Affinity-Aware Relation Network for Oriented Object Detection in Aerial Images

Tingting Fang[1,2], Bin Liu[1,2*], Zhiwei Zhao[1,2], Qi Chu[1,2], and Nenghai Yu[1,2]

[1] School of Cyber Science and Technology,
University of Science and Technology of China
[2] Key Laboratory of Electromagnetic Space Information,
Chinese Academy of Science
fountain@mail.ustc.edu.cn, flowice@ustc.edu.cn,
zwzhao98@mail.ustc.edu.cn, {qchu, ynh}@ustc.edu.cn

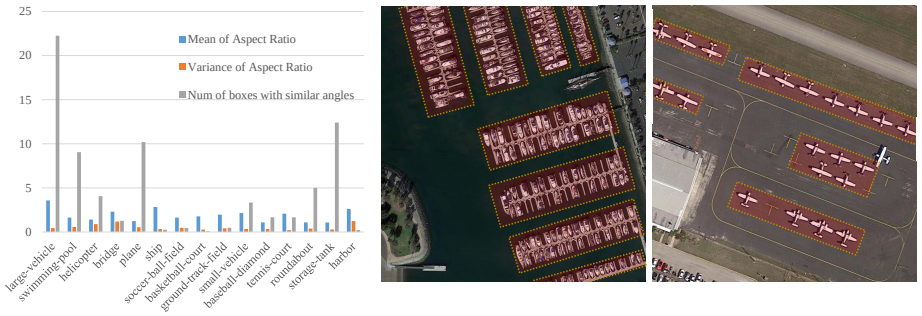
Abstract. Object detection in aerial images is a challenging task due to the oriented and densely packed objects. However, densely packed objects constitute a significant characteristic of aerial images: objects are not randomly scattered around in images but in groups sharing similar orientations. Such a recurring pattern of object arrangement could enhance the rotated features and improve the detection performance. This paper proposes a novel and flexible Affinity-Aware Relation Network based on two-stage detectors. Specifically, an affinity-graph construction module is adopted to measure the affinity among objects and to select bounding boxes sharing high similarity with the reference box. Furthermore, we design a dynamic enhancement module, which uses the attention to learn neighbourhood message and dynamically determines weights for feature enhancement. Finally, we conduct experiments on several public benchmarks and achieve notable AP improvements as well as state-of-the-art performances on DOTA, HRSC2016 and UCAS-AOD datasets.

1 Introduction

Oriented object detection of aerial images is a significant yet challenging task in computer vision. Unlike object detection in ordinary scenes, aerial images with high resolution often contain a larger number of densely packed objects. In this case, detection performance of horizontal object detection models [3, 26, 29, 54, 2] deteriorates considerably due to the intersection of axis-aligned receptive fields between objects. Existing methods mainly contribute to solving this challenge from two aspects: One is to optimize the extraction of rotated features [5, 10, 42, 45, 43], such as using rotation-equivariant backbones or enhancing the feature fusion. The other is to perform well-designed bounding box representations [38, 52, 7, 44, 46], such as using eight-parameter or convex-hull to represent boxes.

However, these densely packed objects form a pattern of object arrangements. For each object in each category, we count the average number of its similar

* Corresponding author.



(a) Box distribution (b) Object arrangement with regularity

Fig. 1. (a) For each object in each category, we count the average number (grey pillar) of its similar objects with an angle difference less than 5 degrees inside a 1024 × 1024 image. (b) Objects inside each red box share high similarity in categories and orientations.

(a) Test image (b) ReDet (c) Our method

Fig. 2. (a) The edge of a tennis court is obscured by shadows. (b) and (c) are visualizations of ReDet [10] and our method on DOTA. Here red and green boxes represent predictions and ground truth. ReDet does not perform well on the obscured boundary, while our Affinity-Aware Relation Network perceives correct boundary information of the tennis court.

objects with an angle difference less than 5 degrees inside a 1024 × 1024 image. The result in Fig. 1(a) shows that each object can find 3-5 objects with similar categories and orientations on average, and some categories can even find more than 20 objects. Fig. 1(b) displays the recurring pattern of object arrangement. Each object can be allocated to an imaginary red box, such that objects inside the same box share high similarity in categories and orientations. Therefore, the semantic information of one object can imply information of other objects in the same box, which can be utilized as an enhancement for the detection task.

This paper proposes an Affinity-Aware Relation Network(AARN) based on the two-stage detector, which aims to enhance the Rotated Region of Interest(RRoI) Align feature for classification and regression in the second stage. Specifically, the proposed AARN consists of two modules. One is a graph construction module, which measures the affinity among objects and dynamically se-

lects bounding boxes sharing high similarity with the reference box. The other is a dynamic enhancement module, which use the attention module to learn neighbourhood message and dynamically determines weights for feature enhancement.

The effectiveness of the method can be simply illustrated by Fig. 2. The input image, detection results on ReDet [10] and results on our method are respectively shown in Fig. 2(a), (b) and (c). The edge of the tennis court on the right in Fig. 2(a) is obscured by the shadows. Fig. 2(b) shows that ReDet cannot perceive the object shape correctly in this case. However, Fig. 2(c) shows our method performs well in understanding the accurate boundary of the tennis court, based on a semantic feature implying the height and width information from the other two tennis courts. Therefore, it is meaningful to construct a relation graph among objects and enhance the current object's feature using extra information aggregating from objects with high affinity. Our contributions can be summarized as follows:

- We propose an Affinity-Aware Relation Network, using the affinity among densely packed oriented objects to improve detection performance.
- A Graph Construction Module is proposed, designing KFIoU similarity to measure the affinity among objects and selecting high-quality neighbours for subsequent feature enhancement in a dynamic way.
- A Dynamic Enhancement Module is proposed, using the attention module to learn neighbourhood message and dynamically determining the weight for feature enhancement.
- Extensive experiments are conducted to show that the proposed two modules can notably improve detection performance based on two-stage methods.

2 Related work

2.1 Oriented Object Detection

Existing oriented object detection methods mainly improve the detection accuracy from three aspects: enhancing rotated features, designing sampling assignment strategy and exploring the representation of bounding box.

Feature enhancement mainly aims at densely packed objects with arbitrary orientations. RoI Transformer [5] and ReDet [10] respectively design a detector with rotation-invariance and rotation-equivariance. R3Det [42] proposes a Feature Refinement Module (FRM), improving the single-stage method performance to a level comparable to two-stage ones. Mask OBB [32], CenterMap Net [33], SCRDet [45], SCRDet++ [43] introduce pixel-level semantic information and provides more granular feature fusion branch.

Well-designed assigner alleviates the inconsistency between classification and regression task. Both DAL [22] and CFC-Net [20] incorporate the Intersection of Union(IoU) [11] metric, which directly reflects the localization capability of predicted boxes, into the assignment strategy of positive samples. SASM [13] dynamically selects the IoU threshold for each object according to its shape.

Oriented RepPoints[17] selects sample points not only from the classification and localization but also from the orientation and point-wise feature correlation.

Studies on box representation and loss function mainly contribute to solving the boundary problem in regression-based methods. BBAVector [49] and PolarDet [51] represent the bounding box in coordinate systems. CFA [7] proposes a convex hull representation method. Gliding Vertex [38] predicts quadrilateral by learning the offset of the four corners of the horizontal bounding boxes. RIL [21] adopts the Hungarian loss. CSL [41] and DCL [40] transform the regression into a classification problem. GWD [44], KLD [46] and KFIoU [47] model the oriented object as a Gaussian distribution to construct a new loss function. P2PLoss [48] describes the spatial distance and morphological similarity of two convex polygons. Unlike our approach, none of these methods consider learning additional information from the affinity among objects for feature enhancement.

2.2 Graph Convolutional Neural Networks

The graph convolutional neural network extends the convolutional neural network to the non-Euclidean space. The graph convolutions fall into two categories: spectral [1, 4, 15, 16, 35] and spatial [23, 6, 8, 28] methods.

The spectral methods define the convolution in the spectral domain via the convolution theorem. The first graph convolutional neural network SCNN [1] defines its operator in the spectral domain. ChebNet [4] and GCN [15] parameterize the convolution kernel, significantly reducing the time and space complexity. The spatial methods define the node correlation in the spatial domain. GNN [12] selects a fixed number of neighbour nodes by a random walk algorithm. GraphSAGE [8] divides the convolution process into sampling and aggregation. GAT [31] uses the attention mechanism to differentiate the aggregation of neighbour nodes. PGC [39] defines convolution as the sum of a specific sampling function multiplied by a particular weight function. Our approach uses the idea of graph convolution for neighbour message learning and feature aggregation.

3 Methods

3.1 Overview

An overview of the proposed Affinity-Aware Relation Network is illustrated in Fig. 3. The model consists of a basic two-stage detector, a Graph Construction Module(GC-Module) and a Dynamic Enhancement Module(DE-Module).

An image is first fed into the pipeline of the basic detector. The GC-Module uses the proposal quintuples from RPN as well as RRoI features from RRoI Align to calculate the affinity matrix and dynamically determines the threshold to filter out low-quality neighbours. For each proposal, GC-Module selects proposals(neighbour) sharing high similarity with the current proposal(reference).

The DE-Module consists of neighbour message learning and feature enhancement weight learning. Neighbour message learning performs an attention mech-

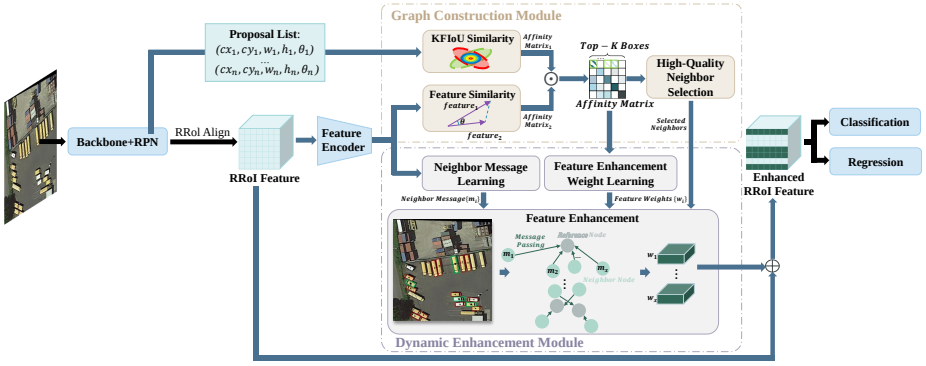


Fig. 3. An overview of the proposed AARN. Our approach is based on the basic two-stage detector ReDet.

anism over the high-quality neighbours to obtain messages and weights for aggregation. Feature enhancement weight learning determines the feature enhancement factor in consideration of the proposal aspect ratios. Then the neighbour message is used for node aggregation to get the aggregation feature of each node, and the feature enhancement weight is used to dynamically add the aggregation feature to the original feature to obtain the final enhanced feature. Finally, the detection result is achieved after classification and box regression branches of the basic detector. Our proposed method is based on the two-stage model ReDet [10], which in fact can be easily applied to various modern two-stage detectors.

3.2 Graph Construction Module

This module aims to construct a graph to represent the relationship between proposal regions. Formally, given N_r proposal regions of the input image, the relationship among regions can be modeled as an undirected graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$, where v_i in vertex set $\mathbf{V} = \{v_i\}_{i=1}^{N_r}$ corresponds to the i-th proposal and e_{ij} in $\mathbf{E} \in \mathbb{R}^{N_r \times N_r}$ quantifies the relationship between v_i and v_j. GC-Module calculates the affinity between proposal regions to filter out neighbours with low-similarity for each reference node, and then retains only the edges with high affinity in $\mathbf{G}$.

Affinity Matrix Calculation Affinity matrix $\mathbf{M} \in \mathbb{R}^{N_r \times N_r}$ reflects the similarity between proposals. We should consider two aspects when calculating the affinity: the semantic similarity inside the proposal and the shape similarity of the bounding box.

Feature Similarity. RRoI features characterize object semantics. Given a visual feature $\mathbf{F} \in \mathbb{R}^{N_r \times D}$ extracted by RRoI Align, we first employ a nonlinear transformation $\psi(\cdot) : \mathbb{R}^{N_r \times D} \rightarrow \mathbb{R}^{N_r \times L}$, projecting $\mathbf{F}$ into the latent semantic space denoted by

$$\mathbf{F}' = \psi(\mathbf{F}) \tag{1}$$

where $\mathbf{F}' \in \mathbb{R}^{N_r \times L}$. We adopt a simple form of $\psi(\cdot)$ which is implemented by a stack of two fully-connected layers followed by layer normalization and ReLU in order. Each row $f_i' \in \mathbf{F}'$ corresponds to a proposal's latent semantic feature. Then we apply the cosine similarity to calculate the semantic affinity matrix $\mathbf{M}_1$ between $f_i'(i = 1, 2, .., N_r)$, as shown in Eq. (2).

$$\mathbf{M_1}[i][j] = \frac{f_i' f_j'}{\|f_i'\| \, \|f_j'\|} \tag{2}$$

where $\|\cdot\|$ is a modulus operation.

KFIoU Similarity. The calculation of shape similarity should involve the height, width and rotation angle of objects. As an evaluation metric, IoU well combines these factors. To overcome the high computational complexity of Skew-IoU, we approximate oriented boxes as Gaussian distributions and use the overlap of two Gaussian distributions to measure the shape similarity, as shown in Fig. 4. The conversion from a rotated box to a Gaussian distribution has been discussed in some previous works [44, 47, 14], described as follows.

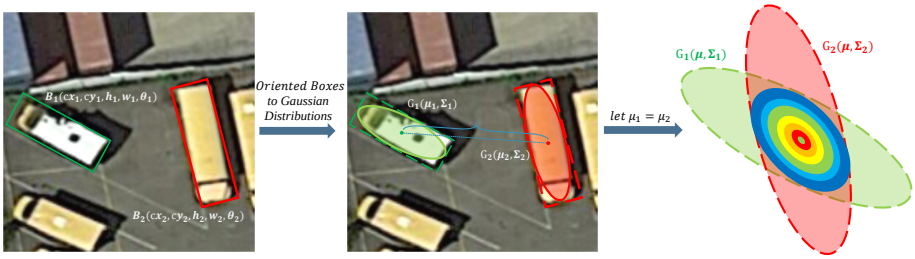


Fig. 4. First, we convert the oriented bounding boxes to Gaussian distributions. Then we make two Gaussian distributions be concentric and introduce Kalman Filter to simulate the distribution overlapping.

The oriented box can be represented by a quintuple $\mathbf{B}(cx, cy, h, w, \theta)$, where (cx, cy) are the center point coordinates. h,w and θ respectively refer to the height, width and rotation angle. The transformation from the proposal quintuple to the Gaussian distribution $\mathbf{N}(\mu, \Sigma)$ is shown in Eq. (3).

$$\Sigma^{\frac{1}{2}} = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix} \begin{pmatrix} \frac{w}{2} & 0 \\ 0 & \frac{h}{2} \end{pmatrix} \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix}, \qquad \mu = (cx, cy) \tag{3}$$

After obtaining a 2D Gaussian distribution, we can easily calculate the box area from its covariance of the corresponding distribution.

$$\mathbf{S_B}(\Sigma) = 4\sqrt{\prod eig(\Sigma)} = 4 \cdot |\Sigma|^{\frac{1}{2}} \tag{4}$$

Then the overlapping distribution can be intuitively derived by multiplication of two Gaussian distributions. The probability density function of multiplying two Gaussian distributions $\mathbf{N_1}(\mu_1, \Sigma_1)$ and $\mathbf{N_2}(\mu_2, \Sigma_2)$ can be expressed as

$$\mathbf{f_1}(X)\mathbf{f_2}(X) = \mathbf{S_g} \cdot \frac{1}{\sqrt{2\pi\Sigma}} e^{-\frac{1}{2}(\mathbf{X}-\mu)^T \Sigma^{-1}(\mathbf{X}-\mu)} \tag{5}$$

$$\mathbf{S_g} = \frac{1}{\sqrt{2\pi(\Sigma_1+\Sigma_2)}} e^{-\frac{1}{2}(\mu_1-\mu_2)^T(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)} \tag{6}$$

$$\mu = (\mu_2\Sigma_1 + \mu_1\Sigma_2)(\Sigma_1+\Sigma_2)^{-1}, \qquad \Sigma = \Sigma_1\Sigma_2(\Sigma_1+\Sigma_2)^{-1} \tag{7}$$

That is, the multiplication of two Gaussian distributions is equal to a compressed or enlarged Gaussian distribution. The constant **Sg** is a scaling factor.

Inspired by [47], we perform Kalman Filter to calculate the overlapping areas. Unlike the loss design in [47], the similarity should not be affected by the center distance. Therefore, we let $\mu_1 = \mu_2 = \mu$ to make two Gaussian distributions be concentric. In this case, **Sg** is decoupled from center points and IoU similarity can be calculated as Eq. (8).

$$\mathbf{IoU}(\mathbf{N_1}, \mathbf{N_2}) = \frac{\mathbf{S_B}(\Sigma)}{\mathbf{S_{B1}}(\Sigma_1) + \mathbf{S_{B2}}(\Sigma_2) - \mathbf{S_B}(\Sigma)} = \frac{\Sigma^{\frac{1}{2}}}{\Sigma_1^{\frac{1}{2}} + \Sigma_2^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}} \tag{8}$$

The shape affinity matrix $\mathbf{M_2} \in \mathbb{R}^{N_r \times N_r}$ is obtained by

$$m_{ij} = \mathbf{IoU}(\varphi(cx_i, cy_i, h_i, w_i, \theta_i), \varphi(cx_j, cy_j, h_j, w_j, \theta_j)) \tag{9}$$

where $m_{ij} \in \mathbf{M_2}$ and φ represents the box to Gaussian distribution function.

We use min-max normalization to scale the value of $\mathbf{M_1}$ and $\mathbf{M_2}$ ranging from 0 to 1. The final affinity matrix $\mathbf{M}$ satisfies $\mathbf{M} = \mathbf{M_1} \odot \mathbf{M_2}$, where $\odot$ represents the point-wise multiplication.

High-Quality Neighbour Selection Similar to ATSS [50], High-Quality Neighbour Selection is proposed to dynamically select high-quality neighbour nodes according to their statistical characteristics. We first keep the top-k largest values of each row in affinity matrix $\mathbf{M}$ for each proposal, and then use the mean and standard deviation of selected proposals' affinity values to determine the threshold $\mathbf{\Gamma_i}$ for i-th proposal.

$$\mathbf{\Gamma_i} = u_i + \sigma_i \tag{10}$$

$$u_i = \frac{1}{k}\sum_{j=idx_1}^{idx_k} m_{ij}, \qquad \sigma_i = \sqrt{\frac{1}{k}\sum_{j=idx_1}^{idx_k}(m_{ij}-u_i)^2} \tag{11}$$

where $m_{ij} \in \mathbf{M}$ and idx_i indicates the index of selected k boxes.

For each proposal, absorbing neighbours with inaccurate positions and shapes will degrade its detection performance due to the introduction of noise. Therefore, we perform a non-maximum suppression(NMS) on the neighbour nodes

according to the score from the RPN, so that the aggregation nodes tend to be samples from different positions rather than overlapping proposals from the adjacent center points.

3.3 Dynamic Enhancement Module

After determining the reference nodes and neighbour nodes, we design a dynamic enhancement module consisting of neighbour message learning and feature enhancement weight learning. The former uses the attention to learn neighbour message for node aggregation, and the latter dynamically determines the weight for feature enhancement.

Neighbour Message Learning We use an attention mechanism drawing global dependencies to learn the weighted messages between neighbour and reference nodes. As shown in Fig. 5, this module is implemented based on the Multi-Head Attention in [30]. Embedding Feature $\mathbf{F'} \in \mathbb{R}^{N_r \times L}$ is used as the query (Q), key (K) and value (V) of the Multi-Head Attention in [30] .

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \tag{12}$$

where the d_k is the channel dimension. $\mathbf{A} = \mathbf{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}) \in \mathbb{R}^{N_r \times N_r}$ represents the attention weight matrix used for neighbourhood aggregation subsequently. It is worth noting that Eq. (12) only displays the structure of the single head. In practice, multiple heads are concatenated to get the *Multi-Head Attention*$(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \in \mathbb{R}^{N_r \times L}$.

The final enhanced node features are obtained by residual connections, as shown in Eq. (13). Then the message m_{ij} delivered from the j-th neighbour to the i-th reference node can be expressed as Eq. (14).

$$Enhanced\_Node\_Feature = Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + Embedding\_Feature \tag{13}$$

$$m_{ij} = \mathbf{A}_{ij} \cdot Enhanced\_Node\_Feature_i \tag{14}$$

Feature Enhancement Weight Learning Fig. 6 displays two objects with the same angle offset ω. However, the box_1 with a lower aspect ratio outperforms box_2 on the IoU metric, indicating that objects with high aspect ratio are more sensitive to the angular deviation. Therefore, it is necessary to treat objects with high aspect ratios more cautiously in neighbourhood aggregation.

Intuitively, the message delivered to a reference node with high aspect ratio should be assigned a smaller weight before enhancement. Furthermore, for objects with drastic changes in aspect ratio, it tends to be difficult to learn a universal feature generalizing characteristic of all neighbour nodes. We should also tone down the enhancement of features from these objects.

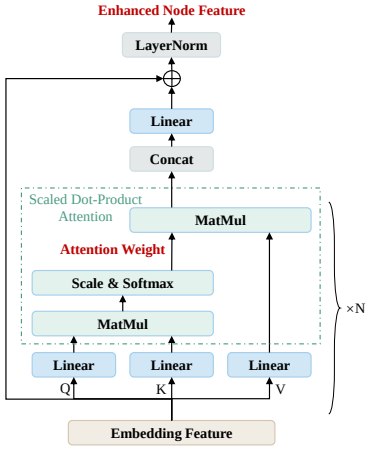


Fig. 5. Flowchart of the Multi-Head Attention Module. Enhanced Node Feature and Attention Weight Matrix respectively represent the neighbour message and the weight used for aggregation.

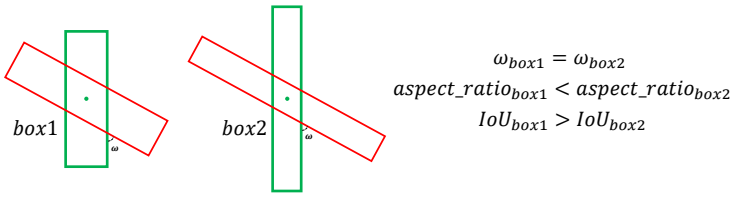


Fig. 6. The red and green box represent prediction and ground truth. The box_1 with a lower aspect ratio outperforms box_2 on the IoU metric, indicating that objects with high aspect ratio are more sensitive to the angular deviation.

In response, we design the Feature Enhancement Weight Learning, which can dynamically adjust enhancement weight w_i according to the i-th object's aspect ratio, as shown in Eq. (15)-(16).

$$u_i^{ratio} = \frac{1}{N_r} \sum_{j=idx_{i1}}^{idx_{ik}} r_j, \qquad \sigma_i^{ratio} = \sqrt{\frac{1}{Nr} \sum_{j=idx_{i1}}^{idx_{ik}} (r_j - u_i)^2} \tag{15}$$

$$w_i = (\alpha - e^{\frac{u_i^{ratio}}{\beta}}) \cdot e^{-\sigma_i^{ratio}} \tag{16}$$

where $idx_i = \{idx_{ij}\}_{j=1}^{k}$ denotes indices of k boxes most relevant to the i-th reference box, selected in High-Quality Neighbour Selection. And r_j is the aspect ratio of the j-th proposal. Eq. (15) computes the mean and standard deviation of the aspect ratios of the top-k boxes. The mean value reflects the estimated aspect ratio and the standard deviation implies the fluctuation of aspect ratio. Given $\alpha > 0$ and $\beta > 0$, w_i decreases as the mean or standard deviation increases.

Feature Enhancement and Final Prediction The enhanced feature of the i-th proposal is obtained by combining the original feature with the aggregation feature, described as

$$Enhanced\_feature_i = feature_i + \varepsilon \cdot \frac{w_i}{N_i} \sum_{j=1}^{N_i} m_{ij} \tag{17}$$

where N_i is the number of selected neighbours of the i-th reference proposal. The meaning of m_{ij} and w_i are as same as mentioned above. ε is a learnable parameter with an initial value of 1.0, in order to implement a dynamic residual connection. Finally, the enhanced features are fed into the classification and regression branches of the basic detector to get prediction results.

4 Experiments

4.1 Experimental Setup

Datasets *DOTA-v1.0*[36] is a large-scale dataset for oriented objects detection in aerial images, which contains 2806 images ranging from 800× 800 to 4k× 4k pixels, 188,282 instances and 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, validation and testing set are 1/2, 1/6, and 1/3, respectively. All images of training and validation set are split into 1024× 1024 with an overlap of 200 pixels during training.

HRSC2016[18] is a high-resolution optical remote sensing dataset for ship recognition, which contains 1061 images (436 for training, 181 for validation and 444 for testing) ranging from 300×300 to 1,500×512 pixels. All images of training and validation set are resized to 800×512 pixels during training.

UCAS-AOD[53] contains 1,510 images with approximately 659×1280 pixels, 14,596 instances and two categories: plane and car. Like other works[43, 44, 47], we randomly select 1100 images for training and 400 for testing.

Implementation Details We use a two-stage detector ReDet[10] as our baseline and ReResNet-50 pretrained on ImageNet[27] following ReDet as our backbone. All modules before the RRoI Align follow the settings of ReDet.

As for the implementation of AARN, we first use two linear layers of size 512 ($L = 512$) to learn the latent feature $\mathbf{F}'$ in Eq. (1). Then top $k = 9$ largest values of each row in the affinity matrix are kept to determine the threshold $\mathbf{\Gamma_i}$ for i-th proposal in Eq. (11). NMS with threshold = 0.1 is performed over selected neighbour nodes before DE-M to avoid the introduction of noise. For the neighbour message learning in DE-M, all linear layers of $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ produce outputs of dimension $d_{model} = L = 512$. And we employ $h = 8$ parallel attention

heads so $d_k = d_{model}/h = 64$ in Eq. (12). For the feature enhancement weight learning, α and β in Eq. (16) are set to 2 and 3.5 respectively.

In the inference phase, RPN will generate 2000 proposals. If such a large number of proposals are input into AAFN, great noise will be introduced. Therefore, we set filter threshold as 0.9 in line with scores from RPN stage, so that only boxes with high confidence can participate in graph construction and feature enhancement. The weights of modules before RRoI Align are frozen during training. We adopt a stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.0001, the momentum of 0.9 and weight decay of 0.0001. We train the model for 12, 40, 120 epochs on the DOTA, HRSC2016 and UCAS-AOD datasets. We use 2 TITAN RTX GPUs with a total batch size of 4 for training and one TITAN RTX GPU for inference.

4.2 Comparisons with the State-of-the-Art

Table 1 compares our method with the state-of-the-art detectors on DOTA-v1.0. Without random rotation and multi-scale data augmentation, we improve by **1.07%** AP over the baseline ReDet. Especially categories with low aspect ratios or less semantic information achieve more notable AP improvements: **4.7%** on roundabout (RA), **3.77%** on helicopter (HC), **2.21%** on soccer-ball-field (SF), and **2.04%** on ground-track-field (GTF). For multi-scale training with random rotation, our method achieves the state-of-the-art **80.79%** AP and the best performance on 6 categories. Fig. 7 displays results of ReDet, results of our method, and visualization of Graph Construction Module results on DOTA. In Fig. 7(a) and (b), we mark some instances which are accurate under our method but inaccurate under ReDet with white circles. It shows a better performance of our method. In Fig. 7(c), the reference box (green) is connected to its selected neighbour boxes (red). It can be found that a reference box always share high similarity in category and orientation with its neighbour boxes.

Table 2 lists the performances of our method and state-of-the-art detectors on HRSC2016. Our method achieves the best performance of **90.57%** under the VOC2007 metric. Table 3 shows results on UCAS-AOD. Our method achieves the state-of-the-art **89.94%** and **97.45%** mAP under VOC2007 and VOC2012 metrics respectively, and the mAP of VOC2012 improves by **1.22%**.

4.3 Ablation Study

To prove the effectiveness of our proposed method, we choose ReDet as our baseline and perform a detailed ablative analysis on DOTA-v1.0 test set. Following previous works, random horizontal flipping without any other tricks is applied for data augmentation. Ablation study result is shown in Table 4, which demonstrates the effectiveness of each module.

Affinity Matrix Calculation Affinity Matrix Calculation consists of feature similarity and shape similarity. As shown in Table 5, the absence of either com-

Table 1. AP for each class and AP_{50} on DOTA-v1.0. R-50, RX-101 and H-104 respectively stand for ResNet-50, ResNeXt-101 and Hourglass-104. MS/RR denotes random rotation and multi-scale used for augmentation during training.

Method	Backbone	MS/RR	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	AP_{50}
Single-stage/Anchor-free																		
Oriented RepPoints[17]	R-101		89.53	84.07	59.86	71.76	79.95	80.03	87.33	90.84	87.54	85.23	59.15	66.37	75.23	73.75	57.23	76.52
CFA[7]	R-152		89.08	83.20	54.37	66.87	**81.23**	80.96	87.17	90.21	84.32	86.09	52.34	69.94	75.52	**80.76**	67.96	76.67
O^2-DNet[34]	H-104	✓	89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03	71.04
DRN[24]	H-104	✓	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23
BBAVectors[49]	R-101	✓	88.63	84.06	52.13	69.56	78.26	80.40	88.06	90.87	87.23	86.39	56.11	65.62	67.10	72.08	63.96	75.36
CSL[41]	R-152	✓	**90.25**	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93	76.17
PolarDet[51]	R-101	✓	89.65	87.07	48.14	70.97	78.53	80.34	87.45	90.76	85.63	86.87	61.64	70.32	71.92	73.09	67.15	76.64
SASM[13]	R-101	✓	89.54	85.94	57.73	78.41	79.78	84.19	**89.25**	90.87	85.80	87.27	63.82	67.81	78.67	79.35	69.37	79.17
RetinaNet-P2P[48]	R-101	✓	89.22	86.12	55.23	**81.39**	80.34	83.45	88.25	90.87	86.63	87.08	71.74	69.87	77.34	76.01	59.59	79.15
Two/Refined-stage																		
Gliding Vertex[38]	R-101		89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
SCRDet++[43]	R-101		89.77	83.90	56.30	73.98	72.60	75.63	82.82	90.76	87.89	86.14	65.24	63.17	76.05	68.06	70.24	76.20
ReDet[10](baseline)	ReR-50		88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.89	87.78	85.75	61.76	60.39	75.96	68.07	63.59	76.25
Oriented R-CNN[37]	R-101		88.86	83.48	55.27	76.92	74.27	82.10	87.52	90.90	85.56	85.33	65.51	66.82	74.36	70.15	57.28	76.28
KFIoU[47]	R-101		89.04	84.04	52.98	73.00	78.69	83.60	87.61	90.79	85.97	85.47	64.77	63.29	69.18	76.38	65.63	76.70
Mask OBB[11]	RX-101	✓	89.56	85.95	54.21	72.90	76.52	74.16	85.63	89.85	83.81	86.48	54.89	69.64	73.94	69.06	63.32	75.33
S^2A-Ne[9]	R-50	✓	89.07	82.22	53.63	69.88	80.94	82.12	88.72	90.73	83.77	86.92	63.78	67.86	76.51	73.03	56.60	76.38
RSDet-II[25]	R-152	✓	89.93	84.45	53.77	74.35	71.52	78.31	78.12	91.14	87.35	86.93	65.64	65.17	75.35	79.74	63.31	76.34
R^3Det[42]	R-152	✓	89.80	83.77	48.11	66.77	78.76	83.27	87.84	90.82	85.38	85.51	65.67	62.68	67.53	78.56	72.62	76.47
DAL[22]	R-50	✓	89.69	83.11	55.03	71.00	78.30	81.90	88.46	90.89	84.97	87.46	64.41	65.65	76.86	72.09	64.35	76.95
DCL[40]	R-152	✓	89.26	83.60	53.54	72.76	79.04	82.56	87.31	90.67	86.59	86.98	67.49	66.88	73.29	70.56	69.99	77.37
OSKDet[19]	R-101	✓	90.04	**87.25**	54.41	79.48	72.66	80.29	88.20	90.84	83.91	86.90	63.39	**71.76**	75.63	72.59	69.75	77.81
GWD[44]	R-152	✓	89.66	84.99	59.26	82.19	78.97	84.83	87.70	90.21	86.54	86.85	**73.47**	67.77	76.92	79.22	74.92	80.23
KLD[46]	R-152	✓	89.92	85.13	59.19	81.33	78.82	84.38	87.50	89.80	87.33	87.00	72.57	71.35	77.12	79.34	**78.68**	80.63
AARN(Ours)	ReR-50		89.18	84.31	52.65	76.04	78.22	84.29	88.29	**91.87**	86.82	86.85	63.97	65.09	74.64	70.33	67.36	77.32
AARN-MS(Ours)	ReR-50	✓	89.60	85.72	**62.11**	81.18	78.98	**86.01**	88.68	90.90	**89.13**	**88.23**	69.90	68.68	**79.12**	78.72	74.89	**80.79**

Table 2. Performances of AARN and state-of-the-art detectors on HRSC2016.

Method	RoI-Trans[5]	Gliding Vertex[38]	R^3Det[42]	CFC[20]	DAL[22]	GWD[44]
mAP(07)	86.20	88.20	89.26	89.70	89.77	89.85
Method	KLD[46]	S^2A-Net	Oriented RepPoints[17]	ReDet[10]	Oriented R-CNN[37]	AARN(Ours)
mAP(07)	89.97	90.17	90.38	90.46	90.50	**90.57**

ponent results in a lower performance than baseline. We also discuss the effectiveness of different ways to compute shape similarity. Theta similarity refers to angle cosine similarity. SkewIoU refers to the regular IoU calculation between skewed boxes. KFIoU refers to the Gaussian distribution overlapping method in this paper, which achieves the highest **77.32%** mAP. It shows that KFIoU similarity can better describe the affinity of objects.

Table 3. Performances of AARN and state-of-the-art detectors on UCAS-AOD.

Method	VOC2007			VOC2012		
	Car	Plane	mAP	Car	Plane	mAP
RIDet-O[21]	88.88	90.35	89.62	-	-	-
DAL[22]	89.25	90.49	89.87	-	-	-
R^3Det[42]	-	-	-	94.14	98.20	96.17
SCRDet++[43]	-	-	-	94.97	98.93	96.95
OSKDet[19]	-	-	-	95.29	**99.09**	97.18
ReDet[10]	88.00	90.30	89.15	94.10	98.30	96.23
AARN(Ours)	**89.10**	**90.80**	**89.94**	96.30	98.60	**97.45**

Table 4. Ablation study for High-Quality Neighbour Selection (HQNS), Neighbour Message Learning (NML) and Feature Enhancement Weight Learning (FEWL) on DOTA-v1.0 test set.

Method	HQNS	NML	FEWL	AP_{50}	Individual Improvement	Total Improvement
ReDet	×	×	×	76.25	-	-
ReDet-AARN	✓	×	×	76.63	+0.38	+0.38
	✓	✓	×	77.04	+0.41	+0.79
	✓	×	✓	76.98	+0.35	+0.73
	✓	✓	✓	**77.32**	-	**+1.07**

High-Quality Neighbour Selection We discuss the effects of different selection strategies and NMS thresholds on the performance of objects with different aspect ratios. We collect the aspect ratio distribution for each category in Fig. 1. As shown in Table 6, the dynamic selection strategy contributes to reducing the

(a) (b) (c)

Fig. 7. (a) ReDet (the basic detector) detection results. (b) ReDet with AARN detection results. (c) Visualization of Graph Construction Module results.

Table 5. Performance of Affinity Matrix Calculation Module and comparisons of different methods to compute shape similarity.

Method	Feature Similarity	Shape Similarity			HQNS	NML	FEWL	AP_{50}
		Theta Similarity	SkewIoU Similarity	KFIoU Similarity				
ReDet	×	×			×	×	×	76.25
ReDet-AARN	✓	✓	×	×	✓	✓	✓	76.95
		×	✓	×				77.11
		×	×	✓				**77.32**
	✓	×						76.13
	×	✓						75.98

sensitivity of high aspect ratio objects, such as harbor (HA) and basketball-court (BC), to noisy neighbours. Furthermore, a lower neighbour NMS value, which corresponds to a more strict NMS strategy, aims to filter out more low-quality neighbour boxes with high aspect ratios.

Neighbour Message Learning We compare the performance of three methods to learn message passing weight in Table 7. Gaussian refers to using a Gaussian distribution on similarity to model the edge weight as [23]. Affinity refers to aggregation with only affinity values. Attention refers to the aggregation with a multi-head attention, which achieves the highest **77.32%** mAP and **0.21%** mAP improvements than gaussian modeling. It shows that the attention module can better perceive neighbour features and represent neighbour messages.

Feature Enhancement Weight Learning We compare the performance of different values of the two hyperparameters in Eq. (16). As shown in Table 8, the best performance is **77.32%** mAP when $\alpha = 2.0$ and $\beta = 3.5$. Especially for objects with low values and variances of aspect ratio such as roundabout (RA) and storage-tank (ST), the AP improvement is more obvious, with an increase of **1.12%** and **0.62%** AP compared to the situation without Feature Enhancement Weight Learning. It shows this dynamic feature enhancement strategy is especially effective for objects with low aspect ratios and little semantic information.

Table 6. Comparisons of different selection strategies and effects of different NMS thresholds on detection performance.

Method	HQNS			NML	FE WL	$\frac{\text{width}}{\text{height}} \gg 1$		$\frac{\text{width}}{\text{height}} \approx 1$		AP_{50}
	Threshold		Neighbour NMS							
	Dynamic Threshold	Fixed Threshold				HA	BC	RA	BD	
ReDet	×	×	×	×	×	**75.96**	**87.78**	**60.39**	**82.64**	**76.25**
ReDet-AARN	×	0.5	×			71.45	83.72	62.89	83.22	75.97
	×	0.7	×			72.50	85.96	62.12	82.93	76.23
	×	0.9	×			73.69	86.24	61.37	82.76	76.31
	✓	×	×			**74.13**	**86.52**	**63.47**	**83.62**	**76.52**
	✓	×	0.5			74.40	86.60	63.63	83.78	76.56
	✓	×	0.3			74.36	86.69	63.65	83.82	76.58
	✓	×	0.1			**74.51**	**86.73**	**63.70**	**83.81**	**76.63**

Table 7. Comparisons of different methods learning neighbourhood message.

Method	HQNS	NML			FEWL	AP_{50}
		Gaussian	Affinity	Attention		
ReDet-AARN	✓	×	✓	×	✓	76.98
		✓	×	×		77.11
		×	×	✓		**77.32**

Table 8. The performances on different value of two hyperparameters in Eq. (16) and effects of Feature Enhancement Weight Learning (FEWL) on objects with low aspect ratios.

Method	HQNS	NML	FEWL		$\frac{\text{width}}{\text{height}} \approx 1$ & std($\frac{\text{width}}{\text{height}}$) is low			AP_{50}
			Dynamic Feature Enhancement					
			α	β	RA	BD	ST	
ReDet-AARN	✓	✓	×		63.97	83.87	86.23	77.04
			2.0	4	64.93	84.35	86.73	77.27
				3.5	**65.09**	**84.31**	**86.85**	**77.32**
				3.0	65.00	84.04	86.94	77.23
			1.5	3.5	65.11	83.99	86.51	77.13
			2.5		65.07	84.29	86.32	77.09

5 Conclusions

In this paper, we propose an Affinity-Aware Relation Network, using the affinity among densely packed oriented objects, which consists of two parts: an affinity-graph construction module selecting bounding boxes sharing high similarity with the reference box, and a dynamic enhancement module using the attention module to learn neighbourhood message and dynamically determining the weight for feature enhancement. We conduct experiments on several public benchmarks and achieve the state-of-the-art performance.

Acknowledgements This work is supported by the Natural Resources Science and Technology Project of Anhui Province (Grant No. 2021-K-14).

References

1. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203 (2013)
2. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A language modeling framework for object detection. arXiv preprint arXiv:2109.10852 (2021)
3. Chen, X., Gupta, A.: An implementation of faster rcnn with study for region sampling. arXiv preprint arXiv:1702.02138 (2017)
4. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems **29** (2016)
5. Ding, J., Xue, N., Long, Y., Xia, G.S., Lu, Q.: Learning roi transformer for oriented object detection in aerial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2849–2858 (2019)
6. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E.: Neural message passing for quantum chemistry. In: International conference on machine learning. pp. 1263–1272. PMLR (2017)
7. Guo, Z., Liu, C., Zhang, X., Jiao, J., Ji, X., Ye, Q.: Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8792–8801 (2021)
8. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Advances in neural information processing systems **30** (2017)
9. Han, J., Ding, J., Li, J., Xia, G.S.: Align deep features for oriented object detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–11 (2021)
10. Han, J., Ding, J., Xue, N., Xia, G.S.: Redet: A rotation-equivariant detector for aerial object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2786–2795 (2021)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
12. Hechtlinger, Y., Chakravarti, P., Qin, J.: A generalization of convolutional neural networks to graph-structured data. arXiv preprint arXiv:1704.08165 (2017)
13. Hou, L., Lu, K., Xue, J., Li, Y.: Shape-adaptive selection and measurement for oriented object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022)
14. Huang, Z., Li, W., Xia, X.G., Tao, R.: A general gaussian heatmap label assignment for arbitrary-oriented object detection. IEEE Transactions on Image Processing **31**, 1895–1910 (2022)
15. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
16. Klicpera, J., Bojchevski, A., Günnemann, S.: Predict then propagate: Graph neural networks meet personalized pagerank. arXiv preprint arXiv:1810.05997 (2018)
17. Li, W., Chen, Y., Hu, K., Zhu, J.: Oriented reppoints for aerial object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1829–1838 (2022)
18. Liu, Z., Yuan, L., Weng, L., Yang, Y.: A high resolution optical satellite image dataset for ship recognition and some new baselines. In: International conference on pattern recognition applications and methods. vol. 2, pp. 324–331. SciTePress (2017)

19. Lu, D.: Oskdet: Towards orientation-sensitive keypoint localization for rotated object detection. arXiv preprint arXiv:2104.08697 (2021)
20. Ming, Q., Miao, L., Zhou, Z., Dong, Y.: Cfc-net: A critical feature capturing network for arbitrary-oriented object detection in remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing (2021)
21. Ming, Q., Miao, L., Zhou, Z., Yang, X., Dong, Y.: Optimization for arbitrary-oriented object detection via representation invariance loss. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2021)
22. Ming, Q., Zhou, Z., Miao, L., Zhang, H., Li, L.: Dynamic anchor learning for arbitrary-oriented object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2355–2363 (2021)
23. Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., Bronstein, M.M.: Geometric deep learning on graphs and manifolds using mixture model cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5115–5124 (2017)
24. Pan, X., Ren, Y., Sheng, K., Dong, W., Yuan, H., Guo, X., Ma, C., Xu, C.: Dynamic refinement network for oriented and densely packed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11207–11216 (2020)
25. Qian, W., Yang, X., Peng, S., Yan, J., Guo, Y.: Learning modulated loss for rotated object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 2458–2466 (2021)
26. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
27. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
28. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE transactions on neural networks **20**(1), 61–80 (2008)
29. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
32. Wang, J., Ding, J., Guo, H., Cheng, W., Pan, T., Yang, W.: Mask obb: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. Remote Sensing **11**(24), 2930 (2019)
33. Wang, J., Yang, W., Li, H.C., Zhang, H., Xia, G.S.: Learning center probability map for detecting objects in aerial images. IEEE Transactions on Geoscience and Remote Sensing **59**(5), 4307–4323 (2020)
34. Wei, H., Zhang, Y., Zhonghan, C., Li, H., Wang, H., Sun, X.: Oriented objects as pairs of middle lines. ISPRS Journal of Photogrammetry and Remote Sensing **169**, 268–279 (11 2020). https://doi.org/10.1016/j.isprsjprs.2020.09.022
35. Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., Weinberger, K.: Simplifying graph convolutional networks. In: International conference on machine learning. pp. 6861–6871. PMLR (2019)

36. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dota: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3974–3983 (2018)
37. Xie, X., Cheng, G., Wang, J., Yao, X., Han, J.: Oriented r-cnn for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3520–3529 (2021)
38. Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.S., Bai, X.: Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE transactions on pattern analysis and machine intelligence **43**(4), 1452–1459 (2020)
39. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
40. Yang, X., Hou, L., Zhou, Y., Wang, W., Yan, J.: Dense label encoding for boundary discontinuity free rotation detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15819–15829 (2021)
41. Yang, X., Yan, J.: Arbitrary-oriented object detection with circular smooth label. In: European Conference on Computer Vision. pp. 677–694. Springer (2020)
42. Yang, X., Yan, J., Feng, Z., He, T.: R3det: Refined single-stage detector with feature refinement for rotating object. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 3163–3171 (2021)
43. Yang, X., Yan, J., Liao, W., Yang, X., Tang, J., He, T.: Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
44. Yang, X., Yan, J., Ming, Q., Wang, W., Zhang, X., Tian, Q.: Rethinking rotated object detection with gaussian wasserstein distance loss. In: International Conference on Machine Learning. pp. 11830–11841. PMLR (2021)
45. Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., Fu, K.: Scrdet: Towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8232–8241 (2019)
46. Yang, X., Yang, X., Yang, J., Ming, Q., Wang, W., Tian, Q., Yan, J.: Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. Advances in Neural Information Processing Systems **34** (2021)
47. Yang, X., Zhou, Y., Zhang, G., Yang, J., Wang, W., Yan, J., Zhang, X., Tian, Q.: The kfiou loss for rotated object detection. arXiv preprint arXiv:2201.12558 (2022)
48. Yang, Y., Chen, J., Zhong, X., Deng, Y.: Polygon-to-polygon distance loss for rotated object detection. AAAI Conference on Artificial Intelligence (2022)
49. Yi, J., Wu, P., Liu, B., Huang, Q., Qu, H., Metaxas, D.: Oriented object detection in aerial images with box boundary-aware vectors. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2150–2159 (2021)
50. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9759–9768 (2020)
51. Zhao, P., Qu, Z., Bu, Y., Tan, W., Guan, Q.: Polardet: a fast, more precise detector for rotated target in aerial images. International Journal of Remote Sensing **42**(15), 5831–5861 (2021)

52. Zhou, L., Wei, H., Li, H., Zhao, W., Zhang, Y., Zhang, Y.: Arbitrary-oriented object detection in remote sensing images based on polar coordinates. IEEE Access **8**, 223373–223384 (2020)
53. Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., Jiao, J.: Orientation robust object detection in aerial images using deep convolutional neural network. In: 2015 IEEE International Conference on Image Processing (ICIP). pp. 3735–3739 (2015)
54. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)