IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 February 2024, accepted 14 March 2024, date of publication 19 March 2024, date of current version 25 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3379362

RESEARCH ARTICLE

A2Net: An Anchor-Free Alignment Network for Oriented Object Detection in Remote Sensing Images

QINGYONG YANG[1,2], LIKUN CAO[1], CHENCHEN HUANG[1], QI SONG[1], AND CHUNMIAO YUAN[1]

[1]School of Software, Tiangong University, Tianjin 300387, China
[2]School of Software and Communication, Tianjin Sino-German University of Applied Sciences, Tianjin 300350, China

Corresponding author: Chunmiao Yuan (cm_yuan@tiangong.edu.cn)

ABSTRACT Object detection in remote sensing images is crucial for identifying and locating objects in the field, holding significance in remote sensing. Oriented object detection employs oriented bounding boxes to locate objects with varying orientations, achieving recent advancements. However, challenges persist due to vast variations in object scale and orientation. Existing methods use Intersection over Union (IoU) to measure bounding box quality but often ignore shape information. Unlike horizontal detectors, oriented detectors always incorporate an angle parameter. Yet, objects with different shapes exhibit varying angle sensitivity. For objects with the same angle but different shapes, their IoU can differ significantly. We argue that relying solely on IoU is not comprehensive. To address this, we propose the Shape-aware IoU Score (SaIS), considering shape information and IoU for each bounding box. We use SaIS to enhance the dynamic soft label assignment strategy, resulting in an improved Shape-aware Label Assignment (SaLA). SaLA aids the detector in selecting more suitable samples. Leveraging RTMDet-R and S2ANet strengths, we design an Anchor-free Alignment Network (A2Net) for oriented object detection. A2Net features two detection heads: the initial head and the refinement head. Utilizing alignment convolution (AlignConv) between these heads obtains aligned features. We validate the proposed approach's effectiveness on the DOTA dataset and DIOR-R dataset.

INDEX TERMS Oriented object detection, shape-aware, remote sensing, deep learning.

I. INTRODUCTION

Object detection in remote sensing images is a crucial technique for interpreting such images, finding applications in terrain surveying, intelligence reconnaissance, disaster rescue, and more, thus attracting increasing attention. Objects in remote sensing images often exhibit characteristics of significant scale variations, arbitrary orientations, and dense arrangements, making object detection in these images challenging. Unlike horizontal detectors that regress a four-dimensional vector for object localization, oriented object detectors additionally predict a parameter representing the angle of the bounding box. For objects at arbitrary angles in remote sensing images, horizontal bounding boxes include a considerable amount of background information, while oriented bounding boxes can accurately and effectively represent the object's position. In recent years, owing to the rapid development of horizontal object detectors, many oriented object detectors have been proposed. Some methods [1], [2], [3], [4], [5], [6], [7] focus on extracting better features to alleviate feature misalignment problems, while others [8], [9], [10], [11] concentrate on designing new regression loss functions to address issues arising from angle periodicity and edge swapping, achieving improved detection performance. However, existing methods often use intersection over union (IoU) between the bounding box and the ground truth as the metric for measuring the quality of the bounding box, often overlooking the shape information

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

(e.g., aspect ratio) of the objects. As shown in Fig. 1, objects with different shapes, even with the same angle, can have significantly different IoU values. Given this situation, we redesigned the metric for bounding boxes based on IoU. For each bounding box position, we calculate a shape score based on the shape information of the corresponding ground truth and the IoU value between the bounding box and the ground truth. We then combine the shape score and IoU value to compute the final score, referred to as the Shape-aware IoU Score (SaIS). Using SaIS, we improve the dynamic soft label assignment strategy [12], resulting in an enhanced strategy known as the Shape-aware Label Assignment (SaLA). This strategy assists the detector in selecting more suitable samples. Finally, we propose an anchor-free oriented object detection network—Anchor-free Alignment Network (A2Net)—building upon the advancements of the recent state-of-the-art anchor-free oriented object detection network RTMDet-R [12] and incorporating alignment of object features inspired by anchor-based oriented object detection network S2ANet [5]. A2Net comprises two detection heads: the initial detection head and the refinement detection head. The initial detection head outputs initial bounding boxes for each position. Using the information from these bounding boxes, alignment convolution is applied to the feature map from the feature pyramid network to obtain aligned object features. The refinement detection head then utilizes these improved features to output a set of scale factors refining the initial bounding boxes at each position. The refined bounding boxes serve as the final prediction results. We validate the effectiveness of the proposed approach through extensive experiments on the DOTA [13] and DIOR-R [14] datasets. Our main contributions are summarized as follows:

- We propose a novel method for measuring the quality of bounding boxes, termed Shape-aware IoU Score (SaIS). SaIS combines the shape information of the object and the IoU value.
- Based on SaIS, we propose the Shape-aware Label Assignment (SaLA), representing an improvement over the existing dynamic soft label assignment strategy. SaLA enhances the detector's ability to select more suitable samples.
- We design an anchor-free oriented object detector, A2Net, which employs AlignConv [5] for aligning features and incorporates a refinement detection head to enhance the results from the initial detection head. In comparison with other state-of-the-art methods, our A2Net achieves competitive performance.
- We conduct extensive experiments on the DOTA and DIOR-R datasets to verify the effectiveness of our method.

II. RELATED WORK

In this part, we will initially provide an overview of the iconic object detection algorithms in Section II-A, specifically focusing on horizontal object detectors. Subsequently, in Section II-B, we will introduce the existing oriented object detection algorithms. These methods typically utilize an oriented bounding box with an angle parameter to determine the object's position. Finally, Section II-C outlines various label assignment strategies currently employed in oriented object detectors.

A. HORIZONTAL OBJECT DETECTION

Currently, horizontal object detection has achieved significant advancements, and numerous sophisticated object detectors are widely employed. These detectors are primarily categorized into two types: one-stage object detectors and two-stage object detectors. The two-stage object detector, exemplified by the R-CNN series [16], [17], [18], [19], [20], generates high-quality proposal regions in the initial stage, which are then refined in the subsequent stage. In contrast, single-stage object detectors like SSD [21] and YOLO series [22], [23], [24], [25], [26], prioritize the real-time performance of the network. These detectors directly regress the target's location and predict its class. Typically, the performance of one-stage object detectors is lower than that of two-stage detectors. To mitigate the computational load associated with anchors, anchor-free object detectors [27], [28], [29], [30] have emerged. CornerNet [27] predicts the top-left and bottom-right key points of the target, while CenterNet [28] treats the object as a point, predicting key points through a heatmap. FCOS [29], similar to the original YOLO [22], assigns a category label and regresses bounding box coordinates at each spatial location of the feature map. However, FCOS regresses the distance from the current location to the four edges of the bounding box for object localization. RepPoints [30] employs sets of points to represent objects. Moreover, some works [31], [32], [33] based on the Transformer [34] structure, rather than the traditional CNN structure, eliminate certain additional computations such as anchors and post-processing to achieve a true end-to-end approach.

B. ORIENTED OBJECT DETECTION

Benefiting from the remarkable advancements in horizontal object detection technology, oriented object detection has progressed rapidly. Similar to horizontal detectors, existing oriented object detectors are categorized into one-stage and two-stage detectors. In the realm of two-stage oriented object detection, RoI Transformer [1] utilizes the RRoI learner to supervise the learning of transformation parameters, converting horizontal RoIs to oriented RoIs. This approach avoids the need for numerous anchors. Simultaneously, it utilizes RRoI Align to extract enhanced features, mitigating the issue of feature misalignment. GlidingVertex [6] employs regression of four length ratios to represent relative sliding offsets corresponding to each edge, facilitating the learning of offsets. ReDet [4] introduces a rotation-equivariant network (ReCNN) into the detector to extract rotation-equivariant features. It proposes rotation-invariant RoI Align

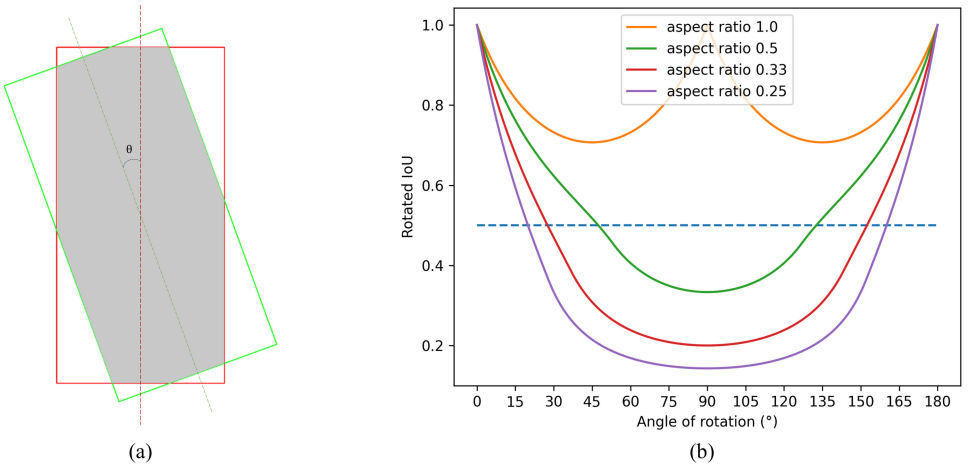


FIGURE 1. Rotated IoU. (a) The red box denotes the prediction box, the green box denotes the ground truth, the gray area denotes the intersection of these two boxes, and θ denotes the rotation angle of the ground truth. (b) The Rotated IoU varies with the rotation angle of the boxes with different aspect ratios. Bounding boxes with smaller aspect ratios are more significantly influenced by the angle. At around a 20-degree rotation, the IoU value of the bounding box with an aspect ratio of 0.25 and the initial horizontal box is already close to 0.5. Typically, 0.5 serves as a threshold to evaluate the bounding box's quality, as represented by a dashed line in the graph. The figure clearly shows that after a 20-degree rotation, bounding boxes with an aspect ratio of 0.5 remain of high quality, while those with an aspect ratio of 0.25 have dropped below the "high-quality boundary." For details of Rotated IoU, please refer to [15].

(RiRoI Align), which adaptively extracts rotation-invariant features from equivariant features based on the orientation of the ROIs. Oriented R-CNN [2] designs a lightweight rotation region proposal network (RPN) that uses the center offset to encode the oriented box, directly generating high-quality oriented region proposals at minimal computational cost. DODet [35] proposes an oriented proposal network (OPN), which generates high-quality oriented proposals via a novel representation scheme of oriented objects, and designed a localization-guided detection head (LDH) that aims at alleviating the feature misalignment between classification and localization. QPDet [36] uses quadrant points in a polar coordinate system to represent bounding boxes, which naturally circumvents the boundary discontinuity problem and enables the production of regular boxes without postprocessing. SFRNet [37] designs two transformer-based branches to perform function-specific feature refinement for fine-grained classification and oriented localization, separately. For single-stage oriented object detectors, BiFA-YOLO [38] proposes a novel bi-directional feature fusion module (Bi-DFFM) to efficiently aggregate features at different resolutions for ship detection. R3Det [3] achieves feature reconstruction and alignment by recording the position information of the current refined bounding box to the corresponding feature points via pixel-wise feature interpolation. S2ANet employs a lightweight anchor refinement network (ARN) to generate high-quality oriented anchors. It then adaptively aligns convolutional features based on the coordinate offsets encoded by these oriented anchors. TCD [39] proposes task collaboration assignment (TCA) and task collaboration header (TCH) to enhance the consistency between classification and localization predictions. Moreover, CP-FCOS [40] designs a category-position (CP) module to optimize the position regression branch features in the FCOS network, which can improve target positioning performance in complex scenes by generating guidance vectors from classification branch features. Mask OBB [41] treats oriented bounding box regression as a pixel-level classification problem, which uses the predicted masks to subsequently generate oriented bounding boxes.

C. LABEL ASSIGNMENT

The label assignment process aims to allocate positive and negative samples for training. In many anchor-based oriented object detectors, such as S2ANet, the MaxIoU matching strategy is commonly employed. This strategy utilizes the IoU value between the anchor box and the ground truth as the matching metric, selecting positive and negative samples based on a predefined IoU threshold. SASM [42] introduces Shape Adaptive Selection (SA-S), which adjusts the IoU threshold according to the shape of the sample. FCOSR [43] enhances the central sampling approach of FCOS by introducing elliptic center sampling. It further addresses the issue of insufficient sampling through a fuzzy sample allocation strategy and a multi-level sampling module. Oriented RepPoints [44] innovates with an adaptive point quality metric and an assignment strategy to allocate high-quality samples. RTMDet-R proposes a dynamic soft label assignment strategy inspired by SimOTA [45]. This strategy utilizes softened classification cost, regression cost, and region cost as matching metrics.

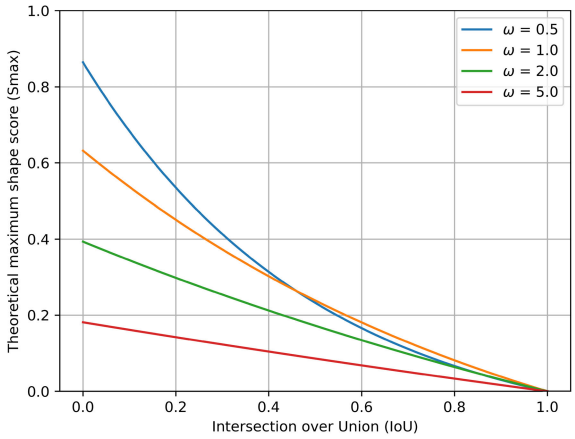


FIGURE 2. The theoretical maximum shape scores for different values of ω. With the increase of ω, S_{max} gradually decreases, and the influence of different IoU values on the upper limit of the theoretical maximum shape score also tends to be "flat."

III. METHOD

This section provides a detailed description of our proposed method. In Section III-A, we define and present the calculation formula for SaIS. The improved shape-aware label assignment strategy is discussed in Section III-B. Section III-C outlines the network structure of A2Net. Finally, an overview of the loss function utilized by the network is presented in Section III-D.

A. SHAPE-AWARE IoU SCORE

The classic IoU-based quality assessment strategy generally performs well for most objects but overlooks those with distinct shapes. As depicted in Fig. 1, for objects with smaller aspect ratios, the IoU values between their bounding boxes and ground truth boxes are more influenced by angles. Relying solely on IoU values is insufficient for a comprehensive evaluation of a bounding box. Consequently, we introduce SaIS, consisting of two components: the shape score and the IoU value. The shape score of a bounding box is computed based on its IoU value with the ground truth and the shape information of the ground truth. The formula for the shape score S is as follows:

$$S_{i,j} = Smax_{i,j} \times (1 - \gamma_j)^2 \tag{1}$$

where $S_{i,j}$ represents the shape score of the i-th bounding box relative to the j-th ground truth. γ_j denotes the aspect ratio of the j-th ground truth, with the longer side defined as the height, thus $\gamma_j \in (0, 1]$. $Smax_{i,j}$ represents the theoretical maximum shape score of the i-th bounding box relative to the j-th ground truth. $Smax_{i,j}$ is defined as follows:

$$Smax_{i,j} = \begin{cases} e^{-\frac{I_{i,j}}{\omega}} - e^{-\frac{1}{\omega}} & \text{if } I_{i,j} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

$I_{i,j}$ represents the IoU value between the i-th bounding box and the j-th ground truth. Since $I_{i,j} \in [0, 1]$, we use $-e^{-\frac{1}{\omega}}$ to constrain the theoretical maximum shape score to be non-negative. ω is a hyperparameter that controls the upper limit of the theoretical maximum shape score and the rate at which this limit increases as IoU decreases. We set the default value of ω to 2. As shown in Fig. 2, for bounding boxes with relatively low IoU values, we assign a higher theoretical maximum shape score, assuming that the low IoU values result from a large aspect ratio and arbitrary orientation. This is also the reason why we refer to it as the theoretical maximum shape score. However, boxes with an IoU of 0 are not considered.

Based on (1) and (2), we get the calculation method of SaIS, which actually simply adds the shape score and the IoU value. We chose not to introduce additional parameters to assign weights to $S_{i,j}$ and $I_{i,j}$, as we have already used hyperparameters in (2) to achieve satisfactory results. The calculation of SaIS is as follows:

$$SaIS_{i,j} = S_{i,j} + I_{i,j} \tag{3}$$

where $SaIS_{i,j}$ represents the SaIS of the i-th bounding box relative to the j-th ground truth, while $S_{i,j}$ and $I_{i,j}$ represent, respectively, the shape score and IoU between the i-th bounding box and the j-th ground truth. We can see that SaIS is jointly influenced by shape information and IoU value. The theoretical maximum shape score and aspect ratio determine the shape score, while the IoU value determines the theoretical maximum shape score.

B. SHAPE-AWARE LABEL ASSIGNMENT

We improve the dynamic soft label assignment strategy, as introduced in Section III-A, based on SaIS. The enhanced label assignment strategy is referred to as the Shape-aware Label Assignment. For convenience, we abbreviate the dynamic soft label assignment strategy as DSLA, while our shape-aware label assignment strategy is abbreviated as SaLA. Similar to DSLA, SaLA utilizes a cost function to compute a cost matrix as the matching criterion. The cost function C is defined as follows:

$$C = \lambda_1 C_{cls} + \lambda_2 C_{reg} + \lambda_3 C_{center} \tag{4}$$

where C_{cls}, C_{reg}, and C_{center} correspond to the classification cost, regression cost, and region cost, respectively. The weights for these three costs are denoted as λ_1, λ_2, and λ_3. By default, following RTMDet-R, we set $\lambda_1 = 1$, $\lambda_2 = 3$, and $\lambda_3 = 1$.

Inspired by GFL [46], DSLA utilizes IoU as a soft label to reweight the classification cost for different regression qualities, avoiding noise and unstable matching caused by binary labels. In SaLA, we employ our SaIS instead of IoU as the soft label. The calculation formula for the classification cost C_{cls} is as follows:

$$C_{cls} = CE(P, SaIS) \times (SaIS - P)^2 \tag{5}$$

where CE denotes cross-entropy loss, P represents the estimated probability.

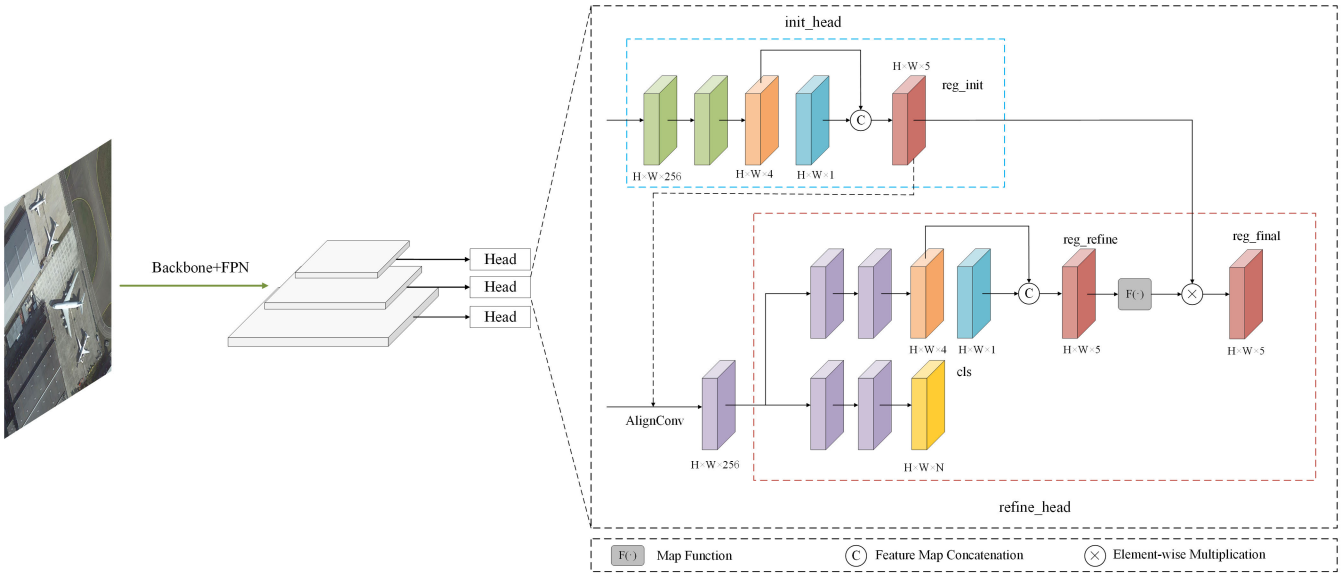


FIGURE 3. A2Net network structure. It is composed of backbone network, feature pyramid network and detection head corresponding to feature map of different layers. By default, we use CSPNeXt series as the backbone network and CSPNeXt-PAFPN as the feature pyramid network. Each detection head comprises an initial detection head and a refinement detection head. AlignConv uses the initial bounding box information output by the initial detection head to apply deformable convolution to the feature map from the feature pyramid network, obtaining axis-aligned features. The refinement detection head utilizes these aligned features to predict category and a set of scaling coefficients at each position on the feature map. After mapping the scaling coefficients through a mapping function, they are multiplied by the initial bounding box to obtain the refined bounding box as the final prediction. We have omitted the classification branch in the initial detection head as it is only used during the training phase. During testing, we only require the results from the regression branch. The detection heads across feature pyramid layers share weights.

Similarly, we apply SaIS for calculating the regression cost. We take the logarithm of SaIS to amplify the maximum difference between the best and worst matches, better distinguishing high-quality matches from low-quality matches. The calculation formula for the regression cost C_{reg} is as follows:

$$C_{reg} = -\log(SaIS). \tag{6}$$

The region cost C_{center}, like DSLA, utilizes central priors to assign lower matching costs to bounding boxes closer to the center of the ground truth. The calculation formula is as follows:

$$C_{center} = \alpha^{|x_{pred} - x_{gt}| - \beta} \tag{7}$$

where x_{pred} and x_{gt} denote the center coordinates of the predicted bounding box and ground truth, respectively. α and β are hyperparameters, and by default, we set $\alpha = 10$ and $\beta = 3$.

Based on the cost C, we assign the samples through a dynamic top-k strategy at different iterations. Please refer to Dynamic k Estimation strategy in OTA [47] for more details. By default, we set $k = 13$.

C. A2Net

We have developed an anchor-free alignment network, A2Net, for oriented object detection in remote sensing images, building upon the anchor-free oriented object detection network RTMDet-R and drawing inspiration from the alignment of features in the anchor-based oriented object detection network S2ANet. The network structure is illustrated in Fig. 3. A2Net's head consists of two parts: the initial detection head and the refinement detection head. Each detection head comprises a classification branch and a regression branch. The regression branch of the initial detection head outputs a five-dimensional vector (l, t, r, b, θ) at each spatial position, representing the initial oriented bounding box. AlignConv uses these initial oriented bounding boxes to apply deformable convolution to align features from the feature pyramid network. The classification branch of the refinement detection head outputs a vector of N elements at each spatial position (where N is the total number of classes), while the regression branch at each position outputs a set of scaling coefficients $(\Delta l, \Delta t, \Delta r, \Delta b, \Delta\theta)$. The final oriented bounding box $(l', t', r', b', \theta')$ is calculated using the following formula:

$$(l', t', r', b', \theta') = (l, t, r, b, \theta) \times \Delta s \tag{8}$$

where Δs represents the scaling factor output by the refinement detection head, and we adjust the initial box within the scaling range of $[0.5, 1.5]$. The formula for calculating Δs is as follows:

$$\Delta s = 1 + \frac{2 \times sigmoid(\Delta\cdot) - 1}{2} \tag{9}$$

here, $\Delta\cdot$ represents $(\Delta l, \Delta t, \Delta r, \Delta b, \Delta\theta)$, and $sigmoid$ is a linear mapping function that maps its input to the range $[0, 1]$.

TABLE 1. Ablation experiments of SaLA. The base detector is A2Net, using CSPNeXt-s as the backbone network. λ_1, λ_2, and λ_3 denote the weight of the classification cost, the regression cost, and the region cost, respectively.

λ_1	1	0	0	1	1	0	1	1	1	1
λ_2	0	1	0	1	0	1	1	2	3	4
λ_3	0	0	1	0	1	1	1	1	1	1
mAP (%)	-	66.19	67.71	65.82	57.55	69.28	68.80	70.05	69.55	69.27

D. LOSS FUNCTION

Each detection head in A2Net assigns a category label at each position in its feature map and regresses the object's position. The total loss is composed of two parts: the loss from the initial detection head and the loss from the refinement detection head. The total loss function is defined as follows:

$$L_{total} = L_{init} + L_{refine}. \tag{10}$$

The loss for each detection head includes both classification loss and regression loss. As their calculation formulas are the same, we use $L\cdot$ to represent the loss for any detection head, specifically defined as follows:

$$L\cdot = L_{cls} + \lambda L_{reg} \tag{11}$$

where the classification loss L_{cls} is calculated using Quality Focal Loss (QFL) [46], and the regression loss L_{reg} is computed using Rotated IoU Loss [15]. The parameter λ is a hyperparameter used to balance the classification and regression losses, and we set $\lambda = 2$ by default.

IV. EXPERIMENT

A. DATASET

DOTA is a challenging dataset for large-scale aerial image object detection. It consists of 2,806 aerial images of 188,282 objects of different scales, orientations, and shapes, each of which ranges in size from 800 × 800 to 20,000 × 20,000 pixels. The dataset is divided into 15 categories: plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). The proportions of the training set, validation set, and testing set in DOTA are 1/2, 1/6, and 1/3, respectively.

DIOR-R is a large-scale benchmark dataset for object detection in optical remote sensing images, which consists of 23,463 images and 192,518 object instances annotated with oriented bounding boxes. The dataset is an extended version of DIOR [48] annotated with oriented bounding boxes, which shares the same images with DIOR. It is categorized into 20 classes, including bridge (BR), ship (SH), airplane (APL), stadium (STA), tennis court (TC), chimney(CH), overpass (OP), vehicle (VE), baseball field (BF), airport (APO), basketball court (BC), ground track field (GTF), expressway service area (ESA), train station (TS), golf field (GF), dam (DAM), expressway toll station (ETS), storage tank (STO), harbor (HA), and windmill (WM).

B. IMPLEMENTATION DETAILS

For the DOTA dataset, the training and validation sets are used for training our model, and the testing set is used for evaluation. We crop the original images into a set of 1024 × 1024-sized images with a 200-pixel overlap. By default, we use CSPNeXt series [12] as the backbone and CSPNeXt-PAFPN [12] as the neck. We employ the stochastic gradient descent (SGD) optimizer for training, with weight decay and momentum set to 0.0001 and 0.9, respectively. All models are trained for 12 epochs with an initial learning rate of 0.01, reducing the learning rate by a factor of 10 at the 8th and 11th epochs. During training, we randomly apply horizontal, vertical, or diagonal flips to the images with a probability of 0.75. Our approach was implemented based on mmrotate [49], and all experiments were performed on a single NVIDIA A100 PCIe with a batch size of 8 per training iteration. For the DIOR-R dataset, we simply randomly flipped the images horizontally, vertically, or diagonally with a probability of 0.75, otherwise remaining consistent with DOTA.

C. ABLATION STUDIES

1) SHAPE-AWARE LABEL ASSIGNMENT

The impact of various weights in the cost function on the experimental results was investigated. Table 1 displays the outcomes, indicating that the optimal performance was achieved when $\lambda_1 = 1$, $\lambda_2 = 2$, and $\lambda_3 = 1$. Since the default settings are close to optimal performance, we are consistent with RTMDet-R and set $\lambda_1 = 1$, $\lambda_2 = 3$, and $\lambda_3 = 1$. We used RTMDet-R and our A2Net as the benchmark network to verify the effectiveness of SaLA. As shown in Table 2, the performance of RTMDet-R and A2Net using DSLA on the testing set of DOTA is 70.54% mAP and 72.17% mAP, respectively, and after using SaLA as the label assigner, they reached 72.73% mAP and 73.75% mAP, respectively, an increase of 2.19% mAP and 1.58% mAP. This verifies the effectiveness of SaLA.

2) A2Net

We studied the impact of individual components of A2Net, and the results are shown in Table 3. The first

TABLE 2. Comparison of SaLA and DSLA. The base detectors are RTMDet-R and our A2Net, both using CSPNeXt-l as the backbone network and CSPNeXt-PAFPN as the feature pyramid network. DSLA represents the dynamic soft label assignment strategy, while SaLA represents the shape-aware label assignment strategy as described in Section III-B.

Model	Label Assigner	mAP (%)
RTMDet-R	DSLA	70.54
RTMDet-R	**SaLA**	**72.73**
A2Net	DSLA	72.17
A2Net	**SaLA**	**73.75**

TABLE 3. Ablation experiments of A2Net. ✓ indicates that the corresponding module is used. Use CSPNeXt-l as the backbone network and use CSPNeXt-PAFPN as the feature pyramid network.

SaLA	Refinement Head	mAP(%)
		70.54
✓		72.73
	✓	72.17
✓	✓	**73.75**

TABLE 4. More comparisons of SaLA and DSLA. All these methods utilize CSPNeXt-PAFPN as the feature pyramid network.

Model	Bakebone	Label Assigner	DOTA mAP (%)	DIOR-R mAP (%)
RTMDet-R	CSPNeXt-tiny	DSLA	64.62	51.82
RTMDet-R	CSPNeXt-tiny	**SaLA**	**66.93 (+2.31)**	**53.82 (+2.00)**
RTMDet-R	CSPNeXt-s	DSLA	68.27	54.23
RTMDet-R	CSPNeXt-s	**SaLA**	**68.95 (+0.68)**	**56.89 (+2.66)**
RTMDet-R	CSPNeXt-m	DSLA	68.81	58.68
RTMDet-R	CSPNeXt-m	**SaLA**	**71.01 (+2.20)**	**59.81 (+1.13)**
RTMDet-R	CSPNeXt-l	DSLA	70.54	60.90
RTMDet-R	CSPNeXt-l	**SaLA**	**72.73 (+2.19)**	**62.64 (+1.74)**
A2Net	CSPNeXt-tiny	DSLA	65.94	55.08
A2Net	CSPNeXt-tiny	**SaLA**	**67.14 (+1.20)**	**56.42 (+1.34)**
A2Net	CSPNeXt-s	DSLA	68.20	58.50
A2Net	CSPNeXt-s	**SaLA**	**69.55 (+1.35)**	**58.21 (-0.29)**
A2Net	CSPNeXt-m	DSLA	71.88	60.90
A2Net	CSPNeXt-m	**SaLA**	**72.68 (+0.80)**	**61.83 (+0.93)**
A2Net	CSPNeXt-l	DSLA	72.17	62.63
A2Net	CSPNeXt-l	**SaLA**	**73.75 (+1.58)**	**63.77 (+1.14)**

row shows the performance of our baseline method RTMDet-R, which has the same network structure as A2Net without the refinement detection head and achieves 70.54% mAP. After replacing DSLA with our SaLA, there is an improvement of 2.19% mAP, reaching 72.73% mAP. Adding the refinement detection head increases the performance further improves to 73.75% mAP. With these components, our A2Net achieves a 3.21% mAP performance improvement at a small cost compared to RTMDet-R.

TABLE 5. Parameters versus accuracy on the testing set of DOTA. RTMDet-R uses the default DSLA for label assignment, while A2Net uses our SaLA. All models use CSPNeXt-PAFPN as the feature pyramid network.

Model	Bakebone	Params	GFLOPs	mAP (%)
RTMDet-R	CSPNeXt-tiny	4.87 M	20.45	64.62
A2Net	CSPNeXt-tiny	5.29 M	27.63	**67.14 (+2.52)**
RTMDet-R	CSPNeXt-s	8.85 M	37.62	68.27
A2Net	CSPNeXt-s	9.60 M	50.36	**69.55 (+1.28)**
RTMDet-R	CSPNeXt-m	24.66 M	99.76	68.81
A2Net	CSPNeXt-m	26.33 M	128.39	**72.68 (+3.87)**
RTMDet-R	CSPNeXt-l	52.25 M	204.21	70.54
A2Net	CSPNeXt-l	55.21 M	255.06	**73.75 (+3.21)**

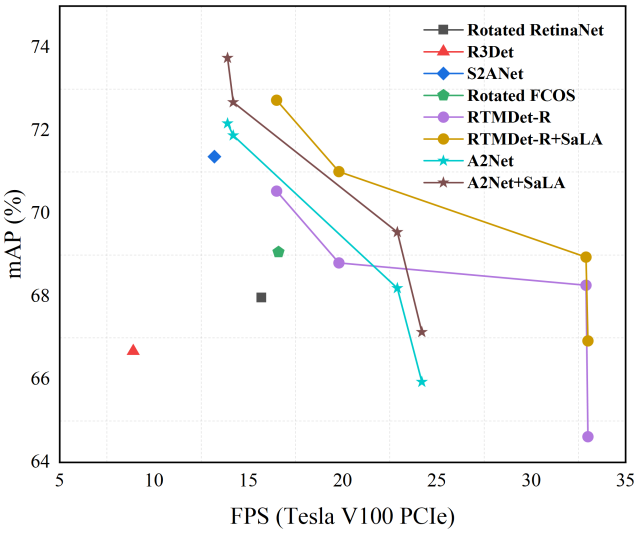


FIGURE 4. Speed versus accuracy on the testing set of DOTA. All experiments were performed on a single Tesla V100 PCIe.

3) MORE BACKBONES AND DATASETS

We conducted additional experiments using different backbone networks on the DOTA and DIOR-R datasets. The results, shown in Table 4, indicate that all models, except for A2Net, which uses CSPNeXt-s as the backbone and drops by 0.29% mAP on the DIOR-R dataset, have more or less performance improvement on both datasets, with the maximum improvement obtained on the DOTA dataset being 2.31% mAP and on the DIOR-R dataset being 2.66% mAP.

4) PARAMETERS VERSUS ACCURACY

We analyze the complexity of our algorithm in Table 5. Compared to RTMDet-R, the extra computation comes from the refinement head we added, and SaLA is cost-free. The refinement head is lightweight, and it is worthwhile that we obtain a performance improvement of at least 1.28% mAP at a small cost, with a maximum improvement of 3.87% mAP.

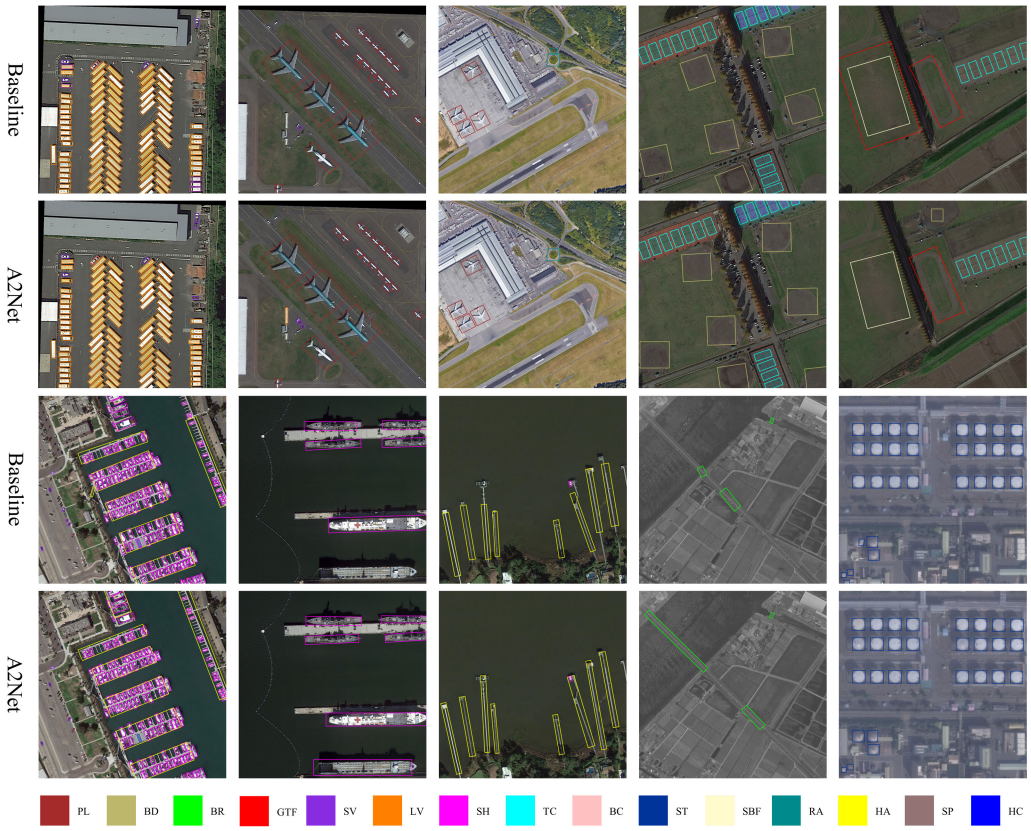


FIGURE 5. Some comparison results between A2Net and the baseline method on the testing of DOTA. The baseline model is RTMDet-R, and the confidence threshold for visualization results is set to 0.3.

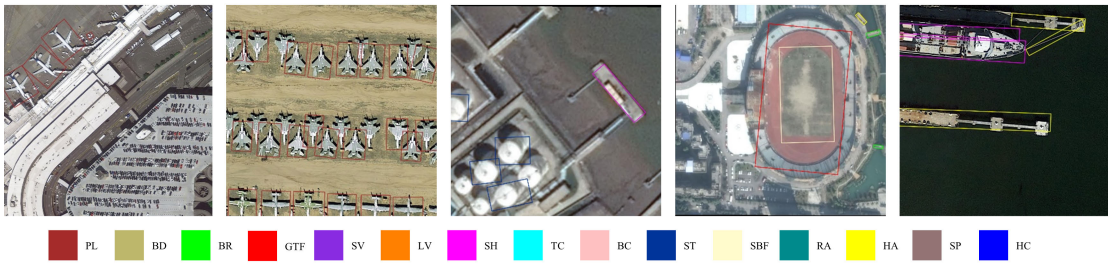


FIGURE 6. Some typical failure predictions of our method on the testing set of DOTA.

5) SPEED VERSUS ACCURACY

Fig. 4 illustrates the trade-off between speed and accuracy. All methods use ResNet-50 [50] and FPN [51] except RTMDet-R and our A2Net which uses CSPNeXt and CSPNeXt-PAFPN. The hardware platform for testing is a single Tesla V100 PCIe. The batch size for testing is 1 and the input size is 1024 × 1024. We iterated 2000 times to get more accurate results. From Fig. 4, we can find that our SaLA enables RTMDet-R to outperform the single-stage detectors in both speed and accuracy, with accuracy up to 72.73% mAP and speed up to 33 FPS. A2Net is a compromise that guarantees faster speed than the single-stage detectors while obtaining higher accuracy than RTMDet-R. The method has an accuracy of up to 73.75% mAP and a speed of up to 24.2 FPS.

D. COMPARISONS WITH STATE-OF-THE-ART

We conducted a performance comparison of our proposed A2Net with other state-of-the-art methods on DOTA. To ensure a fair comparison, we retrained all methods using the same training strategy on our machine. All experiments were conducted based on mmrotate. As shown in Table 6, our approach achieved a mAP of 73.75%, exhibiting a 3.21% mAP improvement compared to the baseline method (RTMDet-R). Simultaneously, A2Net demonstrated competitive performance compared to other anchor-based methods. We visualize some high-quality results with

TABLE 6. Comparisons with state-of-the-art methods on the testing set of DOTA. R50 denotes ResNet-50. ReR50, similar to R50, employs a rotation-equivariant convolutional network as utilized in ReDet. By default, RTMDet-R and A2Net use CSPNeXt-PAFPN, while other methods adopt FPN. All methods use single-scale training and testing. ∗ indicates that the results are from the original paper. † represents our SaLA as the label assignment strategy.

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP (%)
Anchor-base (two-stage):																	
Rotated Faster-RCNN [18]	R50	88.49	79.10	51.27	68.90	78.46	74.62	86.37	90.68	79.96	84.78	55.11	65.34	66.82	69.65	51.63	72.75
RoI Transformer	R50	88.59	83.32	55.15	69.57	78.93	83.20	88.08	90.82	86.60	85.54	63.12	60.98	76.76	72.13	50.32	75.54
ReDet	ReR50	88.22	83.91	54.41	73.26	79.21	83.55	88.36	90.65	87.32	85.90	64.03	60.77	76.69	71.05	62.71	76.67
AOPG* [14]	R50	89.27	83.49	52.50	69.97	73.51	82.31	87.95	90.89	87.64	84.71	60.01	66.12	74.19	68.30	57.80	75.24
DODet*	R50	89.34	84.31	51.39	71.04	79.04	82.86	88.15	90.90	86.88	84.91	62.69	67.63	75.47	72.22	45.54	75.49
Oriented R-CNN	R50	88.80	80.83	54.28	72.96	78.79	82.57	87.94	90.74	85.85	85.57	64.44	62.78	73.25	68.80	50.87	75.23
QPDet*	R50	89.55	83.66	54.06	73.93	78.93	83.08	88.29	90.89	86.60	84.80	62.03	65.55	74.16	70.09	58.16	76.25
Anchor-base (one-stage):																	
Rotated RetinaNet [52]	R50	87.54	78.97	33.98	62.87	78.01	60.53	76.85	90.88	82.26	81.61	58.00	61.10	54.09	64.81	48.08	67.97
R3Det	R50	89.04	73.00	34.82	58.13	77.84	73.80	83.62	90.88	73.51	83.23	51.99	60.25	57.41	57.16	35.55	66.68
S2ANet	R50	88.26	74.14	47.52	67.46	79.24	79.25	87.56	90.90	81.21	85.18	54.42	62.08	66.34	65.71	41.25	71.37
TCD*	R50	70.89	65.75	56.91	89.27	70.67	76.54	76.76	60.16	83.79	71.95	90.88	72.13	71.21	86.95	83.81	75.18
Anchor-free:																	
Rotated FCOS [29]	R50	88.44	67.64	43.35	59.11	79.98	78.02	87.32	90.88	77.65	83.41	51.07	59.65	63.64	64.63	41.23	69.07
RTMDet-R	CSPNeXt-tiny	88.20	61.70	33.98	53.91	77.74	74.94	85.75	90.86	76.17	84.34	37.13	53.97	53.95	61.40	35.26	64.62
RTMDet-R	CSPNeXt-s	88.76	69.80	37.16	57.63	78.73	78.09	87.35	90.85	81.48	85.03	46.50	55.39	60.83	63.98	42.54	68.27
RTMDet-R	CSPNeXt-m	89.43	60.88	39.87	60.76	79.58	78.79	88.28	90.90	79.58	84.70	50.81	57.54	64.09	67.25	39.69	68.81
RTMDet-R	CSPNeXt-l	89.17	73.38	41.30	62.29	79.54	80.63	88.25	90.89	80.68	86.39	49.57	57.44	65.90	63.61	49.04	70.54
A2Net†	CSPNeXt-tiny	88.26	66.75	36.43	58.39	78.23	77.66	86.91	90.87	78.31	84.28	45.38	52.53	64.45	61.63	37.05	67.14
A2Net†	CSPNeXt-s	88.46	68.82	39.13	62.01	78.58	79.61	87.47	90.87	77.85	86.39	48.79	56.69	65.69	67.02	45.88	69.55
A2Net†	CSPNeXt-m	89.40	76.07	41.41	69.75	79.93	81.21	88.04	90.91	84.32	85.00	56.57	57.41	68.55	69.69	51.92	72.68
A2Net†	CSPNeXt-l	89.46	74.79	44.96	68.18	79.84	81.16	88.39	90.89	82.98	84.84	61.98	60.04	70.18	69.52	59.07	73.75

better localization of our method compared to the baseline in Fig. 5.

V. CONCLUSION

In this paper, we propose the Shape-aware IoU Score (SaIS), which integrates IoU and object shape information to alleviate differences caused by the varying sensitivity to angles in objects with different shapes. Building upon SaIS, we enhance the dynamic soft label assignment strategy and introduce the Shape-aware Label Assignment strategy (SaLA), which incorporates classification cost, regression cost, and region priors to allocate more appropriate samples for the training process. Additionally, we devise an anchor-free alignment network, A2Net, for object detection in remote sensing images. A2Net is an improvement over RTMDet-R, utilizing AlignConv and introducing a refined detection head to address the feature misalignment issue arising from the arbitrary orientation of objects. We conducted extensive experiments on large-scale aerial image datasets, DOTA and DIOR-R, to validate the effectiveness of our approach. A2Net with SaLA achieved competitive performance in the DOTA OBB task. Furthermore, compared to anchor-free methods, our approach demonstrated state-of-the-art performance.

Future work: As shown in Fig. 6, our method fails for some small and dense objects. It also performs poorly for some ring-shaped objects (e.g., storage tanks), and for long and narrow objects, it also appears to detect parts of them as objects. The inability to accurately predict large objects, such as ground track fields, also occurs. Therefore, in our future work, we plan to explore the following aspects:

- We will learn from and enhance ERF-RTMDet [53], which focuses on high-precision detection of small objects, to improve accuracy of small objects in aerial images.
- Localization based on rotated bounding boxes often lacks precision due to objects having arbitrary orientations and dense arrangements. Some works [41], [54], [55] have demonstrated the effectiveness of generating bounding boxes by predicting masks. We will conduct further research in this direction.
- Objects in aerial images exhibit significant scale variations. In the future, we will focus on high-performance multi-scale rotated object detection.

REFERENCES

[1] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2844–2853.

[2] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3500–3509.

[3] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 3163–3171.

[4] J. Han, J. Ding, N. Xue, and G.-S. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2785–2794.

[5] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3062048.

[6] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[7] Z. Guo, C. Liu, X. Zhang, J. Jiao, X. Ji, and Q. Ye, "Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8788–8797.

[8] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Computer Vision—ECCV 2020*. Cham, Switzerland: Springer, 2020, pp. 677–694.

[9] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian Wasserstein distance loss," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11830–11841.

[10] X. Yang, X. Yang, J. Yang, Q. Ming, W. Wang, Q. Tian, and J. Yan, "Learning high-precision bounding box for rotated object detection via kullback–leibler divergence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 18381–18394.

[11] X. Yang, Y. Zhou, G. Zhang, J. Yang, W. Wang, J. Yan, X. Zhang, and Q. Tian, "The KFIoU loss for rotated object detection," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022, pp. 1–17.

[12] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "RTMDet: An empirical study of designing real-time object detectors," 2022, *arXiv:2212.07784*.

[13] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[14] G. Cheng, J. Wang, K. Li, X. Xie, C. Lang, Y. Yao, and J. Han, "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3183022.

[15] D. Zhou, J. Fang, X. Song, C. Guan, J. Yin, Y. Dai, and R. Yang, "IoU loss for 2D/3D object detection," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 85–94.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

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

[18] 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. 28, 2015, pp. 1–9.

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

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision—ECCV*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 21–37.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[23] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[24] C.-Y. Wang, A. Bochkovskiy, and H. M. Liao, "Scaled-YOLOV4: Scaling cross stage partial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13024–13033.

[25] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOV7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[26] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13034–13043.

[27] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[28] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Computer Vision—ECCV 2020*. Cham, Switzerland: Springer, Aug. 2020, pp. 474–490.

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

[30] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9656–9665.

[31] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision—ECCV 2020*. Cham, Switzerland: Springer, 2020, pp. 213–229.

[32] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–16.

[33] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–23.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[35] G. Cheng, Y. Yao, S. Li, K. Li, X. Xie, J. Wang, X. Yao, and J. Han, "Dual-aligned oriented detector," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618111.

[36] Y. Yao, G. Cheng, G. Wang, S. Li, P. Zhou, X. Xie, and J. Han, "On improving bounding box representations for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3231340.

[37] G. Cheng, Q. Li, G. Wang, X. Xie, L. Min, and J. Han, "SFRNet: Fine-grained oriented object recognition via separate feature refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3277626.

[38] Z. Sun, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 21, p. 4209, Oct. 2021.

[39] C. Zhang, B. Xiong, X. Li, and G. Kuang, "TCD: Task-collaborated detector for oriented objects in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3244953.

[40] Z. Sun, M. Dai, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "An anchor-free detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 14, pp. 7799–7816, 2021.

[41] J. Wang, J. Ding, H. Guo, W. Cheng, T. Pan, and W. Yang, "Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images," *Remote Sens.*, vol. 11, no. 24, p. 2930, Dec. 2019.

[42] L. Hou, K. Lu, J. Xue, and Y. Li, "Shape-adaptive selection and measurement for oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, 2022, pp. 923–932.

[43] Z. Li, B. Hou, Z. Wu, B. Ren, and C. Yang, "FCOSR: A simple anchor-free rotated detector for aerial object detection," *Remote Sens.*, vol. 15, no. 23, p. 5499, Nov. 2023.

[44] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented RepPoints for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1819–1828.

[45] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[46] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21002–21012.

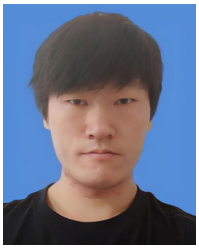
[47] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, "OTA: Optimal transport assignment for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 303–312.

[48] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[49] Y. Zhou, X. Yang, G. Zhang, J. Wang, Y. Liu, L. Hou, X. Jiang, X. Liu, J. Yan, C. Lyu, W. Zhang, and K. Chen, "MMRotate: A rotated object detection benchmark using PyTorch," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 7331–7334.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[51] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[52] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[53] S. Liu, H. Zou, Y. Huang, X. Cao, S. He, M. Li, and Y. Zhang, "ERF-RTMDet: An improved small object detection method in remote sensing images," *Remote Sens.*, vol. 15, no. 23, p. 5575, Nov. 2023.
[54] V. Schmidt and M. Kada, "Object detection of aerial image using mask-region convolutional neural network (mask R-CNN)," in *IOP Conference Series: Earth and Environmental Science*, vol. 500. Bristol, U.K.: IOP Publishing, 2020, p. 012090.
[55] K. A. Hashmi, A. Pagani, D. Stricker, and M. Z. Afzal, "BoxMask: Revisiting bounding box supervision for video object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2029–2039.

QINGYONG YANG was born in 1974. He received the B.S. and M.S. degrees in computer software and computer science and technology from Xidian University, Xi'an, China, in 1997 and 2004, respectively. He is currently a Professor with the School of Software and Communication, Tianjin Sino-German University of Applied Sciences, Tianjin, China. He is also a Master's Supervisor with Tiangong University. His research interests include artificial intelligence and computer vision.



LIKUN CAO was born in 1998. He received the bachelor's degree in software engineering from Xi'an University of Posts and Telecommunications, in 2021. He is currently pursuing the master's degree in software engineering with Tiangong University. His research interests include computer vision, deep learning, and oriented object detection.

CHENCHEN HUANG was born in 1996. He received the bachelor's degree in software engineering from the School of Computer Science and Technology, Tiangong University, in 2021. He is currently pursuing the master's degree with Tiangong University. His current research interests include deep learning and small object detection.

QI SONG was born in 1999. She received the bachelor's degree in software engineering from Taiyuan University of Science and Technology, in 2021. She is currently pursuing the master's degree in software engineering with Tiangong University. Her research interests include image recognition and object detection.

CHUNMIAO YUAN was born in 1975. She received the B.S. and M.S. degrees in computer software and computer science and technology from Xidian University, Xi'an, China, in 1998 and 2003, respectively, and the Ph.D. degree in computer application technology from Tianjin University, Tianjin, China, in 2013. She is currently an Associate Professor with the School of Software, Tiangong University. Her research interests include artificial intelligence and computer vision.