AOGC: Anchor Free Oriented Object Detection Based on the Gaussian Centerness

Zechen Wang , Chun Bao , Jie Cao * , Qun Hao

Posted Date: 14 August 2023

doi: 10.20944/preprints202308.1011.v1

Keywords: remote sensing images; orientated object detection; one-stage; anchor-free; Gaussian kernal

Article

AOGC: Anchor Free Oriented Object Detection Based on the Gaussian Centerness

Zechen Wang [1,2], Chun Bao [1], Jie Cao [1,2,*] and Qun Hao [1,2]

[1] School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China; wangzechen@bit.edu.cn (Z.W.); baochun@bit.edu.cn (C.B.); qhao@bit.edu.cn(Q.H.)

[2] Yangtze River Delta Research Institute (Jiaxing), Beijing Institute of Technology, Jiaxing, Zhejiang 314003, China

* Correspondence: caojie@bit.edu.cn

Abstract: Oriented object detection is a challenging task in scene text detection and remote sensing image analysis, which has attracted extensive attention due to the development of deep learning in recent years. Currently, mainstream oriented object detectors are based on preset anchor boxes. This method increases the computational load of the network and causes a large amount of anchor box redundancy. In order to address this issue, we propose anchor-free oriented object detection method based on the Gaussian centerness(AOGC), a single-stage anchor-free detection method. Our method uses contextual attention FPN(CAFPN) to obtain the contextual information of the target. Then we design a label assignment method for oriented objects. Finally, we develop a Gaussian kernel-based centerness branch, which can effectively determine the significance of different anchors. AOGC achieves mAP of 74.30% on the DOTA-1.0 datasets and 89.80% on the HRSC2016 datasets, respectively. AOGC exhibits superior performance to other methods in oriented anchor-free object detection methods.

Keywords: remote sensing images; orientated object detection; one-stage; anchor-free; Gaussian kernal

1. Introduction

Due to the emergence of the CNN[1], the object detection method has developed rapidly in recent years and has reached a relatively mature stage. Object detection generally involves using the horizontal bounding box(HBB) to detect the targets. According to whether the anchor box is preset, it can be divided into anchor-based methods, such as Faster R-CNN[2], YOLO series[3–6], RetinaNet[7], and anchor-free based methods such as FCOS[8], CenterNet[9], CornerNet[10], etc. In recent years, oriented object detection(OBB) technology for remote sensing images has attracted extensive attention from researchers[11]. Still, this challenging task usually has the following problems: (1) When using the HBB method to detect objects, there is often significant overlap between bounding boxes due to their dense arrangement, as shown in Figure 1(a). (2) Most remote sensing images in size are large-scale images with many tiny targets, making detection difficult. (3) The aspect ratio of the target varies greatly. Due to remote sensing image data characteristics, the detected target usually has a large aspect ratio. The detection method using HBB often results in a small proportion of the target pixels in the bounding boxes, as shown in Figure 1(b). These problems have made it difficult for the HBB method to detect targets effectively, so we design an OBB method to detect targets in any direction.

The prevalent methods for oriented object detection are mostly anchor-based techniques. These anchor-based methods first preset many dense anchor boxes, and then the detector predicts the deviation between the preset box and the ground truth. RoI Transformer[12] is a typical two-stage algorithm based on anchor boxes. It borrows from the framework of Faster R-CNN[2]. The first stage will preset dense anchor boxes on the feature map, output horizontal proposals, and extract the features of the horizontal proposal through RoI Align. Unlike Faster R-CNN, it designs RRoI Learner to extract rotated features from horizontal features and use them for the second stage of learning.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

Remote sensing object detection is typically a large image with high resolution and a large number of small targets. The anchor-based methods often need to lay dense anchor boxes, which will cause the imbalance of positive and negative samples and the redundancy of anchor boxes in the first detection stage.

(**a**) (**b**)

Figure 1. Problems existing in remote sensing object detection using horizontal bounding boxes. (a) shows the overlapping phenomenon of anchor boxes when using horizontal bounding boxes to detect objects in aerial images. The green box is the detected horizontal bounding box. (b) shows the problem that the proportion of target pixels is less when using horizontal bounding boxes to detect objects with large aspect ratios in aerial images. The red area in the figure is the target pixel, and the green area is the background pixel.

Recently, some anchor-free detectors have also been developed. CenterNet[9] proposes a new method to directly predicts the bounding box of the target through the center point. To achieve non maximum suppression(NMS) free, they get the target by predicting the center points in each heatmap. However, CenterNet will require more effort to obtain effectual output for multiple targets whose center points coincide. FCOS[8] is a feature point-based detector that only predicts the vector of the bounding box on the feature points and requires less computation cost. To distinguish the quality of different feature points, FCOS also design a centerness branch to represent the distance from the feature point to the target center point. Previous work has proved that FCOS has achieved excellent results in HBB object detection. In order to enhance the precision of object detection in remote sensing images., FCOS requires better detection capabilities in this field. And the existing label assignment method and centerness branch for HBB are not suitable for detecting OBB objects.

Based on the above conclusions, we propose a novel anchor-free detector based on the baseline of FCOS. We use Residual Network(ResNet)[13] as the backbone, and to better extract the feature information of target, we design a novel Feature Pyramid Networks for Object Detection(FPN)[14] structure that uses the attention mechanism to extract context information. To adapt to the detection of oriented targets, we design a label assignment suitable for oriented targets in the detection head part and use a two-dimensional Gaussian kernel function to design the centerness branch. Finally, we conduct extensive experiments on two public oriented object detection datasets, DOTA and HRSC2016.

Our contributions are as follows:

(1) We propose a new anchor-free detector anchor-free oriented object detection based on the Gaussian centerness（AOGC), which uses FCOS as the baseline and adds a detection branch for oriented objects, and our model has a solid ability to detect oriented objects.

 doi:10.20944/preprints202308.1011.v1

(2) We design a FPN structure based on attention mechanism, which can effectively extract the targets' contextual information and improve the network's feature expression ability. This method is suitable for object detection in remote sensing images with more background pixels.

(3) We design a label assignment method suitable for rotating boxes, which can efficiently divide positive and negative samples and adapt to targets with large aspect ratios; secondly, we also design a Gaussian-based kernel function for oriented detection tasks. The centerness branch is used to determine the significance of different anchor points and improve the detection quality of the network.

(4) Our method achieves mAP of 74.30% and 89.80% on the DOTA and HRSC2016 datasets. The experimental results show that our method shows substantial improvement compared to the baseline method, surpassing most anchor-free and single-stage oriented object detection approaches.

2. Materials and Methods

2.1. Related Works

2.1.1. Horizontal Object Detection

As the CNN network continues to develop, the performance of object detectors also improves. Object detection generally refers to horizontal object detection, that is, to detect and locate the desired target with a horizontal bounding box. Mainstream horizontal object detection methods can be broadly classified according to the following criteria: two-stage and one-stage object detection.

Two-stage object detectors, such as the Faster R-CNN series[2,15,16], first generate RoIs, which can be roughly divided into background class and objects to be detected, and then extract RoI features in the second stage to perform fine classification and localization. Object detectors with two stages can offer higher detection accuracy, but they may have slower inference speeds. Object detectors like YOLO series[3–6], SSD[17], and RetinaNet[7] are single-stage detectors that predict the full detection results in one step. Single-stage detectors have faster real-time detection speed. However, they have lower accuracy compared to two-stage detectors. Due to the dense arrangement of remote sensing image targets and the sharp changes in oriented detection tasks, these horizontal object detectors often need help with problems such as a large proportion of background pixels and overlapping detection bounding boxes. Therefore, when performing object detection on remote sensing images, the oriented object detection method has far more advantages than the horizontal object detection method.

2.1.2. Oriented Object Detection

To solve the problem of the dense arrangement of targets in natural scenes and the rapid change of detection target size, oriented object detection has begun to appear and has received significant attention in natural scene texts and remote sensing images [18–35]. Typically, object detectors that orient objects use a basic object detector as a starting point and then incorporate specialized modules to estimate OBB from HBB.

For example, Rotation Region Proposal Networks(RRPN)[18] detects oriented objects by directly preset rotating anchor boxes. Rotational Region CNN(R^2CNN)[19] uses Faster R-CNN as the baseline, and RPN will generate rotating proposals for subsequent detection. RoI Transformer[12] learns oriented RoIs from horizontal RoIs through an RRoI learner. R^3Det[20] added a feature optimization module, which reconstructs the feature map through bilinear interpolation to solve the problem of feature misalignment caused by the change of the bounding box position. SCRDet[24] mitigates the effect of angular periodicity by designing a novel Intersection over Union(IoU) smoothing L1 loss. S^2A-Net[21] developed a new convolution method, which is different from the random offset of separable convolution, but first predicts a rotation box and then calculates the difference between the rotation box and the preset box to obtain the offset value of the convolution kernel achieves the alignment of the rotated features. Oriented object detection methods designed for dense detection

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

tasks have solved some existing problems to a certain extent, but the redundancy problem of anchor-based methods still exists.

2.1.3. Anchor-Free Detection

The utilization of a generalized design for anchors is a crucial component of Faster R-CNN. It's imperative to note that in anchor-based detectors, anchor boxes are considered predetermined sliding windows or proposal boxes, which must be categorized as either positive or negative. The network refines the bounding box location by predicting an additional offset regression. The object detector based on the design of the anchor box has become mainstream.

Anchor-based detectors are widely used in both horizontal and oriented object detection, as they hold a prominent position. But the abundance of preset anchor boxes on the feature map can lead to redundancy and an increase in the subsequent regression tasks and NMS calculations. Therefore, a corresponding anchor-free detector is designed to directly localize objects without manually defining anchor boxes. For example, FCOS[8] generates feature points on the feature map, and each feature point will return its distance vector to the target box and design a centerness branch to determine the significance of different feature points. CornerNet[10] obtains the positions of the upper left and lower right corners of the target through the Hourglass network, and then pairs the corner points through the embedding layer. CenterNet[9] directly returns the center point of the target, and then predicts the length and width of the target for the center point to obtain the final HBB. BBAVectors[22] is an oriented anchor-free detection method based on the CenterNet. It first predicts a heat map to indicate the position of the center point, then predicts the distance from the center point to the oriented boxes, and adds the length and width of the HBB bounding box. IENet[23] added two offsets, w and h, to realize the regression of the oriented boxes based on the FCOS regression horizontal boxes. Anchor-free methods have fast inference speeds and achieve competitive detection results compared to anchor-based object detection methods.

For oriented object detection, the anchor-based method must preset oriented anchor boxes with multiple angles, as shown in Figure 2. Presetting a large number of anchor boxes will lead to a very unbalanced ratio of positive and negative samples, and for densely arranged targets in remote sensing images, anchor boxes are prone to overlap, so oriented anchor-free detection methods are becoming more and more popular.

Figure 2. Anchor-based detector preset anchor box redundancy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

2.2. Method

2.2.1. Network Architecture

The framework of our proposed AOGC network is shown in Figure 3. AOGC is a single-stage anchor-free oriented object detector. It consists of ResNet, contextual attention FPN and Gaussian kernel anchor-free detection head. The last three feature maps C_3、C_4 and C_5 of ResNet are used as the input of the contextual attention FPN. The contextual attention FPN will establish the connection between the three feature layers and set a specific attention relationship to obtain the contextual information of the target to be detected. We will introduce the particular implementation of contextual attention FPN in Section 3.2. Contextual attention FPN will generate three feature layers P_3、P_4 and P_5, then P_6 and P_7 are obtained by downsampling from P_5. The head part will output three branches, namely the regression branch, the classification branch and the Gaussian centerness branch. The classification branch and the regression branch get the classification features and regression features after performing four 3 × 3 convolution layers on the feature map. The regression feature passes through a 1 × 1 convolution layer to obtain an output of $H \times W \times 5$, where H and W mean the length and width of the feature map. Each point in the feature map will output five regression vectors, which are the four distance vectors from the center point to left, top, right, and bottom sides of the bounding box (l, t, r, b). and a vector α representing the angle of the target predicted by the network. The categorical feature passes through a 1 × 1 convolutional layer to output the confidence of each category at each point. Finally, the Gaussian centerness branch we designed is generated from the regression branch feature and obtained through a convolution layer. It will determine the significance of each anchor point on the feature map.

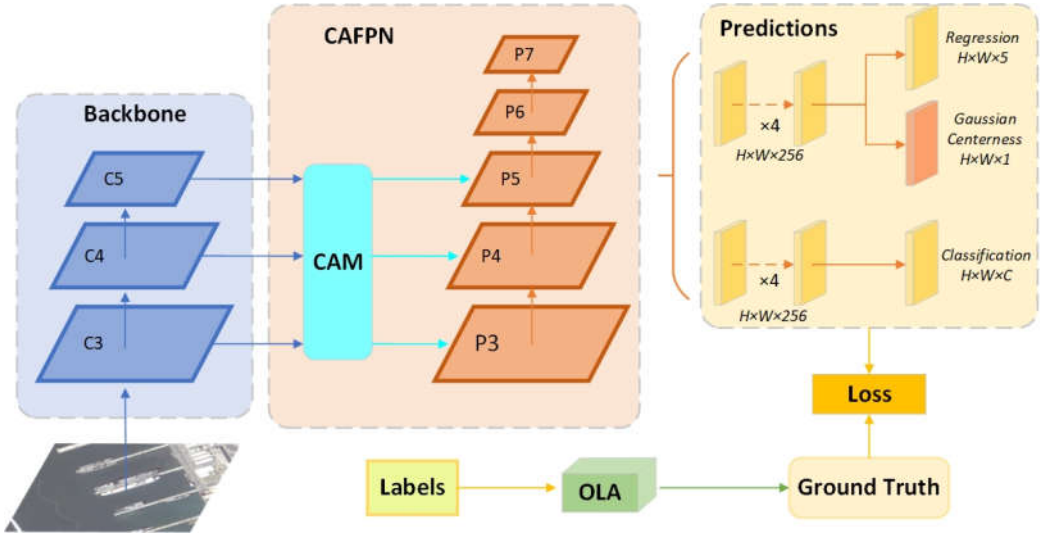


Figure 3. The framework of the AOGC network, where C_3, C_4, and C_5 represent the output features of the backbone, and CAM is the contextual attention module. H and W are the width and height of the feature map, and C is the number of categories in the network classification. OLA is the oriented label assignment method we designed, and Gaussian centerness is a centerness branch based on a two-dimensional Gaussian kernel.

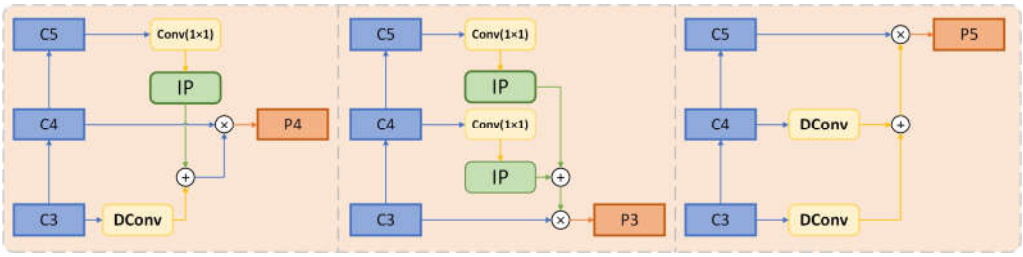
2.2.2. Contextual Attention FPN(CAFPN)

Remote sensing images usually have the characteristics of dense distribution of targets, small target size, and large proportion of background, which make it difficult to detect targets. Background information often contains a large amount of prior knowledge, for example, airplanes generally appear in airports, and ships typically appear in harbors. Such prior knowledge often plays a vital role in object detection. Many research studies have verified the importance of background

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

information in remote sensing object detection[36–38]. Attention mechanisms have also shown promise by obtaining the contextual information of the target and extracting the association between pixels in oriented object detection.

In order to better fuse contextual information of oriented objects, we design a contextual attention FPN structure. The structure is shown in Figure 4. We obtain the feature maps,C_3, C_4, and C_5, of the last three layers from the backbone as the input of the contextual attention module. Their feature maps can be expressed as $C_i \in \mathbb{R}^{H\times W\times C}$, i = 3, 4, 5, and finally, output the feature layer $P_i \in \mathbb{R}^{H\times W\times C}$, i = 3, 4, 5, 6, 7.



(**a**) (**b**) (**c**)

Figure 4. CAFPN structure diagram. Among them, DConv represents dilated convolution, IP means interpolation operation and Conv (1×1) represents 1×1 convolution for channel conversion.

Specifically, for each input feature map layer $C_i \in \mathbb{R}^{H\times W\times C}$, i = 3, 4, 5, we will get the contextual attention information from the remaining two feature maps $C_j \in \mathbb{R}^{H\times W\times C}$, $j \neq i$ and multiplied it with the feature map C_i to obtain P_i., the following formulas can describe the expression of:

$$P_4 = C_4 \cdot (I(C_5) + Dc(C_3)) \quad (1)$$

$$P_3 = C_3 \cdot (I(C_5) + I(C_4)) \quad (2)$$

$$P_5 = C_5 \cdot (Dc(C_3) + Dc(C_4)) \quad (3)$$

where the upsampling operation I is used to obtain the contextual information of the next-level feature map. Dc stands for dilated convolution [43]. The dilated convolution is employed because it has the ability to widen the range of the convolution layer's receptive field, thus resulting in the extraction of more comprehensive spatial information. We fused the upper and lower levels of information as a spatial attention mechanism and then multiplied it with the original feature layer to obtain context information. To match the channels of the fused feature map, we added a 1 × 1 convolution before the upsampling operation to perform channel conversion.

The acquisition of P_3 is shown in Figure 4b. We obtained the upsampling information from layer C_4 and layer C_5, then multiplied it with C_3. The acquisition of P_4 is shown in Figure 4a. We used dilated convolution to extract extensive receptive field features from the C_3 layer to fuse the upsampling information of the C_5, and then we multiplied it with C_4. The corresponding feature P_5 corresponds to the C_5 layer, as shown in Figure 4c. We extracted extensive receptive field features from the C_5 and C_4 layers by dilated convolution and multiplied them with C_5. In general, we extracted information from the other two feature layers using dilated convolutions from the layer with a low sampling rate and upsampling from the layer with a high sampling rate.

After extracting features through the contextual attention layer, our contextual attention FPN outputs three feature layers: P_3, P_4, and P_5. The remaining two feature layers, P_6 and P_7, were obtained by downsampling from the P_5 layer following the FCOS method. Finally, our contextual attention FPN outputs the feature layer, $P_i \in \mathbb{R}^{H\times W\times C}$, i = 3, 4, 5, 6, 7, for the classification and regression tasks of the subsequent Gaussian detection head.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

2.2.3. Oriented Bounding Box Label Assignment (OLA)

The anchor-based detector usually calculates the IoU between the preset box and ground truth and then assigns the sample as a positive sample or a negative sample by the size of the IoU value. The anchor-free detection network often divides the anchor points into positive samples and negative samples. For example, CenterNet defines that the center point of the ground truth falls on the heatmap as a positive sample, the regression label is 1, and other position points are negative samples, and the regression label is obtained according to the Gaussian distribution. FCOS divides the anchor points falling in the ground truth as positive sample points, and the rest of the points are divided into negative sample points. These label assignment methods based on horizontal boxes are no longer applicable for oriented object detection. As shown in Figure 5, it can be seen that the positive sample point area divided by the FCOS method and the ground truth are misaligned due to the change of angle. Therefore, we design an oriented label assignment method for oriented object detection.

Figure 5. The FCOS positive and negative sample division method does not match the OBB detection. The blue area in the figure is the positive sample point area of FCOS, and the red box is the ground truth.

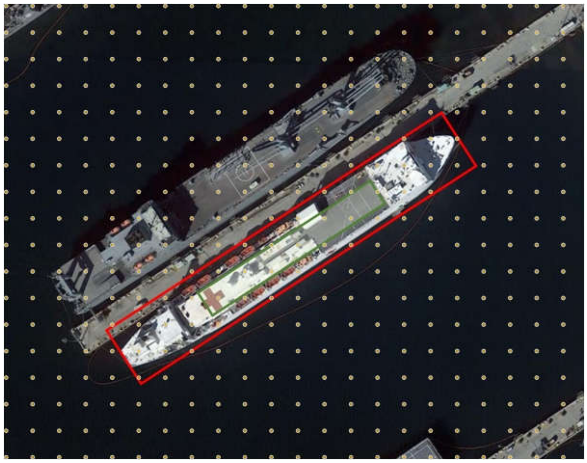
We defined the ground truth as $(x^*, y^*, h^*, w^*, \theta^*)$. (x^*, y^*) represent the coordinates of the center point of the ground truth. h^* and w^* represent the long and short sides of the ground truth, respectively. θ^* is the rotation angle of the ground truth, which means the angle between the long side h^* of the ground truth and the x-axis, and its scope is $(-\pi/2, \pi/2)$. For an anchor point (x, y) on the feature map, we first calculated the distance (x^*, y^*) from it to the ground truth center point D, and then we obtained the distance D' after rotating this distance map through affine transformation $D' = (x', y')$. The transformation process can be expressed as Equation (4).

$$D' = \begin{pmatrix} cos\theta^* & -sin\theta^* \\ sin\theta^* & cos\theta^* \end{pmatrix} D \tag{4}$$

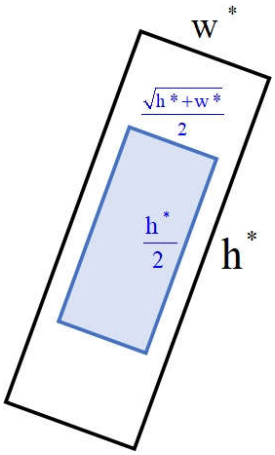
After obtaining the rotation coordinate distance (x', y') from the anchor point to the ground truth center point, we could divide the positive and negative samples by this distance. To improve the quality of positive samples, we first located the division range of positive and negative samples as $\frac{h^*}{2}$ and $\frac{w^*}{2}$, that is, the judge points that satisfy $x' < \frac{h^*}{2}$ and $y' < \frac{w^*}{2}$ as positive samples. The remaining points are positioned as negative sample points. However, we have observed that targets with large aspect ratios make up a significant proportion of remote sensing images, and the distance from the short side w of these large aspect ratio targets to the center point is particularly small.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

Therefore, setting the positive sample range as $\frac{w^*}{2}$ on the short side will cause the large aspect ratio target to have fewer positive sample points, which leads to the imbalance of positive and negative samples. Figure 6a shows that the green area is the positive sample sampling area. To solve this problem, we defined the division range at the short side as $\frac{\sqrt{h^* \cdot w^*}}{2}$, which effectively alleviates the pain of fewer positive sample points for large aspect ratio targets. Therefore, in the end, we judged the points that satisfy $x' < \frac{h^*}{2}$ and $y' < \frac{\sqrt{h^* \cdot w^*}}{2}$ as positive samples, and the rest of the points were positioned as negative samples. The representation of the sampling area is shown in Figure 6b.



(**a**)



(**b**)

Figure 6. Oriented label assignment method and its modification for large aspect ratio objects. (a) shows that after reducing the sampling area of the positive sample, the number of positive sample points of the target with a large aspect ratio is too small, (b) shows the sampling area after correction, and the blue area is the sampling area of the positive sample.

2.2.4. Gaussian Centerness Branch(GC)

Anchor-based detectors usually filter out low-quality anchor boxes by IoU thresholding. And anchor-free detectors do not produce anchor boxes; as such, methods based on IoU are not generally used to filter out low-quality anchor boxes. In fact, many low-quality prediction bounding boxes are often generated at positions far from the center of the target. FCOS designs a centerness branch to filter out these low-quality anchor boxes. Centerness describes the normalization from this position to the target center responsible for this position. This method has been adopted in HBB networks such as FCOS [5] and YOLOX [14]. Moreover, it has steadily improved detection accuracy, but it no longer applies to OBB detection tasks. Therefore, we proposed a novel centerness branch based on a two-dimensional Gaussian distribution. We used OBB parameters (x, y, h, w, θ) to define a two-dimensional Gaussian distribution:

$$\mathbf{\Sigma} = RAR^T \tag{5}$$

where $\mathbf{\Sigma}$ is the covariance matrix, R is the rotation transformation matrix formed by the sine and cosine angles of the target (its function is shown in (6)), R^T is the transpose of the rotation transformation matrix, and A represents the matrix of the covariance matrix produced by eigenvalue decomposition. In addition, the elements on the diagonal are eigenvalues arranged from the largest to the smallest, and its function is shown in (7).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

$$R = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix} \tag{6}$$

$$A = k \begin{pmatrix} w^2 & 0 \\ 0 & h^2 \end{pmatrix} \tag{7}$$

where k is a hyperparameter (which we set to 0.1 in this paper), and h and w represent the long and short sides of the target. Finally, we uses a normalized 2D Gaussian kernel function $g(X)$ to represent the centerness of the OBB task:

$$g(X) = exp(-\frac{1}{2}(X - \mu)^T \mathbf{\Sigma}^{-1}(X - \mu)) \tag{8}$$

where X represents the preset anchor point coordinates, μ represents the center point coordinates of the ground truth, and $exp(\ )$ is an exponential function. We used the normalized Gaussian centerness $g(X) \in (0, 1)$. Gaussian centerness will generate an elliptical area with the center point of the ground truth as the core. The closer to the center point, the higher its value; the closer to the boundary, the lower the value. Therefore, two-dimensional Gaussian kernel functions can effectively reflect the significance of the anchor point to the ground truth; furthermore, adding the Gaussian centerness branch can effectively filter out certain unimportant anchor points. Figure 7 shows a Gaussian ground truth heat map of an oriented object detection target.

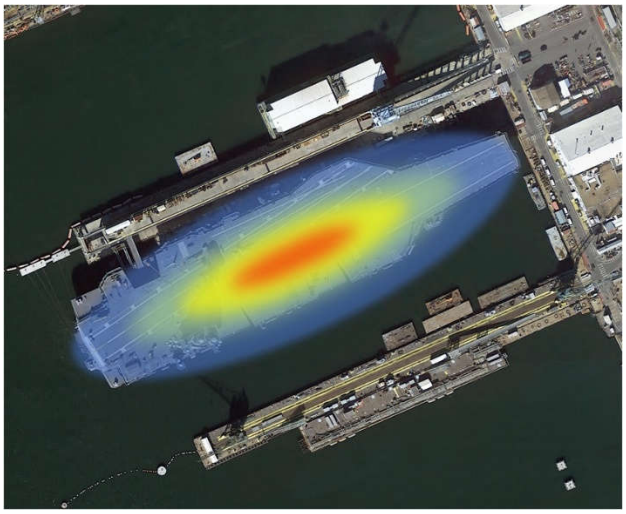


Figure 7. The representation heat map of Gaussian centerness.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

2.2.5. Loss Function

The final loss function of AOGC includes regression loss, classification loss, and Gaussian centerness loss. The function for calculating the total loss is displayed in equation (9).

$$Loss = \frac{1}{N_{pos}} \sum_{i} w_{GC} L_{reg} + \frac{\lambda_1}{N} \sum_{l} L_{cls} + \frac{\lambda_2}{N_{pos}} \sum_{i} L_{GC} \quad (9)$$

where N represents the number of all the ground truth, N_{pos} represents the number of positive samples in the ground truth, and i represents the preset anchor point. λ_1 and λ_2 are hyperparameters used to adjust the loss ratio, and we ended up setting them to 1 in our experiment. w_{GC} is the weight represented by the Gaussian centerness, which is used to adjust the size of the regression loss at different positions. L_{cls} means classification loss. We used focal loss [17] to address sample imbalance, and its expression is as follows:

$$L_{cls} = -\alpha(1-p)^{\gamma} log(p) \quad (10)$$

where α and γ represent the hyperparameters of the balanced sample, which we set to 0.25 and 2, respectively. *P* represents the probability that the model predicts a certain category. For regression loss, we used SkewIoU [18] loss to express the following:

$$L_{reg} = -log(S_{IoU}(R, R^*)) \quad (11)$$

3. Results

3.1. datasets

We evaluate our method on DOTA-1.0 and HRSC2016 datasets. DOTA-1.0[40] dataset is a collection of remote sensing images that are used for detecting oriented objects. Consisting of nearly three thousand aerial images with diverse scales, orientations, and object shapes, these images come from a range of sensors and platforms. Their resolution varies between 800 × 800 and 4000 × 4000. Notably, the fully annotated images contain 188,282 instances. DOTA-1.0 has 15 categories: Plane (PL), Baseball Field (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). DOTA images involve various large and small objects. The DOTA data set is divided into training set, validation set, and test set, and the ratios are 1/2, 1/6, and 1/3, respectively. We train on the training set and validation set and, test on the test set, then send the final results to the DOTA evaluation service for evaluation. We crop the original image into 1024 × 1024 patches with a gap of 200. Only random horizontal flips are used during training. For multi-scale training and testing, we choose three scales (0.5, 1.0, and 1.5) to resize the original images and then crop them into 1024 × 1024 patches with a gap of 500 while training with random rotations.

HRSC2016[41] is a dataset focuses on ship detection and includes 1061 images of rotating ships with different aspect ratios. The images were gathered from six well-known ports and include ships both at sea and offshore. This data set has the following detection difficulties: 1. There are numerous ships present on the shore, displaying densely arranged and distributed characteristics. The labeling frames overlap to a great extent. 2. The background of remote sensing images is intricate, and the texture of the ship to be analyzed is comparable to that of the nearby shoreline. 3. The scale of ships varies greatly, with different sizes visible in the same image. 4. There are multiple ship types, with dozens of different variations, making detection and classification challenging. 5. Problems such as cloud and fog occlusion make detection difficult. The dataset consists of images with varying pixel ranges from 300 × 300 to 1500 × 900. And the ground sample distance varies between 2 meters and 0.4 meters. Following R²CNN[19], we split the dataset into three sets: training, validation, and test. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

training set has 436 images with 1207 instances, the validation set has 181 images with 541 instances, and the test set has 444 images with 1228 instances. The training set and validation set are used for training, and the test set is used for evaluation. We evaluate our results using PASCAL VOC07 and VOC12[42] metrics. We resize all images to 800 × 800 without changing the aspect ratio.

3.2. Implementation Details

Our AOGC network uses ResNet50[13] as the backbone. Our experiments are performed with a batch size of 2 on a computer equipped with two 3080Ti GPUs. We use the ResNet50 model trained on ImageNet[43] as the pre-trained model during training. We use the Stochastic Gradient Descent (SGD) optimizer to train our models with an initial learning rate of 0.005, a momentum of 0.9, and a weight decay of 0.0001. And we train on the DOTA dataset for 12 epochs and reduce the learning rate to one-tenth of the original at the end of epoch 8 and epoch 11. We trained on HRSC2016 for 36 epochs and reduced the learning rate to one-tenth of the original at the end of epoch 24 and epoch 33. During testing, the confidence threshold is set to 0.1. We have implemented our training using mmdetection[44].

3.3. Ablation Studies

We conducted a series of ablation experiments on the DOTA-1.0 test set to evaluate the effectiveness of the proposed method. We use FCOS with ResNet50 as the baseline. We add an angle branch to achieve oriented object detection to predict the angle. We named it FCOS-R and our experimental results are shown in Table 1.

Table 1. Ablation experimental results on the DOTA dataset. Where, CAFPN represents the contextual attention FPN module we designed, OLA represents the oriented label assignment module, and GC represents the Gaussian centerness module

Methods	Backbone	CAFPN	OLA	GC	mAP (%)
FCOS-R	Resnet50				69.58
		√			72.19
		√	√		73.39
		√	√	√	74.30

As shown in Table 1, the mAP of FCOS-R at baseline is 69.58%. When added to our CAFPN, the accuracy on DOTA1.0 reaches 72.19%. When adding our positive and negative sample division method OLA module, the detection accuracy increased to 73.39%. Finally, adding our Gaussian centerness, the detection accuracy reached 74.30%.

3.4. Comparison with the State-of-the-art methods

The test results on the DOTA-1.0 dataset are shown in Table 2, and we compare them with some state-of-the-art one-stage, two-stage, and anchor-free oriented object detection methods, respectively. Our AOGC achieves 74.30% mAP on the DOTA dataset, surpassing most of the anchor-free and single-stage detection models, and has similar accuracy to some two-stage detection models. After adding multi-scale training and random rotation, our accuracy reached 76.55% mAP. Furthermore, our model achieves state-of-the-art results on challenging object categories, such as bridges, harbors, and swimming pools. Some of our test results on DOTA are shown in Figure 8.

The dataset labeled as HRSC2016 is comprised of multiple ship instances situated closely together, featuring varying orientations and significant aspect ratios. The test results of HRSC2016 are shown in Table 3, and our AOGC method works well on HRSC2016. Our AOGC achieves 89.80% mAP(07) and 95.20% mAP(12) o4wn the HRSC2016 dataset, surpassing most current anchor-free and single-stage detection models, which is a result of comparing state-of-the-art methods. Some of our test results on HRSC2016 are shown in Figure 9.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

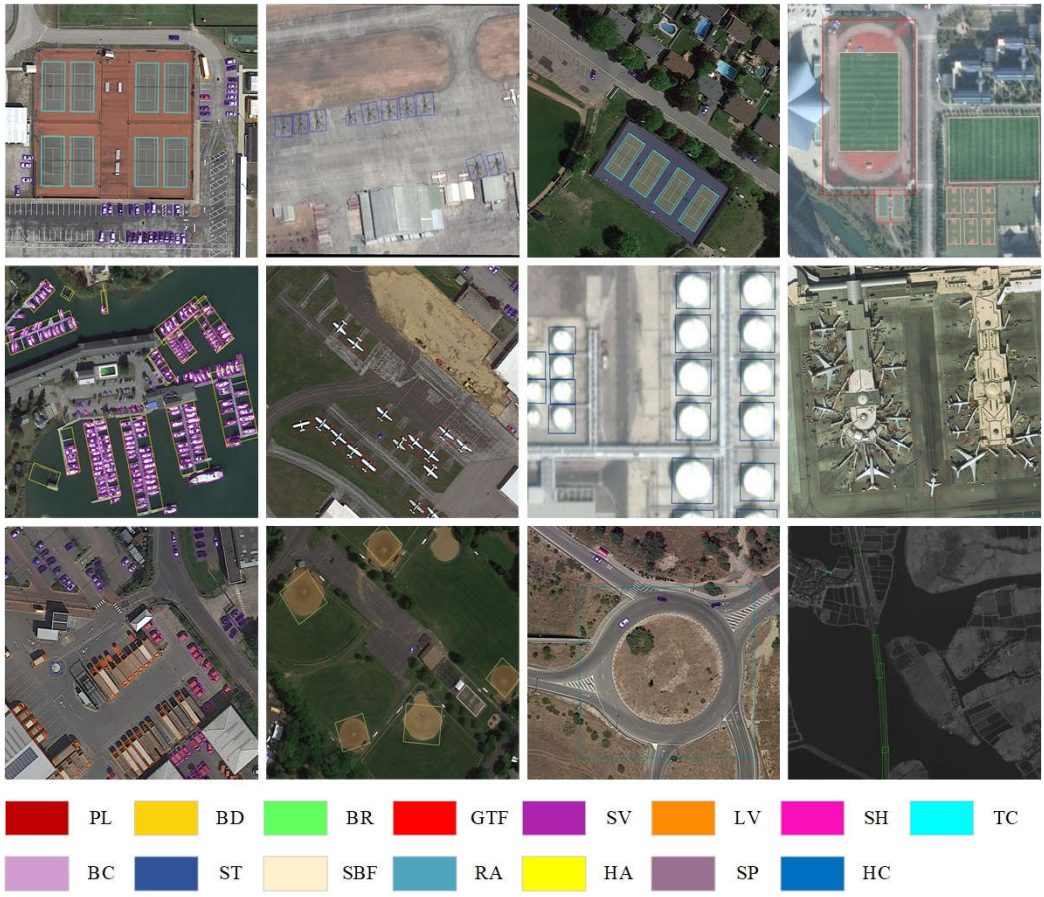


Figure 8. Partial visualization results of our method on the DOTA-1.0 dataset.

Table 2. Comparison with DOTA1.0 test results of other state-of-the-art methods. * indicates multi-scale training.

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP(%)
One-stage																	
Retina-Net-O	R-50	88.67	77.62	41.81	58.17	74.58	71.64	79.11	90.29	82.18	74.32	54.75	60.60	62.57	69.67	60.64	68.43
S2A-Net[21]	R-50	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83	84.90	85.64	60.36	62.60	65.26	69.13	57.94	74.12
DAL[25]	R-50	88.68	76.55	45.08	66.80	67.00	76.76	79.74	90.84	79.54	78.45	57.71	62.27	69.05	73.14	60.11	71.44
R3Det[20]	R-101	88.76	83.09	50.91	67.27	76.23	80.39	86.72	90.78	84.68	83.24	61.98	61.35	66.91	70.63	53.94	73.79
Two-stage																	
RRPN[18]	R-101	88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
RoI transformer[12]	R-101	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
SCRDet[24]	R-101	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	**87.94**	**86.86**	**65.02**	66.68	66.25	68.24	**65.21**	72.61
Gliding Vertex[26]	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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

Anchor-free																	
IENet[23]	R-101	88.15	71.38	34.26	51.78	63.73	65.63	71.61	90.11	71.07	73.63	37.62	41.52	48.07	60.53	49.53	61.24
Axis learning[27]	R-101	79.53	77.15	38.59	61.15	67.53	70.49	76.30	89.66	79.07	83.53	47.27	61.01	56.28	66.06	36.05	65.98
BBAVectors[22]	R-101	88.35	79.96	50.69	62.18	78.43	78.98	87.94	90.85	83.58	84.35	54.13	60.24	65.22	64.28	55.70	72.32
DRN[28]	H-104	88.91	80.22	43.52	63.35	73.48	70.69	84.94	90.14	83.85	84.11	50.12	58.41	67.62	68.60	52.50	70.70
O2-DNet[29]	H-104	89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.26	60.93	60.17	58.21	66.98	61.03	71.04
ProbIoU[30]	R-50	89.09	72.15	46.92	62.22	75.78	74.70	86.62	89.59	78.35	83.15	55.83	64.01	65.50	65.46	46.32	70.04
AOGC(ours)	R-50	84.04	80.61	52.23	67.23	80.64	81.75	87.81	90.91	82.81	84.91	56.55	65.73	73.50	72.50	53.32	74.30
AOGC*(ours)	R-50	83.44	80.29	**54.06**	70.90	**81.52**	**83.42**	**88.24**	90.88	83.02	86.84	60.34	64.90	**75.36**	**80.23**	64.80	**76.55**

Table 3. Comparison with HRSC2016 test results of other state-of-the-art methods. mAP(07) and mAP(12) represent VOC2007 index and VOC2012 index respectively.

Method	Backbone	mAP(07)	mAP(12)
R²CNN[19]	R-101	73.07	79.73
RRPN[18]	R-101	79.08	85.64
Axis Learning[27]	R-101	78.20	-
BBAVectors[22]	R-101	88.60	-
PIoU[31]	DLA-34	89.20	-
RoI Transformer[12]	R-101	86.20	-
DAL[25]	R-101	88.60	-
S²A-Net[21]	R-101	**90.17**	95.01
AOGC(ours)	R-50	89.80	**95.20**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

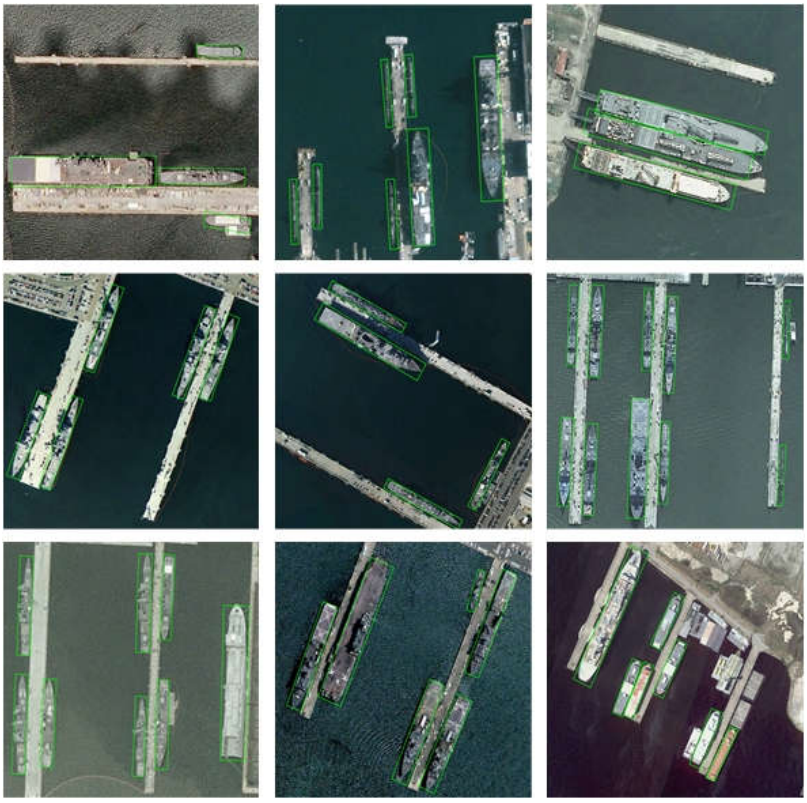


Figure 9. Partial visualization results of our method on the HRSC2016 dataset.

4. Discussion

4.1. Effect of the Proposed CAFPN

The input image is a series of feature maps of different scales obtained through the backbone network, but the feature information in the feature maps of different scales needs to be more balanced. In general, the deep feature map contains a large number of semantic features and less positioning information. In comparison, the shallow feature map carries more position information, but the semantic features are weaker, so the features need to be further enhanced. To achieve multi-scale informative feature fusion, FPN creates high-level semantic feature maps at all scales using a top-down architecture and lateral connections. This structure can integrate deep features and underlying information, strengthen the relationship between features of different scales, and each pyramid feature layer is only responsible for detecting objects within a specific scale, which improves the efficiency of detection tasks. However, FPN is a top-down network that can only transfer deep semantic information to shallow layers. Although multi-scale semantic expression is added, the positioning information between features needs to be more effectively circulated. In addition, the previous work[45] has proved that the top-down structure of The FPN feature will enhance the network's focus on smaller targets and lose the feature information of large targets in the process of gradient backpropagation.

The CAFPN structure we designed based on the above conclusions will effectively avoid these problems. Each feature layer of our CAFPN structure will fuse the shallow semantic information obtained in its adjacent upper-level feature layer with the deep semantic information obtained in its adjacent lower-level feature layer. The structure we designed achieves sufficient semantic information fusion and will not cause information loss due to too much attention to a particular category during the gradient backpropagation process. In order to reduce the parameter amount of the CAFPN network, we canceled the top-down pathway of the FPN structure. The ablation experiments for the FPN structure are shown in Table 4. We conduct our ablation experiments on the DOTA-1.0 dataset. The baseline is the FCOS-R network mentioned in Section 4.1. The experimental

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2023 doi:10.20944/preprints202308.1011.v1

results show that our CAFPN structure improves mAP by 2.61%, and our network parameters Params and FLOPs only increase by about 10%.

Table 4. Ablation experiment table of CAFPN structure.

Method	Backbone	Neck	mAP(%)	Params(MB)	FLOPs(GB)
Baseline	ResNet50	FPN	69.58	206.91	31.92
Proposed Method		CAFPN	72.19	220.74	35.07

4.2. Effect of the Proposed Gaussian kernel anchor-free detection head

We have modified the FCOS method for detecting oriented objects in remote sensing images. Our approach involves adding a Gaussian kernel anchor-free detection head with a specially designed oriented label assignment and Gaussian centerness branch to the existing FCOS detection head.

Before training an object detector, it is a necessary process to determine which ground truth(or background) each anchor should be assigned to, and the positive and negative sample division methods will directly affect the performance of the object detector. Anchor-based detectors usually use a certain threshold of IoU as the allocation criterion, while anchor-free detectors define positive and negative samples by directly assigning anchor points. FCOS directly assigns anchor points inside the bounding box as positive samples, showing good detection performance on horizontal object detection. To enable FCOS to have better performance in oriented object detection, we design the oriented label assignment method OLA. It maps the original FCOS positive sample area to the rotation box through affine transformation, and we improve the quality of the positive sample by shrinking the positive sample area. In order to adapt to a large number of large aspect ratio targets in remote sensing images, we also correct the short side of the sampling area. Our approach has proven to be highly effective for objects that possess large aspect ratios, as demonstrated by the outcomes of our testing on the DOTA-1.0 dataset. And our method achieves the best results on large aspect ratio target classes such as small vehicles, large vehicles, and Ships and achieves state-of-the-art.

In the FCOS model, the centerness branch plays a crucial role in determining the significance of anchor points based on their distance from the center point of the target. In order to make FCOS have better oriented object detection performance, we built a new centerness branch based on the principle of the two-dimensional Gaussian kernel function. Using our centerness branch has the following advantages: (1) The 2D Gaussian kernel function has the characteristic that the weight of the pixel increases and decreases monotonically with the distance from the point to the center point, which meets the design requirements of the centerness branch. (2) The 2D Gaussian kernel function has rotational symmetry, and its smoothness in all directions is the same, which is suitable for oriented target detection. (3) The smoothness of the 2D Gaussian kernel function can be adjusted by setting hyperparameters.

5. Conclusions

In this paper, we propose a novel anchor-free object detection method AOGC, which can be widely used in oriented object detection for aerial images. In our method, we propose the CAFPN module to obtain contextual attention information about objects and enhance the network's capability to extract features. Then, aiming at the positive and negative sample division problem in the current anchor-free detection network, a label assignment method suitable for oriented object detection is designed. Finally, we develop a Gaussian centerness branch suitable for oriented object detection to select high-quality anchors. Comprehensive experiments show that our AOGC helps improve detection accuracy. We conduct extensive experiments on DOTA-1.0 and HRSC2016 to validate our method, and the results show that our method outperforms most anchor-free and single-stage

detection methods. In future work, we will continue to improve our method and continue to explore the potential of oriented anchor-free detection methods for object detection in aerial images.

Author Contributions: Conceptualization, Z.W., J.C. and Q.H.; methodology, Z.W.; software, J.C.; validation, Z.W., C.B. and J.C.; formal analysis, C.B. and Z.W.; investigation, Z.W; resources, Q.H.; data curation, Z.W.; writing—original draft preparation, Z.W.; writing—review and editing, C.B and J.C.; visualization, Z.W.; supervision, J.C. and Q.H.; project administration, Z.W.; funding acquisition, J.C. and Q.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (62275022), Beijing Nature Science Foundation of China (No. 4222017) and the funding of Science. And Technology Entry program under grant (KJFGS-QTZCHT-2022-008).

Data Availability Statement: The DOTA and HRSC2016 are available at following https://captainwhu.github.io/DOTA/dataset.html and https://sites.google.com/site/hrsc2016/, respectively.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the Ieee 1998, 86, 2278-2324, doi:10.1109/5.726791.
2. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence 2017, 39, 1137-1149.
3. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. 2015, arXiv:1506.02640, doi:10.48550/arXiv.1506.02640.
4. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. arXiv e-prints 2018.
5. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. 2021.
6. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv e-prints 2022.
7. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis & Machine Intelligence 2017, PP, 2999-3007.
8. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. 2019.
9. Zhou, X.; Wang, D.; Krhenbühl, P. Objects as Points. 2019.
10. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision, 2018.
11. Wen, L.; Cheng, Y.; Fang, Y.; Li, X.Y. A comprehensive survey of oriented object detection in remote sensing images. Expert Systems with Applications 2023, 224, doi:10.1016/j.eswa.2023.119960.
12. Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; Lu, Q. Learning RoI Transformer for Detecting Oriented Objects in Aerial Images. 2018, arXiv:1812.00155, doi:10.48550/arXiv.1812.00155.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. IEEE 2016.
14. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. IEEE Computer Society 2017.
15. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. IEEE Computer Society 2014.
16. Girshick, R. Fast R-CNN. 2015.
17. Berg, A.C.; Fu, C.Y.; Szegedy, C.; Anguelov, D.; Erhan, D.; Reed, S.; Liu, W. SSD: Single Shot MultiBox Detector. 2015.
18. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. 2017, 1-1.
19. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational Region CNN for Orientation Robust Scene Text Detection. 2017.
20. Yang, X.; Liu, Q.; Yan, J.; Li, A.; Zhang, Z.; Yu, G. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. 2019.
21. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align Deep Features for Oriented Object Detection. 2020.
22. Yi, J.; Wu, P.; Liu, B.; Huang, Q.; Qu, H.; Metaxas, D. Oriented Object Detection in Aerial Images with Box Boundary-Aware Vectors. 2020.

23. Lin, Y.; Feng, P.; Guan, J. IENet: Interacting Embranchment One Stage Anchor Free Detector for Orientation Aerial Object Detection. 2019.
24. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Xian, S.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the International Conference on Computer Vision.
25. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L. Dynamic Anchor Learning for Arbitrary-Oriented Object Detection. 2020.
26. 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 2021.
27. Xiao, Z.; Qian, L.; Shao, W.; Tan, X.; Wang, K. Axis Learning for Orientated Objects Detection in Aerial Images. Remote Sensing 2020, 12, 908.
28. 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. 2020, arXiv:2005.09973, doi:10.48550/arXiv.2005.09973.
29. D, H.W.A.B.C.; B, Y.Z.A.; D, Z.C.A.B.C.; B, H.L.A.; B, H.W.A.; D, X.S.A.B.C. Oriented objects as pairs of middle lines. ISPRS Journal of Photogrammetry and Remote Sensing 2020, 169, 268-279.
30. Llerena, J.M.; Zeni, L.F.; Kristen, L.N.; Jung, C. Gaussian Bounding Boxes and Probabilistic Intersection-over-Union for Object Detection. 2021.
31. Chen, Z.; Chen, K.; Lin, W.; See, J.; Yu, H.; Ke, Y.; Yang, C. PIoU Loss: Towards Accurate Oriented Object Detection in Complex Environments. arXiv e-prints 2020.
32. Liu, L.; Pan, Z.; Lei, B. Learning a Rotation Invariant Detector with Rotatable Bounding Box. 2017.
33. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Xian, S.; Fu, K. R2CNN++: Multi-Dimensional Attention Based Rotation Invariant Detector with Robust Anchor Strategy. 2018.
34. Qian, W.; Yang, X.; Peng, S.; Guo, Y.; Yan, J. Learning Modulated Loss for Rotated Object Detection. 2019.
35. Yang, X.; Hou, L.; Zhou, Y.; Wang, W.; Yan, J. Dense Label Encoding for Boundary Discontinuity Free Rotation Detection. 2020, arXiv:2011.09670, doi:10.48550/arXiv.2011.09670.
36. Ye, X.; Xiong, F.; Lu, J.; Zhou, J.; Qian, Y. 3-Net: Feature Fusion and Filtration Network for Object Detection in Optical Remote Sensing Images. Remote Sensing 2020, 12, 4027.
37. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A Context-Aware Detection Network for Objects in Remote Sensing Imagery. 2019.
38. Xu, Z.; Zhang, W.; Zhang, T.; Li, J. remote sensing HRCNet: High-Resolution Context Extraction Network for Semantic Segmentation of Remote Sensing Images. Remote Sensing 2020, 13.
39. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the ICLR, 2016.
40. 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. IEEE 2018.
41. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship Rotated Bounding Box Space for Ship Extraction From High-Resolution Optical Satellite Images With Complex Backgrounds. IEEE Geoscience & Remote Sensing Letters 2017, 13, 1074-1078.
42. Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision 2010, 88.
43. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. Advances in neural information processing systems 2012, 25.
44. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J. MMDetection: Open MMLab Detection Toolbox and Benchmark. 2019.
45. Jin, Z.; Yu, D.; Song, L.; Yuan, Z.; Yu, L. You Should Look at All Objects. 2022, arXiv:2207.07889, doi:10.48550/arXiv.2207.07889.