Polygon-to-Polygon Distance Loss for Rotated Object Detection

Yang Yang, Jifeng Chen, Xiaopin Zhong*, Yuanlong Deng

Lab. of Machine Vision and Inspection, College of Mechatronics and Control Engineering, Shenzhen University, China
1910294008@email.szu.edu.cn, chenjifeng2020@email.szu.edu.cn
xzhong@szu.edu.cn, dengyl@szu.edu.cn

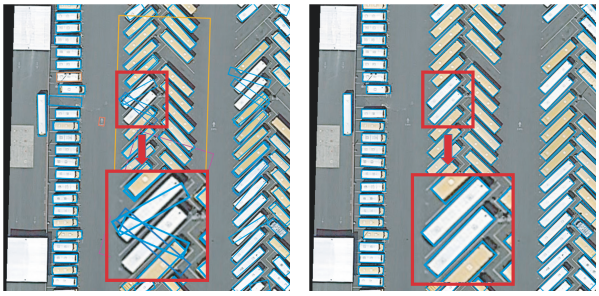
Abstract

There are two key issues that limit further improvements in the performance of existing rotational detectors: 1) Periodic sudden change of the parameters in the rotating bounding box (RBBox) definition causes a numerical discontinuity in the loss (such as smooth$_{L_1}$ loss). 2) There is a gap of optimization asynchrony between the loss in the RBBox regression and evaluation metrics. In this paper, we define a new distance formulation between two convex polygons describing the overlapping degree and non-overlapping degree. Based on this smooth distance, we propose a loss called Polygon-to-Polygon distance loss (P2P Loss). The distance is derived from the area sum of triangles specified by the vertexes of one polygon and the edges of the other. Therefore, the P2P Loss is continuous, differentiable, and inherently free from any RBBox definition. Our P2P Loss is not only consistent with the detection metrics but also able to measure how far, as well as how similar, a RBBox is from another one even when they are completely non-overlapping. These features allow the RetinaNet using the P2P Loss to achieve 79.15% mAP on the DOTA dataset, which is quite competitive compared with many state-of-the-art rotated object detectors.

Introduction

Object detection is a hot topic in computer vision, and horizontal object detectors have achieved promising results in both academic research and industrial applications. However, they are still inadequate because there are oriented and densely packed objects in many cases, such as object detection in aerial images and text recognition in natural scene images. To address this problem, researchers usually incorporate angular parameters in the regression branch as a way to develop rotated object detectors. They reported large improvements in public datasets, such as aerial image datasets DOTA (Xia et al. 2018), DIOR (Li et al. 2020b), HRSC2016 (Liu et al. 2017), scene text datasets ICDAR2015 (Karatzas et al. 2015), ICDAR2017 (Gomez et al. 2017) and face dataset FDDB (Jain and Learned-Miller 2010).

Compared to conventional horizontal object detectors, there emerge at least two issues when including angular parameters. They can be summarized as follows. First, periodic sudden change of the parameters in the rotating bounding box (RBBox) definition causes a numerical discontinuity in the loss. Second, there is a gap of optimization asynchrony between the loss in the RBBox regression and evaluation metrics. We propose a Polygon-to-Polygon distance loss, which can solve the above problems without any cost to obtain better performance of the rotated object detector.



(a) RetinaNet(smooth$_{L_1}$) (b) RetinaNet(P2P Loss)

Figure 1: Visualization comparison of smooth$_{L_1}$ and P2P Loss under the same model. The red box indicates the detrimental effect of boundary discontinuity and gap between loss and metrics on the detector.

The main contributions of this paper are as follows:

- We define a Polygon-to-Polygon distance that can be used to describe the spatial distance and morphological similarity of two convex polygons.
- We proposed a novel Polygon-to-Polygon distance loss (P2P Loss) with differentiable, numerically continuous properties. P2P Loss is free from the impact of periodic sudden change of the RBBox parameters. It is accordingly effective for any bounding box definition. When using our P2P Loss, in other words, a model with different box definitions yields almost the same performance.
- We executed extensive experiments on the commonly-used DOTA and HRSC2016 dataset to demonstrate the performance of P2P Loss and compared it with other key detectors. RetinaNet with P2P Loss can achieve 79.155% mAP on the DOTA dataset, which is preferable over the state-of-the-art rotation detectors.

*indicates the corresponding author.

Related Work

Horizontal Object Detectors

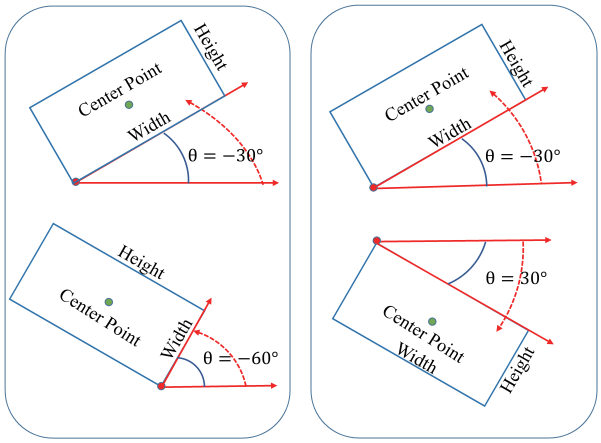
Horizontal object detectors can be classified into two types: two-stage detectors represented by RCNN (Girshick et al. 2014), fast R-CNN (Girshick 2015), faster R-CNN (Ren et al. 2015), and single-stage detectors represented by YOLO series (Redmon et al. 2016; Redmon and Farhadi 2017; Farhadi and Redmon 2018), SSD (Liu et al. 2016), RetinaNet (Lin et al. 2017), FCOS (Tian et al. 2019), etc. In the field of general object detection, the objects are by default considered to be horizontally axis-aligned. However, their performance can be significantly reduced because of the existence of dense adjacent objects in specific areas such as aerial imagery, text recognition, etc. Although some post-processing works, such as soft-NMS (Bodla et al. 2017) and softer-NMS (He et al. 2018), have attempted to address the dense distribution challenge, they are yet inadequate to cope with arbitrarily-oriented objects.

Rotated Object Detectors

Due to the shortage of horizontal detectors, researchers use RBBox to fit objects with arbitrary orientations. Compared with the natural-scene text detection, it is more difficult to deal with aerial images for the sake of the wide distribution of categories, spatial scales, compactness, rotational and translational positions of object instances. To obtain a better rotation detector, according to our investigation, the contributions of the published works so far are focused on the following three aspects.

Novel Feature Extraction Module Feature alignment modules, which are used for feature reconstruction and feature-object alignment, are frequently used in such work, prominently RRPN (Ma et al. 2018), ROI Transformer (Ding et al. 2019), S^2A-Net (Han et al. 2021a), and SCRDet (Yang et al. 2019). Background noise suppression and feature enhancement is also an important research direction, which is implemented in FADet (Li et al. 2019), SCRDet (Yang et al. 2019) and CASCADE-FF (Hou et al. 2020) using the attention mechanism. In SCRDet++ (Yang et al. 2020a), high-performance image-level and instance-level denoising modules are innovatively used. They all report significant performance improvements. Meanwhile, multi-level context information and multi-scale features, which are the key points to challenge the problem of wide distribution in spatial scales, are fully exploited in F^3-Net (Ye et al. 2020) and FFA (Fu et al. 2020).

Novel Regression Branch P-RSDet (Zhou et al. 2020), PolarDet (Zhao et al. 2021) and HRPNet (He et al. 2020) defined the RBBox in polar coordinate instead of in Cartesian coordinate system. CSL (Yang and Yan 2020) and DCL (Yang et al. 2021a) transforms angular prediction from a regression problem to a classification task. Nevertheless, this results in an increase in computational complexity, which requires further trade-offs. Mask OBB (Wang et al. 2019a) and CenterMap (Wang et al. 2020) use a rough instance mask branch to accomplish the fetching of rotating boxes, but bring complicated post-processing.



(a) D_{oc} (b) D_{le}

Figure 2: Bounding box definition. According to the D_{oc}, the parameter $\theta \in [-90°, 0)$ represents the acute angle between the positive half of x-axis and one side of the RBBox. This side is defined as width and the side perpendicular to the width as height. In the D_{le}, the longer side of the RBBox is defined as width and the shorter one as height. The angle between the width and the positive half of x-axis is θ.

Novel Loss Functions Such methods hope to find solutions to the problems posed by the L_n-norm loss in the training of rotation detectors. RSDet (Qian et al. 2021) proposed a modulated loss to smooth the periodic sudden change of the parameters in the L_n-norm loss, but ignoring the gap between the loss and metrics. Other ideas try to use an approximate Intersection over Union (IoU) loss as a way of avoiding the case where the rotating IoU is not differentiable. PIoU (Chen et al. 2020) estimates Pixels-IoU by counting the number of pixels inside the RBBox. The work of Yang Xue et al. (Yang et al. 2019, 2020a) combines IoU and smooth_{L_1} loss, but smooth_{L_1} contributes the main gradient direction. In addition, GWD (Yang et al. 2021c) approximated the RBBox as a two-dimensional Gaussian distribution and settled the non-differentiable behavior by the Gaussian Wasserstein Distance. They all offer some improvements over traditional methods, but further improvements can still be made.

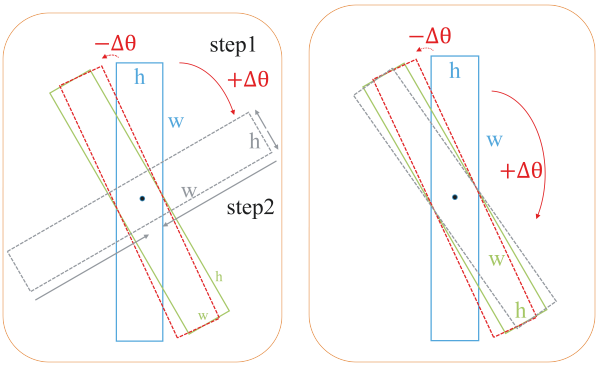
This paper focuses on the perspective of the loss function to improve detectors for rotated objects. In this case, we can transfer the outstanding algorithm for horizontal object detection to the rotational object detection task without making major changes in the network structure and without extra computational cost.

Problem of Loss Function

The defects of loss function used in rotated object detection are discussed in this section.

RBBox Definition

There are two popular RBBox definitions: the definition of OpenCV (D_{oc}) and Long-Edge (D_{le}). As shown in Figure 2, both of them have the defect caused by angular periodicity. In



(a) Case of D_{oc} (b) Case of D_{le}

Figure 3: Boundary discontinuity. The green box and the blue one represent the ground truth box and an anchor box respectively, while the red box is the ideal predicted box and the gray one is the actual predicted box.

particular, in the D_{oc}, a width-height swap occurs when the angle reaches the domain boundary. In the D_{le}, it is difficult for the model to determine which side is the longer one when the RBBox is close to a square. These introduce ambiguity into model training. To use the D_{oc} or the D_{le}, this is a choice dilemma for the pipeline design of most of the existing works. D_{oc} is used in ROPDet (Yang et al. 2020b) and SARD (Wang et al. 2019b) to prevent the problem of width-height illegibility, while D_{le} is adopted for RoI Transformer (Ding et al. 2019) and S^2A-Net (Han et al. 2021a) to avoid width-height swap.

Loss Discontinuity

In the early research on rotated object detection, researchers directly used the loss function from horizontal object detectors, such as smooth_{L_1}. Smooth_{L_1} suffers from numerical discontinuities where the angle parameter crosses through the domain boundary in both cases of using D_{oc} and D_{le}.

For example, in Figure 3(b), the ground truth and anchor box are set to $(0, 0, 60, 10, 89°)$ and $(0, 0, 60, 10, -90°)$. In practice, the difference between the ground truth box and the anchor box is small (this difference is enlarged in the figure for demonstration purposes). But the optimization is to rotate the anchor box clockwise to the gray predicted box. This is because, when the anchor box is rotated by the same angle $\Delta\theta$, the counterclockwise rotation yields predicted box $(0, 0, 60, 10, -90° - \Delta\theta)$ and clockwise rotation yields predicted box $(0, 0, 60, 10, -90° + \Delta\theta)$. The latter has a smaller loss for the ground truth box $(0, 0, 60, 10, 89°)$ than the former. The model prefers to use the latter optimization path increasing the difficulty of training.

In the D_{oc}, this inconsistency becomes more prominent because of the width-height swapping problem. For example, the blue anchor box $(0, 0, 60, 10, -90°)$ and the green ground truth box $(0, 0, 10, 60, -1°)$ are shown in Figure 3(a). The optimization process is divided into two steps. In the first step, the anchor box is rotated to the pose of the gray predicted one that is perpendicular to the ground truth box. Then the gray

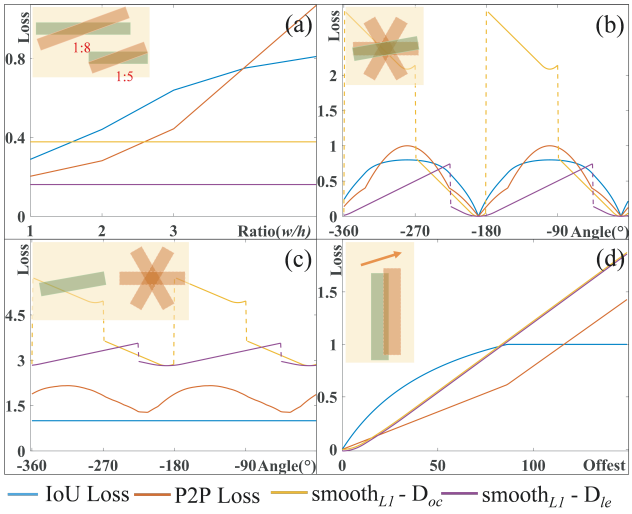


Figure 4: Comparison of different loss. (a) Relation between box ratio and loss. (b) Relation between rotated angle and loss while overlapping. (c) Relation between rotated angle while non-overlapping. (d) Relation between center shifting and loss.

predicted box decreases the width and increases the height to approach the ground truth box.

The discontinuity problem of the loss function appears only when the angle parameter θ moves across the domain boundary, but for limits further performance improvement. The change curve of the loss in Figure 4(b) and 4(c) clearly demonstrates this problem. The failed sample in Figure 1 is also difficult to optimize because of these ambiguities.

Gap Between Metrics and Loss

The head of all object detectors outputs scale-invariant parameterizations like Eq. 9. When using smooth_{L_1}, it is required that all output is properly regularized. However, parameters in different spaces, such as position space, size space, and angle space, are difficult to encode with a particular procedure. The gap between the loss function and the evaluation metrics is thus generated.

To eliminate the gap, in horizontal object detection, many studies have used IoU-based losses and demonstrated that they eliminate gaps and gain performance (Yu et al. 2016; Rezatofighi et al. 2019). But the IoU loss of RBBox is such a non-differentiable loss and cannot be backpropagated, leading to the widespread use of smooth_{L_1}.

P2P Loss was proposed to solve these problems. It is similar to IoU loss when two rotating rectangles intersect and remains optimized when non-overlapping. At the same time, the problem of loss discontinuity can be solved, and the performance is similar regardless of the box definition we use.

The Proposed Method

There are various ways to define the distance between two polygons, such as Euclidean distance, Hausdorff distance (Henrikson 1999) and so on. But these definitions can only

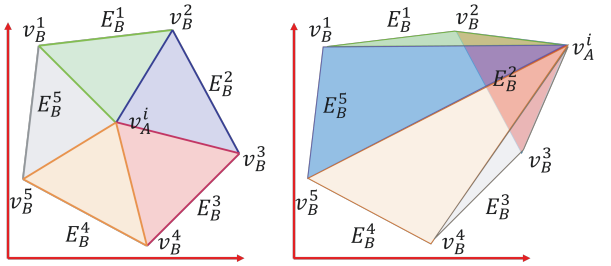


Figure 5: Definition of Vertex-to-Polygon Area. If v_A^i is inside or on one side of polygon B^m, B^m can be divided into m triangles and $S_{VP}\left(v_A^i,\ E_B\right) = S_B$, where S_B is the area of the convex polygon B^m. If v_A^i is outside the polygon B^m, on the contrary, $S_{VP}\left(v_A^i,\ E_B\right) > S_B$. The farther v_A^i is apart from B^m, the larger the $S_{VP}\left(v_A^i,\ E_B\right)$ is. What's more, $S_{VP}\left(v_A^i,\ E_B\right)$ varies continuously with the movement of v_A^i.

measure the distance of their spatial location but not their morphological similarity.

We define an n-vertex convex polygon $A^n = \{V_A, E_A\}$ with vertex set $V_A = \{v_A^1, v_A^2, v_A^3, \cdots, v_A^n\}$ and edge set $E_A = \{E_A^1, E_A^2, E_A^3 \ldots E_A^n\}$. In addition, we denote by S_A the area of convex polygon A^n. We say that $A^n = B^n$ if the vertex set and the edge set are respectively equal in two polygons $A^n = \{V_A, E_A\}$ and $B^n = \{V_B, E_B\}$.

Polygon-to-Polygon Distance

Definition 1: Vertex-to-Polygon Area As shown in Figure 5, we define the area of a vertex $v_A^i \in V_A$ of polygon $A^n = \{V_A, E_A\}$ to polygon $B^m = \{V_B, E_B\}$ as

$$S_{VP}\left(v_A^i, E_B\right) = \sum_{j=1}^{m} S_\Delta\left(v_A^i, E_B^j\right), \tag{1}$$

where $S_\Delta\left(v_A^i, E_B^j\right)$ is the area of the triangle enclosed by the vertex $v_A^i = (x_a, y_a)$ and the edge E_B^j, which can be calculated by the cross product.

Definition 2: Polygon-to-Polygon Area We go one step further and define the area of polygon $A^n = \{V_A,\ E_A\}$ to polygon $B^m = \{V_B,\ E_B\}$ as the area sum of all vertices of A^n to polygon B^m, i.e.,

$$\begin{aligned} S_{PP}\left(A^n, B^m\right) &= \sum_{i=1}^{n} S_{VP}\left(v_A^i,\ E_B\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} S_\Delta\left(v_A^i, E_B^j\right), \end{aligned} \tag{2}$$

likewise for the area of the polygon $B^m = \{V_B,\ E_B\}$ to polygon $A^n = \{V_A,\ E_A\}$.

In general, $S_{PP}\left(A^n, B^m\right) \neq S_{PP}\left(B^m, A^n\right)$. $S_{PP}\left(A^n, B^m\right) = nS_B$ when all of the vertices in V_A lie inside B^m or on the sides of B^m, otherwise $S_{PP}\left(A^n, B^m\right) > nS_B$. Likewise, $S_{PP}\left(B^m, A^n\right) = mS_A$ when the vertices in V_B lie inside A^n or on the sides of A^n, otherwise $S_{PP}\left(B^m, A^n\right) > mS_A$.

Definition 3: Polygon-to-Polygon Distance The distance between polygon $A^n = \{V_A,\ E_A\}$ and polygon $B^m = \{V_B,\ E_B\}$ is formulated as

$$\begin{aligned} d\left(A^n, B^m\right) = &\left(\frac{1}{n} S_{PP}\left(A^n, B^m\right) - S_B\right) \\ &+ \left(\frac{1}{m} S_{PP}\left(B^m, A^n\right) - S_A\right). \end{aligned} \tag{3}$$

Based on Definition 1 and 2, this distance is computed from multiple Vertex-to-Polygon areas, which can capture the spatial movement and shape variation of polygons in a continuous manner. On the basis of the definitions above, we have the following proposition.

Proposition Given two n-vertex convex polygons A^n and B^n, $d\left(A^n, B^n\right)$ is smallest and equal to 0 if and only if $B^n = A^n$. In other words, the B^n that can minimize $d\left(A, B\right)$ is

$$B^* = \arg\min_{B^n} d\left(A^n, B^n\right) = A^n. \tag{4}$$

Proof According to the analysis of definition 2, we know $S_{PP}\left(A^n, B^n\right) \geq \mathrm{nS}_B$, and $S_{PP}\left(A^n, B^n\right) \geq \mathrm{nS}_A$, so the Polygon-to-Polygon distance is non-negative symmetric, i.e., $d\left(A^n, B^n\right) = d\left(B^n, A^n\right) \geq 0$. All possible relative positions between A^n, B^n are discussed as follows.

Case 1 Non-overlapping The intersection of two polygons is empty, and each vertex of one polygon is outside the other polygon. In this case, $S_{PP}\left(A^n, B^n\right) > nS_B$, $S_{PP}\left(B^n, A^n\right) > nS_A$, and $d\left(A^n, B^n\right) = \left(\frac{1}{n} S_{PP}\left(A^n, B^n\right) - S_B\right) + \left(\frac{1}{n} S_{PP}\left(B^n, A^n\right) - S_A\right) > 0$.

Case 2 Partially-overlapping Two polygons partially intersect, then $S_{PP}\left(A^n, B^n\right) > nS_B$, $S_{PP}\left(B^n, A^n\right) > nS_A$. Finally, $d\left(A^n, B^n\right) = \left(\frac{1}{n} S_{PP}\left(A^n, B^n\right) - S_B\right) + \left(\frac{1}{n} S_{PP}\left(B^n, A^n\right) - S_A\right) > 0$.

Case 3 One Completely Enclosing the Other In this regard, one polygon is completely enclosed in the other one and $B^n \neq A^n$. If B^n is enclosed in A^n, $S_{PP}\left(A^n, B^n\right) > nS_B$, $S_{PP}\left(B^n, A^n\right) = nS_A$. If A^n is enclosed in B^n, on the contrary, $S_{PP}\left(A^n, B^n\right) = nS_B$, $S_{PP}\left(B^n, A^n\right) > nS_A$. Above all, $d\left(A^n, B^n\right) = \left(\frac{1}{n} S_{PP}\left(A^n, B^n\right) - S_B\right) + \left(\frac{1}{n} S_{PP}\left(B^n, A^n\right) - S_A\right) > 0$.

Case 4 Enclosing Each Other If the vertices of B^n are one-by-one moved to the those of A^n, then $B^n = A^n$. In this respect, $S_{PP}\left(A^n, B^n\right) = nS_B$, $S_{PP}\left(B^n, A^n\right) = nS_A$. Then $d\left(A^n, B^n\right) = 0$.

Above all, if and only if $B^n = A^n$, $d\left(A^n, B^n\right)$ is minimized to be zero.

Polygon-to-Polygon Distance Loss

With the proposition above, we can declare that the regression between two RBBoxes can be transformed into the optimization of Polygon-to-Polygon distance. The basic P2P Loss between RBBox and is defined as

$$L'_{P2P} = d\left(A^4, B^4\right)/(S_A + S_B). \tag{5}$$

When used individually, in practice, L'_{P2P} is sensitive to some outliers, which including box with error labeling and error gradient caused by underflow during calculation. Thus we have added two additional terms to refine it.

Additional Center Loss

$$L_c = \sum_{i\in\{x,y\}} \text{smooth}_{L_1}\left(t_i, t_i^*\right), \tag{6}$$

where smooth_{L_1} is defined in (Girshick 2015). t and t^* can be found in Eq. 9 for details.

Additional Semiperimeter Loss

$$L_{sp} = \text{smooth}_{L_1}\left(\left(w_A + h_A\right)/\left(w_B + h_B\right), 1\right), \tag{7}$$

where w_A and h_A are the width and height of A^4 (likewise for w_B and h_B), respectively. Semiperimeter is not affected by the width-height swap issue. That is, our box regression loss is defined as

$$L_{P2P} = \alpha L_c + \beta L_{sp} + \gamma L'_{P2P}, \tag{8}$$

where α, β, γ are the trade-off hyperparameters, which are set to $\{1,1,1\}$ by default.

Box Coder and Overall Loss Function

Like in most of the other works, the regression branch of our model finally outputs scale-invariant parameterizations tuple

$$t^* = (t_x^*, t_y^*, t_w^*, t_h^*, t_\theta^*). \tag{9}$$

We refer to the forward computation process of obtaining the corresponding ground truth t as box encode and the inverse computation as box decode. They are together known as the box coder. We tested two different box coders: the coder used in RRPN (Ma et al. 2018) is called Normal Coder and proposed along with RoI Transformer (Ding et al. 2019) is called RTrans Coder.

In addition, The multi-task loss used in our model is defined as

$$L = \frac{1}{N_{pos}}\sum_{i=1}^{N} L_{cls}(p_i, p_i^*) + \frac{\lambda}{N_{pos}}\sum_{i=1}^{N} p_i^* L_{reg}(b_i, g_i), \tag{10}$$

where i is the index of RBBox, p_i yielded by the sigmoid function represents the predicted probability of the i-th box being an object. The ground truth category p_i^* is 0 if this box is negative and 1 if positive. b_i denotes the i-th predicted box, and g_i is the i-th box ground truth. N_{pos} denotes the normalizing hyper-parameters (positive sample size); and λ represents the balancing hyper-parameter between the two losses which is set to 1 in our experiments.

The Properties of P2P Loss

Free From Box Definition The five-parameter-defined RBBox is transformed into a vertex set before P2P Loss is computed. This makes our P2P Loss compatible to any box definitions. In other words, P2P Loss is naturally free from the choice of box definitions.

Continuous and Differentiable As shown in Figure 4, the numerical curve of P2P Loss is always continuous regardless of whether the RBBox varies from the ground truth with respect to spatial or morphological distance. Also, P2P Loss consists of simple arithmetic, but no indexing operations. So it is differentiable and can be easily backpropagated during training.

Consistent with Metrics Before calculating the P2P Loss, the output of the network must be transformed into a set of vertices to form a uniform space with vertex coordinates. On this basis, the P2P Loss actually obtained from the Polygon-to-Polygon area, which is similar to the IoU Loss in both value and definition, i.e., it has a high degree of consistency with the evaluation metrics (as shown in Fig. 4 (a), 4 (b), and 4 (d)). Furthermore, it can measure the spatial as well as the morphological discrepancy of two RBBox when they are non-overlapping, as shown in Figure 4 (c). This idea has frequently appeared in the IoU family of loss functions, such as GIoU (Rezatofighi et al. 2019), DIoU (Zheng et al. 2020), and CIoU (Zheng et al. 2021).

Experiments

Dataset

Aerial image dataset DOTA (Xia et al. 2018) consists of 2806 images ranging in size from 800×800 to 4000×4000 pixels and contains 188282 objects. The ratios of training set, validation set and test set are $1/2$, $1/6$, and $1/3$ respectively. We use the training set and validation set for training and the test set for testing. To facilitate the training process, we crop these images into small 1024×1024 patch with a stride of 824. Eventually, 21046 training images and 10833 testing images are obtained.

HRSC2016 (Liu et al. 2017) is a high-resolution ship detection data set with 436, 181 and 444 images for training, verification and testing, respectively. In training, we maintain the aspect ratio of the images in the training and validation sets and scale them to 800×512.

Baseline and Implementation Details

The RetinaNet (Lin et al. 2017) has a similar structure to most advanced algorithms and is chosen as the baseline for experiments by the vast majority of researchers. We use RetinaNet with ResNet50-FPN as the baseline model for all experiments to make the result comparison more reliable and to generalize our method to other studies more easily. L_{cls} and L_{reg} are Focal loss defined in (Lin et al. 2017) and smooth_{L_1}, respectively. We preset 3 anchors with aspect ratios of 0.5,1.0,2.0 and angle of 0° at each position of the pyramidal features at each level by default, unless otherwise specified. We use 4 GeForce RTX 3090 GPUs with a total of 8 images per mini-batch (2 images per GPU) for training and a single GeForce

Box def.	Box coder	Loss	mAP
D_{oc}	Normal	smooth$_{L_1}$	69.416
D_{le}	Normal	smooth$_{L_1}$	69.916
D_{le}	RTrans	smooth$_{L_1}$	69.880

Table 1: Results of baesline on DOTA dataset with different box definition and box coder.

Box def.	Box coder	P2P Loss			mAP
		α	β	γ	
D_{oc}	Normal				$*$
D_{le}	Normal	1	0	1	70.496
D_{le}	RTrans				$*$
D_{oc}	Normal				70.717
D_{le}	Normal	0	1	1	70.367
D_{le}	RTrans				70.530
D_{oc}	Normal				70.897
D_{le}	Normal	1	1	1	70.911
D_{le}	RTrans				71.055

Table 2: The effect of different hyper-parameter on P2P Loss. The results come from DOTA. $*$ indicates the possibility of random oscillatory non-convergence when training is repeated several times.

RTX 3090 GPU for inference. All experiments are trained using the Adam (Kingma and Ba 2014) optimizer with learning rate 0.0001. We use random flipping to avoid overfitting during training, and no other tricks if not specified.

Ablation Study

Baseline As shown in Table 1, our RetinaNet reimplementation achieves a higher mAP of 69.916% on DOTA, indicating that it is a solid baseline. Even if the model and training details are exactly the same, using different box definitions can yield vastly different results. This result is as expected (as described in Section **Problem of Loss Function**). When using D_{oc}, the smooth$_{L_1}$ loss is affected by two unfavorable factors, the exchangeability of edges and the periodicity of angles, so the performance is 0.5% lower than that of D_{le}.

The Effect of Different Hyper-parameter on P2P Loss Comparison experimental results between different box definition, box coder and P2P Loss hyperparameters are shown in Table 2. The performance improves when the center and semiperimeter losses are added individually, and optimal performance is achieved when both are added simultaneously. Compared with the ordinary loss function, P2P Loss, which is more complex in form and in computation, is sensitive to outliers during training. Therefore, more constraints are needed for stable convergence. With P2P Loss, the model has an improvement of about 0.99% to 1.48% compared to the baseline in Table 1. The model using D_{oc} benefits the most, with a boost of about 1.48%.

In subsequent experiments, the P2P Loss follows the hyperparameter settings of $\alpha = 1$, $\beta = 1$, and $\gamma = 1$. In addition, D_{le} and RTrans coder will be used.

Loss	Epoch	Box def.	Box coder	mAP
IoU-Smooth$_{L_1}$ (Yang et al. 2020a)	40	D_{oc}	Normal	68.65
GWD (Yang et al. 2021c)	40	D_{oc}	Normal	69.92
P2P Loss	12	D_{oc}	Normal	70.90
		D_{le}	Normal	70.91
		D_{le}	RTrans	71.06

Table 3: Comparison using different loss functions on DOTA dataset. The results are obtained with the RetinaNet using ResNet50-FPN as the backbone.

RATSS	QFL	P2P Loss	mAP
✓			70.772
		✓	71.055
✓		✓	71.421
✓	✓	✓	72.200

Table 4: Ablation experiment of P2P Loss# on DOTA. The model and parameters of training are set as default.

Comparison with Other Loss Functions We compared P2P Loss with IoU-Smooth$_{L_1}$ and GWD that tried to solve the aforementioned problem. Table 3 shows that P2P Loss has better consistency and stability for different box definitions and different box coders when the performance exceeds them.

More Appropriate L_{cls} Many current rotation detectors define positive and negative samples based on a fixed iou threshold. However, after using P2P Loss, we cannot explicitly describe the optimal optimization process of the predicted box. In this case, a softer way of assigning positive and negative examples is needed. Horizontal object detector can get performance improvement by ATSS (Zhang et al. 2020). In this work, we adapt the rotation case to the original ATSS by limiting the positive samples' center to the rotation ground-truth box, called Rotation ATSS (RATSS). Finally, we then set L_{cls} to Quality Focal Loss (QFL) (Li et al. 2020c). For convenience, we later refer to the combination of RATSS, QFL and P2P Loss as P2P Loss#. As shown in Table 4, P2P Loss# was able to achieve 72.2% mAP in DOTA without additional data augmentation.

Influence of Anchor Quantity For smooth$_{L_1}$, it depends on the number of anchors and the initial angle. More anchors help to reduce the possibility of anchors at the boundary being selected as the corresponding positive case during the assignment phase. In Table 6, when using smooth$_{L_1}$, an inappropriate number of anchor results in a sharp drop in performance. When using P2P Loss#, the approximate performance can be achieved without pursuing an excessive number of anchors.

Free from Box Definition As shown in Table 2 and 7, using different box definitions and different box coders, all else being equal, has a negligible impact on the final performance. We can even consider them to be equal within the margin of

Method	Backb.	Epoch	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Two Stage																		
SCRDet (Yang et al. 2019)	R-101	–	90.0	80.7	52.1	68.4	68.4	60.3	72.4	90.9	87.9	86.9	65.0	66.7	66.3	68.2	65.2	72.6
GLS-Net (Li et al. 2020a)	R-101	–	88.7	77.4	51.2	71.0	73.3	72.2	84.7	90.9	80.4	85.4	58.3	62.3	67.6	70.7	60.4	73.0
R^4Det (Sun et al. 2020)	R-152	12	89.0	85.4	52.9	73.8	74.9	81.5	80.3	90.8	87.0	85.3	64.1	60.9	69.0	70.6	67.8	75.8
CSL (Yang and Yan 2020)	R-152	20	90.3	85.5	54.6	75.3	70.4	73.5	77.6	90.8	86.2	86.7	69.6	68.0	73.8	71.1	68.9	76.2
RSDet-II (Qian et al. 2021)	R-152	30	89.9	84.5	53.8	74.4	71.5	78.3	78.1	91.1	87.4	86.9	65.6	65.2	75.4	79.7	63.3	76.3
SCRDeT++ (Yang et al. 2020a)	R-101	40	90.1	84.4	55.4	74.0	77.5	71.1	86.1	90.7	87.3	87.1	69.6	68.9	73.7	71.3	65.1	76.8
ReDet (Han et al. 2021b)	ReR-50	12	88.8	82.5	60.8	80.8	78.3	86.1	88.3	90.9	87.8	87.0	68.7	66.9	79.3	79.7	74.7	**80.1**
Single Stage																		
RetinaNet-GWD (Yang et al. 2021c)	R-152	60	87.0	83.9	54.4	77.5	74.4	68.5	80.3	86.6	83.4	85.6	73.5	67.8	72.6	75.8	73.4	76.3
R^3Det (Yang et al. 2021b)	R-152	20	89.8	83.8	48.1	66.8	78.8	83.3	87.8	90.8	85.4	85.5	65.7	62.7	67.5	78.6	72.6	76.5
PolarDet (Zhao et al. 2021)	R-101	360	89.7	87.1	48.1	71.0	78.5	80.3	87.5	90.8	85.6	86.9	61.6	70.3	71.9	73.1	67.2	76.6
R^3Det-GWD (Yang et al. 2021c)	R-152	60	89.3	83.7	59.3	79.9	76.4	83.9	86.5	89.1	85.5	86.5	73.0	67.6	76.9	77.1	71.6	79.1
S^2A-Net (Han et al. 2021a)	R-101	12	89.3	84.1	57.0	79.2	80.2	82.9	89.2	90.9	84.7	87.6	71.7	68.2	78.6	78.2	65.6	**79.2**
RetinaNet-P2P Loss#†	R-50	12	89.0	77.0	48.5	69.9	79.4	80.1	88.1	90.9	85.0	85.5	60.0	62.9	71.9	66.2	56.4	74.0
RetinaNet-P2P Loss#	R-50	12	89.3	85.9	55.4	80.0	79.8	83.0	88.4	90.9	85.6	87.1	68.8	69.9	76.3	74.5	59.8	78.3
RetinaNet-P2P Loss#	R-101D	12	89.2	86.1	55.2	81.4	80.3	83.5	88.3	90.9	86.6	87.1	71.7	69.9	77.3	76.0	59.6	**79.2**

Table 5: Comparisons with the State-of-the-Art methods on DOTA. The corresponding relationship between abbreviations and full name: SBF-Soccer ball field, HC-Helicopter, SP-Swimming pool, RA-Roundabout, LV-Large vehicle, SV-Small vehicle, BR-Bridge, HA-Harbor, GTF-Ground track field, BC-Basketball court, TC-Tennis court, BD-Baseball diamond, ST-Storage tank, SP-Ship, PL-Plane. Data augmentation (random rotation, multi-scale) for training and testing of all models except those with † markers. R-101D indicates ResNet101 with DCN (Dai et al. 2017) added, same for R-50 and R-152.

Loss	Anchor⋆	mAP
smooth_{L_1}	3	73.512
smooth_{L_1}	6	83.532
smooth_{L_1}	9	82.696
P2P Loss#	3	89.177
P2P Loss#	6	88.675
P2P Loss#	9	89.188

Table 6: Comparison of smooth_{L_1} with P2P Loss# for different number of anchors in HRSC dataset. Anchor⋆ represents the number of anchors. The RetinaNet-FPN-R-50 is used and random rotation is applied to data augmentation.

error allowed. These experiments prove that P2P Loss is free from box definition.

Comparison of State-of-the-Art Methods The results are shown in Table 5, where the RetinaNet-P2P Loss# model with ResNet50 and ResNet101-DCN can achieve mAP 78.308% and 79.155% on DOTA dataset, respectively. This performance is also quite competitive among a bunch of State-of-the-Art models, and we only use RetinaNet as the base model without adding any parameters or operations.

Box def.	**Box coder**	**mAP**
D_{oc}	Normal	72.380
D_{le}	Normal	72.211
D_{le}	RTrans	72.200

Table 7: Performance consistency of different box coder and box definition after using P2P Loss# on DOTA.

Conclusion

In this paper, two important issues that impact the performance of rotating object detectors were discussed in detail. It is found that a continuous loss function that can measure the spatial and morphological distance between two polygons is significant to those issues. Therefore, P2P Loss is proposed in this paper. Extensive experiments on DOTA and HRSC datasets demonstrated the effectiveness of P2P Loss.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62171288, and the Shenzhen Science and Technology Program under Grant JCYJ20180305123922293, JCYJ20190808143415801. The authors would also like to thank the anonymous reviewers.

References

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, 5561–5569.

Chen, Z.; Chen, K.; Lin, W.; See, J.; Yu, H.; Ke, Y.; and Yang, C. 2020. PIoU Loss: Towards Accurate Oriented Object Detection in Complex Environments. In *European Conference on Computer Vision*, 195–211. Springer.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; and Lu, Q. 2019. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2849–2858.

Farhadi, A.; and Redmon, J. 2018. Yolov3: An incremental improvement. *Computer Vision and Pattern Recognition, cite as*.

Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; and Sun, X. 2020. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 161: 294–308.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587.

Gomez, R.; Shi, B.; Gomez, L.; Numann, L.; Veit, A.; Matas, J.; Belongie, S.; and Karatzas, D. 2017. Icdar2017 robust reading challenge on coco-text. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, 1435–1443. IEEE.

Han, J.; Ding, J.; Li, J.; and Xia, G.-S. 2021a. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Han, J.; Ding, J.; Xue, N.; and Xia, G.-S. 2021b. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2786–2795.

He, X.; Ma, S.; He, L.; and Ru, L. 2020. High-Resolution Polar Network for Object Detection in Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*.

He, Y.; Zhang, X.; Savvides, M.; and Kitani, K. 2018. Softer-nms: Rethinking bounding box regression for accurate object detection. *arXiv preprint arXiv:1809.08545*, 2: 3.

Henrikson, J. 1999. Completeness and total boundedness of the Hausdorff metric. *MIT Undergraduate Journal of Mathematics*, 1: 69–80.

Hou, L.; Lu, K.; Xue, J.; and Hao, L. 2020. Cascade detector with feature fusion for arbitrary-oriented objects in remote sensing images. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Jain, V.; and Learned-Miller, E. 2010. Fddb: A benchmark for face detection in unconstrained settings. Technical report.

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 1156–1160. IEEE.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, C.; Luo, B.; Hong, H.; Su, X.; Wang, Y.; Liu, J.; Wang, C.; Zhang, J.; and Wei, L. 2020a. Object detection based on global-local saliency constraint in aerial images. *Remote Sensing*, 12(9): 1435.

Li, C.; Xu, C.; Cui, Z.; Wang, D.; Zhang, T.; and Yang, J. 2019. Feature-attentioned object detection in remote sensing imagery. In *2019 IEEE International Conference on Image Processing (ICIP)*, 3886–3890. IEEE.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020b. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159: 296–307.

Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020c. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. *Advances in Neural Information Processing Systems*, 33: 21002–21012.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Liu, Z.; Yuan, L.; Weng, L.; and Yang, Y. 2017. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods*, volume 2, 324–331. SCITEPRESS.

Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; and Xue, X. 2018. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11): 3111–3122.

Qian, W.; Yang, X.; Peng, S.; Yan, J.; and Guo, Y. 2021. Learning Modulated Loss for Rotated Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2458–2466.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 658–666.

Sun, P.; Zheng, Y.; Zhou, Z.; Xu, W.; and Ren, Q. 2020. R4 Det: Refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images. *Image and Vision Computing*, 103: 104036.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9627–9636.

Wang, J.; Ding, J.; Guo, H.; Cheng, W.; Pan, T.; and Yang, W. 2019a. Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sensing*, 11(24): 2930.

Wang, J.; Yang, W.; Li, H.-C.; Zhang, H.; and Xia, G.-S. 2020. Learning center probability map for detecting objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5): 4307–4323.

Wang, Y.; Zhang, Y.; Zhang, Y.; Zhao, L.; Sun, X.; and Guo, Z. 2019b. SARD: Towards scale-aware rotated object detection in aerial imagery. *IEEE Access*, 7: 173855–173865.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, X.; Hou, L.; Zhou, Y.; Wang, W.; and Yan, J. 2021a. Dense label encoding for boundary discontinuity free rotation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15819–15829.

Yang, X.; and Yan, J. 2020. Arbitrary-oriented object detection with circular smooth label. In *European Conference on Computer Vision*, 677–694. Springer.

Yang, X.; Yan, J.; Feng, Z.; and He, T. 2021b. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3163–3171.

Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; and Tian, Q. 2021c. Rethinking rotated object detection with gaussian wasserstein distance loss. *arXiv preprint arXiv:2101.11952*.

Yang, X.; Yan, J.; Yang, X.; Tang, J.; Liao, W.; and He, T. 2020a. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *arXiv preprint arXiv:2004.13316*.

Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; and Fu, K. 2019. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8232–8241.

Yang, Z.; He, K.; Zou, F.; Cao, W.; Jia, X.; Li, K.; and Jiang, C. 2020b. ROPDet: real-time anchor-free detector based on point set representation for rotating object. *Journal of Real-Time Image Processing*, 17(6): 2127–2138.

Ye, X.; Xiong, F.; Lu, J.; Zhou, J.; and Qian, Y. 2020. F3-Net: Feature Fusion and Filtration Network for Object Detection in Optical Remote Sensing Images. *Remote Sensing*, 12(24): 4027.

Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; and Huang, T. 2016. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, 516–520.

Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; and Li, S. Z. 2020. 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*, 9759–9768.

Zhao, P.; Qu, Z.; Bu, Y.; Tan, W.; and Guan, Q. 2021. Polardet: A fast, more precise detector for rotated target in aerial images. *International Journal of Remote Sensing*, 42(15): 5821–5851.

Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12993–13000.

Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; and Zuo, W. 2021. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Transactions on Cybernetics*.

Zhou, L.; Wei, H.; Li, H.; Zhao, W.; Zhang, Y.; and Zhang, Y. 2020. Arbitrary-oriented object detection in remote sensing images based on polar coordinates. *IEEE Access*, 8: 223373–223384.