) is adopted to connect the neck and head, making it easier for the network to pay attention to key information. Thus, it is not easy to lose in the training process. Experiments on challenging public dataset—DOTA dataset- verified the effectiveness of the proposed algorithm. The experimental results show that the mAP of the optimized OAB-YOLOv5 method is improved by 2.79%, the number of parameters is reduced by 13.2%, and the inference time is reduced by 11% compared with the YOLOv5 baseline.

1. Introduction

Object detection plays an important role in the field of computer vision. Remote sensing images have high resolution and optional observation range. Remote sensing object detection provides a new detection method for object detection, which is of significant value in military and national defense security fields. In recent years, object detection has been based on anchored detectors, which can be generally categorized into one-stage detection [1–6] and two-stage detection methods [7–10]. The one-stage method usually places numerous preset-anchor points on the image. Generally, various anchor points with different proportions are preset by clustering, and the coordinates and categories of each anchor box are refined many times. Finally, the screened anchor boxes are considered as the detection results. Compared to the one-stage methods, the two-stage methods refine the anchor boxes with a higher degree and achieve promising results in terms of accuracy, while the one-stage methods maintain faster detection speed. With the emergence of feature pyramid networks (FPN) [11], the accuracy gap between the one-stage and two-stage methods has been narrowed to some extent.

In consideration of numerous preset-anchor boxes by anchor-based detectors, the relevant academic research has gradually shifted from anchor-based detectors to anchor-free detectors. One approach is to locate several predefined or self-learned keypoints and bind the spatial scope of the object, which is called keypoint method [12]. Another approach is to define the positive sample using the center-based or region of the object and predict the four distances (up, down, left, and right) from the positive samples to the object boundary. This type of anchor-free detection is called

a center-based method [13]. It eliminates the hyperparameters related to anchors and has a generalization ability.

However, the performance of anchor-free detectors cannot catch up with anchor-based ones at present. There are two main differences between anchor-based and anchor-free detectors. We take RetinaNet [14], YOLOv3 [6], and FCOS [13] as examples to illustrate the differences between anchor-based and anchor-free detectors. (1) Allocation strategy of positive samples was as follows: RetinaNet based on IoU filter strategy takes positive samples if the IoU value of default anchors and ground truth (GT) IoU is greater than the threshold. YOLOv3 compares the ratio of the width and height of the anchors to that of GT. If the ratio is less than the set hyperparameters of width and height ratio, it is a positive sample. FCOS takes all points in a ground truth bounding box as positive samples. (2) Targets for regression, RetinaNet, and YOLOv3 algorithms are regression of the offset of the border relative to the anchors, while FCOS is to regress the distance of the upper left corner point and the lower right corner point relative to the anchor point. For the moment, anchor-based detectors achieve high performance.

Most anchor-based detectors densely preset anchors at each location of the feature map with three different scales. Particularly, additional anchors are set according to different angle intervals for oriented arbitrary objects with additional angle settings. Numerous preset anchors lead to an extreme imbalance between positive and negative samples. The most common solution is to control the candidate ratio through a specific sampling strategy [15, 16]. Both of them have the problem of uneven positive and negative samples. Some scholars have made some researches on this problem. For example, ATSS [17] and dynamic R-CNN [18] adaptively select high-quality positive samples. However, the study above only considers the noise of positive samples and ignores the potential localization ability of numerous negative samples and the credibility of IoU. HAMBox [19] shows that low-quality negative samples can achieve high-quality positioning. ATSS [17], DAL [20], and FCOS [13] show that adding high-quality positive sample anchors significantly accelerates convergence.

In aerial image scenes, the shooting angle of the image is generally a top-view angle. In contrast, the objects of interest, such as cars, planes, and ships, are usually relatively small and occupy only a few pixels of the image. According to DOTA [21], remote sensing images have the following challenges. (1) Complex background: aerial images usually contain complex scenes, and the target is easily surrounded by scenes, resulting in missed or false detections; (2) huge scale variations: the scale of the target varies greatly; (3) dense arrangement: the detected objects are sometimes densely or sparsely arranged; and (4) small objects. We referred to the MS COCO [11] definition of large, medium, and small targets; approximately 60% of the targets in DOTA have less than 50 pixels.

Due to the complex background and the huge variation in the orientation, scale, and appearance of the object instances in remote sensing images, it is difficult to apply the horizontal detection algorithms to rotated object detection. In order to predict the location and orientation of the rotated objects in remote sensing images, previous rotation detection algorithms [22–29] use preset rotation anchors and additional angle prediction. Owing to changes in orientation, numerous anchors should be preset on the feature map making them spatially aligned with GT boxes. Other methods use horizontal anchor points to detect rotating objects. For example, RoI transformer [23] uses horizontal anchor points but learning the RoI of rotation through spatial transformations reduces the number of predefined anchors to some extent. Rotate-YOLOv5 [29] uses CIoU as the loss function of the bounding box and mosaic data enhancements to improve the detection accuracy on the basis of ensuring the detection speed. R3Det [30] recodes modules using cascading regression and redefinition boxes to achieve high performance. Although this method achieves high performance, it must lay numerous anchor frames on the feature graph. However, there is significant redundancy in the distribution of anchor frames in the rotation scenario.

In this study, the DOTA dataset is representative and challenging, and we discuss the proposed method based on that dataset. The problem discussed is universal in detection algorithms. Inspired by FCOS [13], YOLO [4], ATSS [17], and Rotate-YOLOv5 [29], we analyze not only the characteristics of the existing mainstream algorithms for positive and negative sample sampling strategies but also the advantages of anchor-based and anchor-free methods. Meanwhile, we propose a remote sensing object detection algorithm based on a one anchor-based method. It optimizes the problems of IoU or shape-matching strategy and reduces the design of hyperparameters. Experiments were performed on the DOTA dataset to support the analysis and conclusions. The main contributions of this study are as follows:

(i) The characteristics of the matching strategy based on IoU and shape are analyzed, and it is not necessary to set the anchor frame with multiple proportions on the same anchor point

(ii) Combining the idea of anchor-based and anchor-free methods, a screening strategy for positive and negative samples based on the one anchor-based (OAB) method is proposed

(iii) The self-attention mechanism of the vision transformer is introduced to weaken the complex background information in remote sensing scenes, strengthen the extraction of useful information, and increase the overall detection performance

2. Proposed Methods

2.1. Network. The object detection of remote sensing images must consider both efficiency and accuracy, and the algorithm has good portability. As an improved version of YOLOv3 [6] and YOLOv4 [1], YOLOv5 has similar basic architecture and good algorithm portability. The YOLOv5 method was chosen as the baseline to meet both the detection performance and speed. The pipeline of the network structure is illustrated in Figure 1. We used cross-stage

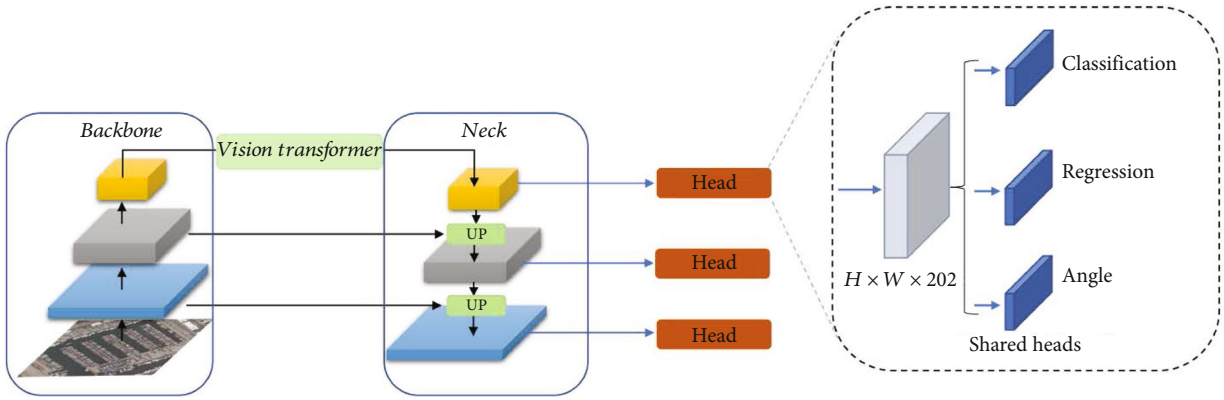


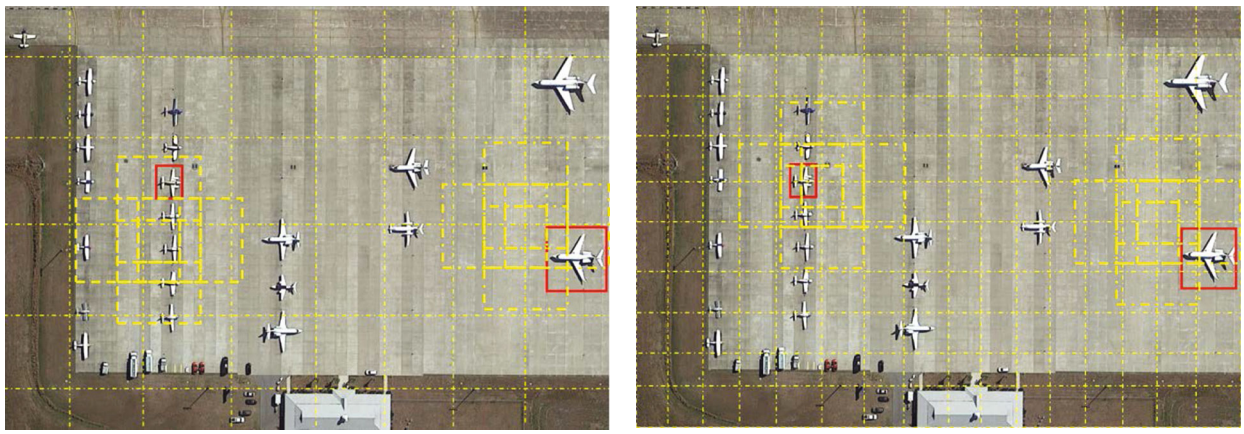
Figure 1: The network architecture of OAB-YOLOv5.

partial connections (CSP) [1] as backbone. At the top of the backbone network, we added a vision transformer (ViT) [31] module to connect to the top of the neck. This allows the network to focus on key information and better learn specific target features. For the detection head part of the network, we added a point-spacing branch to each layer to suppress the regression box far from the center point of GT and improve the detection accuracy. This is similar to the centerness of FCOS [13].

2.2. One-Anchor-Based Method for YOLOv5. One of the important parts of anchor-based target detector is the sampling strategy of positive and negative samples. Currently, there are two mainstream sampling strategies to collect and distinguish it, one is the sampling strategy based on IOU, and the other is the sampling strategy based on shape. The sampling strategy based on IOU sets the IOU threshold and combines the sampling step. When the IOU value of anchor and GT is greater than the set threshold, it is considered to collect a positive sample. The sampling step can control the number of anchors. The smaller the step, the more anchors will be generated, and the more positive samples will be matched, but at the same time, the more redundant negative samples will be collected. The number of positive and negative samples is also smaller, so, the sampling threshold of positive and negative samples based on IOU matching need to be set reasonably. It is hard to do and easy to lead to the loss of small targets, and an imbalance of positive and negative samples exists, especially in remote sensing images. The sampling strategy based on shape matching is relatively simple, but this method is more flexible and has fewer hyperparameters. Due to the unreasonable anchors setting by the sampling strategy based on IOU, a certain GT has no anchor to correspond to and becomes the one that ignores the region. It can be seen that this allocation system will lead to relatively few positive samples. It is guaranteed that each GT box must have a unique anchor by the sampling strategy based on shape. The threshold is not fully considered. By comparing the anchor aspect ratio and threshold, the sample is positive within the maximum IOU value. Even if the maximum IoU is less than ignore threshold, it will not affect the prediction box to be a positive sample. Otherwise, it is negative. However, more anchor frames need to be preset to match targets of different scales. Due to the different sizes of targets in the real environment, a large number of anchor aspect ratios will be set in advance to be more appropriate and real, which will increase the large amount of calculation and result in low calculation efficiency. In this section, we analyze the differences between the IoU and shape label collection methods. Subsequently, we solve the problem of IoU and shape label collection using the OAB method. Finally, we introduce the self-attention mechanism of ViT [31] to enhance the global reasoning ability of the network to the feature map to detect the accuracy.

2.2.1. Label Assignment Based on IoU and Shape Strategy

(1) Based on the IoU Strategy. As shown in Figure 2, red represents GT box, yellow represents grid of feature graph divided according to different sampling stride, and stride represents sampling stride. The FPN generates feature maps of large, medium, and small scales; each scale feature map can predict the target of the corresponding scale. In the sampling process, the sampling step of the anchor frame expands with a decrease in the resolution of the feature graph. Generally, for feature maps of large, medium, and small targets, the sampling step size is set to 8, 16, and 32. This study takes stride = 8 and 16 as examples to analyze the influence of different sampling step sizes on different scale targets. Specifically, when stride = 16 or 8, the preset-anchor frames of different proportions are laid at the center point of each yellow grid, and the IoU between these preset anchors and GT box is calculated. The positive and negative samples are obtained for boundary box regression and classification by setting the IoU threshold to divide the positive and negative samples. The division of positive and negative samples involves two hyperparameters: positive sample IoU threshold (pos_iou_thres) and negative sample IoU threshold (neg_iou_thres). Assume pos_iou_thres = 0.5 and neg_iou_thres = 0.3. In Figure 2(a), most of the IoUs generated by the anchor frame are smaller than (neg_iou_thres) and are regarded as negative samples. In Figure 2(b), the entire graph is divided into denser grids by reducing the sampling step size. When the generated anchor frame is matched with the GT box, more positive samples are matched. However, the number of redundant positive samples increases.



(a) Stride = 16 (b) Stride = 8

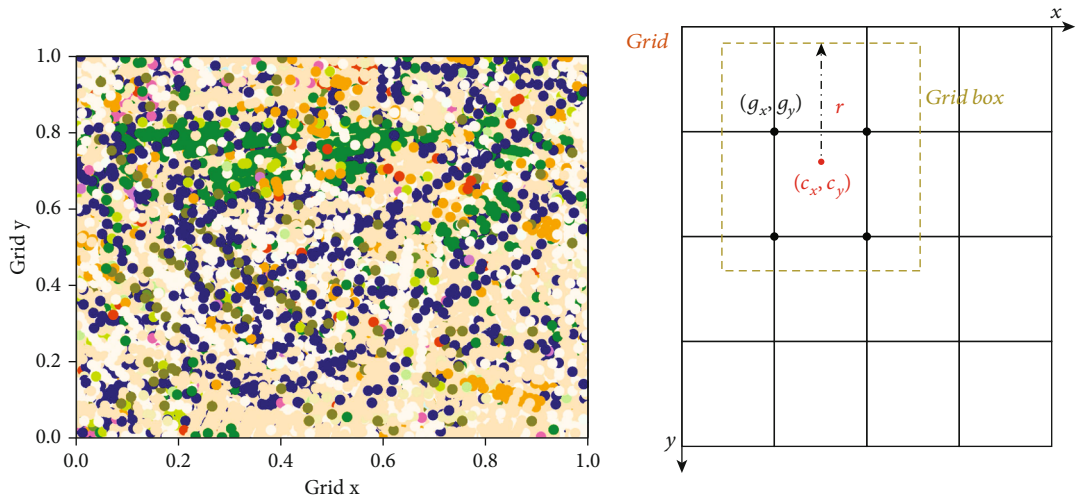
Figure 2: Diagram of matching based on IoU strategy.

Therefore, the setting of IoU hyperparameters significantly affects the number of positive and negative samples. Additionally, with a decrease in the resolution of the feature map, an increase in the sampling step size leads to the loss of small targets, and an imbalance of positive and negative samples exists, especially in remote sensing images.

(2) Based on the Shape Strategy. As shown in Figure 3, two GT boxes with large-scale differences are listed to illustrate the problems existing in the shape-based matching strategy. Red represents GT box. Based on the shape matching strategy, the ratio between the width and height of the preset-anchor frame and that of the GT box is calculated. Subsequently, the hyperparameter threshold (anchor_ratio_thres) is set according to this ratio to divide the positive and negative samples. If the aspect ratio between the preset-anchor frame and GT box is between (1/anchor_ratio_thres, anchor_ratio_thres), this part of the sample is positive. The GT box in the upper left corner is a small target, whereas the lower right corner is a large target. Red represents the default anchor frame. It is discovered that the aspect ratio of the default anchor frame is very different from the red GT box in the upper left corner. Therefore, such small targets are likely to be ignored, resulting in no positive sample to predict them, while the aircraft in the lower right corner is well matched. The shape-based matching strategy matches more positive samples by setting a larger range of aspect ratios. Compared with the IoU-based matching strategy, this method is more flexible and has fewer hyperparameters. However, more anchor frames need to be preset to match targets of different scales. In the real world, especially in aerial images, the target scale varies significantly, and there are targets that are very large or small. Therefore, once the range of the aspect ratio is set improperly, some objects lose positive samples, resulting in poor detection performance of the corresponding categories.

Figure 3: Diagram of matching based on shape strategy.

2.2.2. One-Anchor-Based Sampling Strategy. During data preprocessing, the coordinates of the GT were normalized. We counted the distribution of the coordinates after normalization and filled them into grid points of 1×1. The results are shown in Figure 4(a). We found that most objects were located in the center of the grid. According to this finding, we chose the intersection of grids around each GT center point as the center of the positive sample, instead of the center of each grid, to speed up the convergence rate of regression. As shown in Figure 4(b), the stride size of each layer was set to 1. The feature map of each layer of the FPN was divided into grids $N \times N$ grids, and the center point (g_x, g_y) of each lattice point in the grid was calculated. For the center point (c_x, c_y) of each real label, a rectangle with (fixed value) radius $r = 1$ was generated around it, which is defined as grid box. Furthermore, if the location (g_x, g_y) falls within the range of the grid box, the location is regarded as a positive sample, and the category label of the location is obj (foreground class). Otherwise, it is a negative sample and



(a) Normalized label coordinates of GT (b) OAB matched method

FIGURE 4: The normalized coordinate distribution of GT and matching method of OAB.

obj = 0 (background class). In addition to classification, there is a 5-dimensional real vector $t = (t_x, t_y, t_w, t_h, \text{obj})$ as the regression target for this position. Notably, the coordinate regression range of bounding box (bbox) in YOLOv5 is $-0.5 \sim 1.5$, which was used for sample expansion. In the proposed method, by changing the sampling method, the regression range of (x, y) coordinates is $-1 \sim 1$.

As shown in Figure 5, the regression of width and height follows for YOLOv3 [6]. If the cell is offset from the top-left corner of the image by (cx, cy) and the bounding box prior has p_w,p_h, then the inference regression targets for the location can be formulated as

$$\begin{aligned} b_x &= 2\sigma(t_x) - 1.0 + g_x, \\ b_y &= 2\sigma(t_y) - 1.0 + g_y, \\ b_w &= p_w e^{t_w}, \\ b_h &= p_w e^{t_h}. \end{aligned} \tag{1}$$

2.3. Vision Transformer. Generally, the background of aerial datasets is complex, which reduces the localization ability of the model. The self-attention mechanism of ViT [31] enables the network to perform global reasoning on the image and on the predicted specific target. The model is used to observe other areas of the image to help determine the target in the bounding box. On the contrary, traditional detection models can only predict each target in isolation. Therefore, we introduce ViT [31] to suppress background noise and strengthen the positioning ability of the model.

2.4. Loss Function

2.4.1. Regression Loss. In DOTA [21] dataset, most targets belong to small targets, and they are arranged intensively. While the IoU evaluates the predicted box as a unit of measurement for the whole, the traditional IoU method only

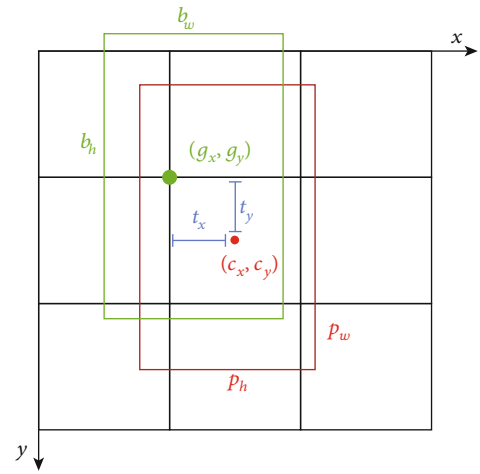


FIGURE 5: Bounding boxes with dimension priors and location prediction.

TABLE 1: Parameter settings.

Parameter name	Parameter value
Initial learning rate	0.01
Epoch	150
Batch size	8
Optimizer	SGD
Momentum	0.937
weight_decay	0.0005
warmup_epochs	3.0
warmup_momentum	0.8
warmup_bias_lr	0.1
lr_factor	0.2
Nms	0.35
Mosaic	1.0

Table 2: Performance comparisons on DOTA-v1.5 test (OBB task) (%).

Methods	Params	OBB results																
		PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC	mAP
DCL [36]	33.4 M	79.09	72.7	37.85	62.67	46.91	50.65	73.21	89.41	72.53	59.62	51.99	68.81	52.36	65.56	56.49	10.23	59.40
RSDet [23]	33.2 M	79.26	79.67	41.61	67.03	48.42	53.41	73.55	89.30	75.08	63.41	50.65	68.50	61.96	64.27	55.08	11.44	64.40
GWD [37]	33.1 M	79.39	74.46	41.95	60.15	49.97	60.03	76.17	90.31	71.75	58.48	48.04	67.8	55.8	65.79	51.35	9.99	60.00
KLD [38]	41.8 M	80.31	74.33	47.87	60.14	65.98	73.11	87.1	89.55	75.38	81.80	49.98	71.8	64.34	71.75	55.97	0.00	65.60
R2-CNN [39]	41.8 M	80.38	78.7	47.95	62.55	65.6	71.33	86.34	89.74	76.29	76.23	49.73	67.55	63.47	73.14	58.44	15.66	66.40
Rotate-YOLOv5 [29]	41.8 M	87.82	78.13	45.60	58.73	66.70	76.21	88.69	90.69	79.59	78.05	41.70	70.40	65.61	75.47	54.53	10.86	66.80
RetinaNet [14]	—	71.43	77.64	42.12	64.65	44.53	56.79	73.31	90.84	76.02	59.96	46.95	69.24	59.65	64.52	48.06	0.83	59.16
MR [40]	—	76.84	73.51	49.9	57.8	51.31	71.34	79.75	90.46	74.21	66.07	46.21	70.61	63.07	64.46	57.81	9.42	62.67
CMR [41]	—	77.77	74.62	51.09	63.44	51.64	72.9	79.99	90.35	74.9	67.58	49.54	72.85	64.19	64.88	55.87	3.02	63.41
FR OBB [21]	—	71.89	74.47	44.45	59.87	51.28	68.98	79.37	90.78	77.38	67.5	47.75	69.72	61.22	65.28	60.47	1.54	62.00
FR H-OBB [21]	—	71.57	74.71	46.39	63.4	51.54	70.11	79.09	90.63	76.81	67.4	48.66	70.9	63.1	65.67	56.66	4.55	62.57
FR OBB + RT [23]	—	71.92	76.07	51.87	69.24	52.05	75.18	80.72	90.53	78.58	68.26	49.18	71.74	67.51	65.53	62.16	9.99	65.03
OAB-YOLOv5 (ours)	**20.3 M**	80.67	77.2	46.37	60.13	66.08	75.80	87.85	90.64	78.12	79.54	46.28	70.58	67.70	75.42	53.63	**28.15**	**68.02**

The short names for categories are defined as follows: PL: plane; BD: baseball diamond; BR: bridge; GTF: ground field track; SV: small vehicle; LV: large vehicle; SH: ship; TC: tennis court; BC: basketball court; ST: storage tank; SBF: soccerball field; RA: roundabout; HA: harbor; SP: swimming pool; HC: helicopter; CC: container crane.

Table 3: Comparisons between baseline method and the proposed YOLOv5m + OAB on DOTA-v1.5 test-dev (OBB task).

		OBB results																
Methods	Params	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC	mAP
YOLOv5m	23.4 M	74.79	74.80	46.99	55.73	66.42	76.44	88.41	86.78	73.53	80.65	40.85	65.30	69.88	76.44	48.56	18.15	65.23
+OAB (without ViT)	22.4 M	79.68	77.2	46.89	53.63	66.29	75.67	88.49	90.54	76.03	78.28	46.65	62.79	67.90	77.50	45.78	22.60	65.81
+OAB (with ViT)	20.3 M	80.67	77.2	46.37	60.13	66.08	75.80	87.85	90.64	78.12	79.54	46.28	70.58	67.70	75.42	53.63	28.15	68.02

OAB: one-anchor-based method; ViT: vision transformer.

Table 4: Inference time evaluation for the three components on DOTA-v1.5 test Results from RTX 3090.

Method	Image size	Inference	NMS	Total
YOLOv5m	1024×1024	12.4	12.3	24.7 ms
OAB-YOLOv5m (ours)	1024×1024	12.4	9.6	22.0 ms

considers the overlap area. According to the characteristics of the DOTA [21] dataset, the overlap area, central point distance, and aspect ratio of the bounding boxes are considered comprehensively. Therefore, CIoU loss [32] is adopted to perform the regression of the boundary boxes, and the loss function can be defined as follows.

$$\mathscr{L}_{\mathrm{CIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2\left(\mathbf{b}, \mathbf{b}^{gt}\right)}{c^2} + \alpha v. \tag{2}$$

And the trade-off parameter α is defined as

$$\alpha = \frac{v}{(1 - IoU) + v}. \tag{3}$$

2.4.2. Angle Loss. Angle regression is a difficult problem in rotation tasks. Therefore, we introduce CSL [33] as the angle regression method and apply it to the baseline YOLOv5 and the proposed method. The CSL [33] method cleverly transforms the angle prediction task from a regression problem to a classification problem to solve the discontinuous boundary problem in a rotating detector. Please refer to [33] for further details. Finally, the expression of the angle regression is as follows:

$$t'_\theta = \left(\theta' - \theta_a\right) \cdot \frac{\pi}{180} \tag{4}$$

Variables θ', θ_a, and t'_θ are for the ground truth angle, anchor angle, and predicted angle, respectively.

3. Experiment Results and Discussion

This section is divided into subheadings. It provides a concise and precise description and interpretation of the experimental results.

3.1. DOTA Dataset and Parameter Settings

3.1.1. DOTA-v1.5. DOTA [21] is a large-scale dataset for object detection in remote sensing images. DOTA-v1.0 contains 2806 large aerial images with the size ranges from 800×800 to 4000×4000 and 188,282 instances among 15 common categories. DOTA-v1.5 uses the same images as DOTA-v1.0 and more extremely small instances (less than 10 pixels). Moreover, a new category "container crane" is added. DOTA-v1.5 contains 403,318 instances in total. Thus, DOTA-v1.5 is more challenging than DOTA-v1.0.

The version of DOTA dataset used in this experiment is DOTA-v1.5. The proportion of the training set, validation set, and test set in DOTA-v1.5 is 1/2, 1/6, and 1/3, respectively. Meanwhile, we crop a series of 1024×1024 patches from the original images with an overlap of 200 pixels by DOTA development kit. Subsequently, the subimages that do not contain the targets are ignored.

3.1.2. Implementation Details. The DOTA-v1.5 is trained by 120 epochs in total with YOLOv5m as the pretraining model. The initial learning rate is 0.01, and the cosine annealing learning rate schedule is utilized to update learning rate. The weight decay is set to 0.0005. The SGD momentum is set to 0.937. Besides, the warm-up strategy is adopted to find a suitable learning rate in the third epoch during training. And other experimental parameters were set as shown in Table 1. The patches of training and test patches were 1024×1024. During inference, we first send patches (the same settings as training) to obtain the detection results before merging, then map the detected results from patch coordinates to the original image coordinates, and perform nonmaximum suppression (NMS) on these results through the original image coordinates. Referring to benchmarks [34, 35], we set different NMS thresholds for each class, "roundabout" is set to 0.1, "tennis-court" is set to 0.3, "swimming-pool" is set to 0.1, "storage-tank" is set to 0.2, "soccer-ball-field" is set to 0.3, "small-vehicle' is set to 0.2, "ship" is set to 0.2, "plane" is set to 0.3, "large-vehicle" is set to 0.1, "helicopter" is set to 0.2,"harbor" is set to 0.0001, "ground-track-field" is set to 0.3, "bridge" is set to 0.0001, "basketball-court" is set to 0.3, "baseball-diamond" is set to 0.3, "container-crane" is set to 0.05, and limit the maximum number of predicted targets for the experiment to 1000. After interference, the detection results are submitted to the DOTA official website at https://captain-whu.github.io/DOTA/evaluation.htmlfor online evaluation of the test dataset and compare with the mainstream SOTA methods. The evaluation index is the average value of each category of average precision (mean average precision, mAP). And the expression of mAP is as follows:

$$\mathrm{AP} = \int_0^1 \mathrm{PR}(r)\mathrm{d}r, \tag{5}$$

$$\mathrm{mAP} = \frac{\sum_K^{i=1} \mathrm{AP}_i}{K}, \tag{6}$$

where AP is the average accuracy of each category, obtained by integrating $\mathrm{PR}(r)$curves which is combined with precision and recall, AP_i represents the average precision in class i, and K represents the number of classes.

3.2. Experiment Result. In this section, we present the training and evaluation of the proposed model using the DOTA-v1.5 dataset. It was deployed in the PyTorch1.7 framework. All experiments were implemented with Intel (*R*) Xeon (*R*) Silver 4114 CPU@ 2.20GHz, one NVIDIA GeForce RTX 3090, and 64 GB of memory.

3.2.1. Result on DOTA-v1.5. The YOLOv5 model with an improved sampling strategy is compared with the mainstream SOTA methods. Table 2 lists the detection performance of the improved YOLOv5 method and the mainstream SOTA method on the DOTA dataset (both one-stage algorithms

(a) Baseline method

(b) Proposal method

Figure 6: Visualization of the results on DOTA-v1.5 between baseline and proposal method.

and two-stage algorithms are included), and the evaluation index is the average value of each category of AP (mAP). We compare the ten peer techniques, including DCL [36], RSDet [23], GWD [37], KLD [38], R2-CNN [39], Rotate-YOLOv5 [29], RetinaNet [14], MR [40], CMR [41], and FR OBB [21] on DOTA-v1.5. Specifically, Zhuang et al. [29] proposed Rotate-YOLOv5 which is one of our most relevant work. We proposed one-anchor-based method as new sampling strategy that can better balance the positive and negative samples of small targets. At the same time, we add a ViT between backbone and neck to reduce background interference and increase the focus on the target, while they used the mosaic data enhancement to enrich the dataset and improve the detection accuracy of small targets. And then, they used the long-edge definition method based on circular smoothing labels to achieve a rotatable bounding box, which solved the effect of angle periodicity on training by converting the regression problem into a classification problem. Finally, they used the CIoU loss as the loss function of the bounding box to improve the detection accuracy on the basis of ensuring the detection speed. In this work, we also care about rotate object detection that is based on the DOTA-v1.5 dataset and evaluated on the official website to ensure the reliability of the experimental results. In addition, the results in Table 2 show the effectiveness and superiority of our proposed method. In contrast, class CC has the lowest number of instances in DOTA-v1.5, where the mAP of the "CC" class for all but the KLD method is very low, near to 0%. We believe that this is because of the imbalance of the positive and negative samples caused by the sampling strategy. The OAB method proposed in this study ensures that the sample ratio of each label is stable and does not cause the instability of the positive and negative sample ratios owing to the setting of the IoU threshold and the object size, thus mitigating the long-tail effect caused by imbalanced sample sizes between categories. As shown in Table 2, OAB-YOLOv5 achieves the highest mAP of 28.15% of the "CC" class. The experimental results show that OAB-YOLOv5 has very excellent performances in the field of oriented object detection.

3.2.2. Ablation Study. To further demonstrate the effectiveness of the proposed sampling strategy and the influence of the ViT [31] module on the overall performance, we compared the influence of the proposed sampling method on the detection performance. Without using the ViT [31] module, the sampling method in this study achieved comparable performance to the baseline and achieved a 5.45% improvement for category "CC." In all, the maps for the 16 categories were the same. In all, the maps for the 16 categories were the same. Thus, we reduced the number of parameters by 1 M and did not need to set additional hyperparameters during

the sampling phase. Therefore, we achieved results similar to that of the baseline method with less complexity, which proves the effectiveness of the proposed sampling method. Finally, we introduce the ViT [31] to reduce the interference of background factors in the remote sensing images, enable the network to study the overall image, and strengthen the ability of the model. The precision of the algorithm is further tested to prove the effectiveness of the overall approach. Table 3 shows the experimental results.

Finally, we retrospectively evaluated the speed of the baseline method and the proposed method. The test resolution was 1024×1024, and the batch size was 8. The results are shown in Table 4; in terms of reasoning time and NMS time, the method in this study reduces the number design of the anchors. Therefore, the reasoning time on NMS is reduced by approximately 11% compared with the baseline method. In conclusion, the proposed method achieves a better balance in speed and accuracy than the baseline method.

4. Detection Effect and Analysis

The detection effect of some categories is visualized at the end of this study. The detection confidence and IoU threshold are set to 0.1 and 0.6, respectively. The specific results are shown in Figures 6(a) and 6(b). Test effect diagram of the proposed method in the right column shows that the test results are clearer and better than the baseline method. In the detection result graph obtained using the baseline method, there are many disorderly anchor frames because each label presets a variety of anchors of different scales. Therefore, there are some redundant detection anchor frames. However, the method in this study is simpler and more efficient.

5. Conclusions

In this study, we proposed a screening strategy based on a single anchor frame to achieve high-performance arbitrary direction remote sensing object detection. Specifically, the characteristics of the two matching methods based on IoU and shape are analyzed, and their shortcomings are identified. Therefore, it is unnecessary to preset multiple anchors. It presets one-anchor-based (OAB) by combining the two ideas of anchor-based and anchor-free and adopts the central point method for sampling. To obtain high-quality samples, the grid points around each real label were calculated as the sampling benchmark, which reduced the hyperparameter design of the matching part and ensured that each GT had a corresponding positive sample for prediction. The validity of this idea was verified using the challenging DOTA dataset.

Data Availability

Datasets are available from https://captain-whu.github.io/DOTA/dataset.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.

[2] H. Huang, Q. Liang, D. Luo, and D. H. Lee, "Attention-enhanced one-stage algorithm for traffic sign detection and recognition," *Journal of Sensors*, vol. 2022, Article ID 3705256, 8 pages, 2022.

[3] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *European Conference on Computer Vision*, pp. 21–37, Springer, 2016.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, 2016.

[5] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, Honolulu, HI, USA, 2017.

[6] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[7] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision*, pp. 1440–1448, Santiago, Chile, 2015.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[11] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, Honolulu, HI, USA, 2017.

[12] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: keypoint triplets for object detection," in *2019 IEEE/CVF International Conference on Computer Vision*, pp. 6569–6578, Seoul, Korea (South), 2019.

[13] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: fully convolutional one-stage object detection," in *2019 IEEE/CVF International Conference on Computer Vision*, pp. 9627–9636, Seoul, Korea (South)., 2019.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision*, pp. 2980–2988, Venice, Italy, 2017.

[15] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: towards balanced learning for object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 821–830, Long Beach, CA, USA, 2019.

[16] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 761–769, Las Vegas, NV, USA, 2016.

[17] S. Zhang, C. Cheng, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9759–9768, Seattle, WA, USA, 2020.

[18] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic r-cnn: towards high quality object detection via dynamic training," in *European Conference on Computer Vision*, pp. 260–275, Springer, 2020.

[19] X. Yang Liu, J. H. Tang, J. Liu, D. Rui, and W. Xiang, "Hambox: delving into mining highquality anchors on face detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13043–13051, Seattle, WA, USA, 2020.

[20] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, pp. 2355–2363, 2021.

[21] G.-S. Xia, X. Bai, J. Ding et al., "Dota: a large-scale dataset for object detection in aerial images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, Salt Lake City, UT, USA, 2018.

[22] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multiclass object detection in unconstrained remote sensing imagery," in *Asian Conference on Computer Vision*, pp. 150–165, Springer, 2018.

[23] J. Ding, N. Xue, L. Yang, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2849–2858, Long Beach, CA, USA, 2019.

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

[25] W. Qian, X. Yang, S. Peng, J. Yan, and Y. Guo, "Learning modulated loss for rotated object detection," *In Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, pp. 2458–2466, 2021.

[26] X. Yang, H. Sun, F. Kun et al., "Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sensing*, vol. 10, no. 1, p. 132, 2018.

[27] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *European Conference on Computer Vision*, pp. 677–694, Springer, 2020.

[28] Y. Xue, J. Yang, J. Yan et al., "Scrdet: towards more robust detection for small, cluttered and rotated objects," in *2019 IEEE/CVF International Conference on Computer Vision*, pp. 8232–8241, Seoul, Korea (South)., 2019.

[29] W. Zhuang, X.-G. Tang, G. Yang, G. Yuan, and Y. Haoyuan, "Remote sensing image object detection based on rotatable bounding box," in *2021 International Conference on Digital Society and Intelligent Systems*, pp. 172–177, Chengdu, China, 2021.

[30] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: a rotation-equivariant detector for aerial object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2786–2795, Nashville, TN, USA, 2021.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: transformers for image recognition at scale," 2020, https://arxiv.org/abs/2010.11929.

[32] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: faster and better learning for bounding box regression," *In Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 7, pp. 12993–13000, 2020.

[33] X. Yang and J. Yan, "On the arbitrary-oriented object detection: classification based approaches revisited," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1340–1365, 2022.

[34] J. Ding, N. Xue, G. S. Xia et al., "Object detection in aerial images: a large-scale benchmark and challenges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7778–7796, 2022.

[35] X. Yang, Y. Zhou, and J. Yan, "Alpharotate: a rotation detection benchmark using tensorflow," 2021, https://arxiv.org/abs/2111.06677.

[36] Y. Xue, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15819–15829, Nashville, TN, USA, 2021.

[37] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with gaussian wasserstein distance loss," in *International Conference on Machine Learning*, pp. 11830–11841, PMLR, 2021.

[38] Y. Xue, X. Yang, J. Yang et al., "Learning high-precision bounding box for rotated object detection via kullbackleibler divergence," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18381–18394, 2021.

[39] Y. Jiang, X. Zhu, X. Wang et al., "R2cnn: rotational region cnn for orientation robust scene text detection," 2017, https://arxiv.org/abs/1706.09579.

[40] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision*, pp. 2961–2969, Venice, Italy, 2017.

[41] K. Chen, J. Pang, J. Wang et al., "Hybrid task cascade for instance segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4974–4983, Long Beach, CA, USA, 2019.