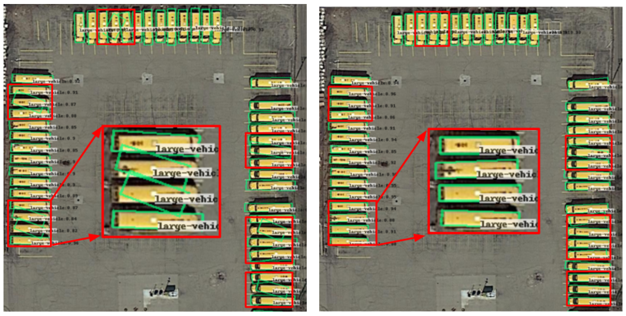
The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

Learning Modulated Loss for Rotated Object Detection

Wen Qian [1,2], Xue Yang[3,4], Silong Peng[1,2,*], Junchi Yan[3,4], Yue Guo[1,2]
[1]Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[4]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{qianwen2018, silong.peng, guoyue2013}@ia.ac.cn, {yangxue-2019-sjtu, yanjunchi}@sjtu.edu.cn

Abstract

Popular rotated detection methods usually use five parameters (coordinates of the central point, width, height, and rotation angle) or eight parameters (coordinates of four vertices) to describe the rotated bounding box and ℓ_1 loss as the loss function. In this paper, we argue that the aforementioned integration can cause training instability and performance degeneration. The main reason is the discontinuity of loss which is caused by the contradiction between the definition of the rotated bounding box and the loss function. We refer to the above issues as rotation sensitivity error (RSE) and propose a modulated rotation loss to dismiss the discontinuity of loss. The modulated rotation loss can achieve consistent improvement on the five parameter methods and the eight parameter methods. Experimental results using one stage and two stages detectors demonstrate the effectiveness of our loss. The integrated network achieves competitive performances on several benchmarks including DOTA and UCAS AOD. The code is available at https://github.com/yangxue0827/RotationDetection.



(a) RetinaNet-H (baseline) (b) The proposed RSDet

Figure 1: Detection results before and after solving the RSE problem with RSDet. The red rectangles in (a) represent failed examples due to the discontinuity of loss.

Introduction

Object detection approaches can generally be divided into horizontal object detectors and rotated object detectors according to the description of the bounding box. Specifically, horizontal detectors, by which all the bounding boxes are set in the horizontal direction, are often more suitable for general natural scene images such as COCO (Lin et al. 2014) and Pascal VOC (Everingham et al. 2010). In contrast, more accurate detection is often needed on occasions such as scene text, aerial imagery, face, and license plate. Until now, many rotated object detection benchmarks such as aerial dataset (DOTA (Xia et al. 2018), DIOR (Li et al. 2019b), HRSC2016 (Liu et al. 2017)), scene text dataset (ICDAR2015 (Karatzas et al. 2015), ICDAR2017 (Gomez et al. 2017)) have been published. The existing region-based rotated object detectors usually regress five parameters (Yang et al. 2019b, 2018b; Jiang et al. 2017; Ma et al. 2018) or eight parameters (Xu et al. 2020; Liao et al. 2018; Zhou et al. 2017; Zhang et al. 2019; He et al. 2017) to describe rotated bounding boxes and use ℓ_1-loss as loss functions. However, these two kinds of detectors both suffer from the loss discontinuity.

Firstly, the discontinuity of loss in five-parameter methods is mainly caused by the angle parameter. The loss value jumps when the angle reaches its range boundary, as shown in Fig. 2: a horizontal rectangle is respectively rotated one degree clockwise and counterclockwise to get the ground truth and the prediction box. The location of the reference rectangle has only been slightly moved, but its angle changes a lot due to the angular periodicity. Moreover, the height and width are also exchanged according to the five-parameter definition method commonly used by OpenCV. Moreover, in the five-parameter system, parameters i.e. angle, width, height, and center point have different measurement units, and show rather different relations against the Intersection over Union (IoU) (see Fig. 5). Simply adding them up for inconsistent regression can hurt performance.

Secondly, the discontinuity of loss also exists in the eight-parameter methods, although parameters in this method denote coordinate value with no ambiguity. The discontinuity of loss in eight-parameter can be seen in Fig. 2: Starting from the most left corner, points are clockwise defined. then we can get $a \rightarrow b \rightarrow c \rightarrow d$ and $d \rightarrow a \rightarrow b \rightarrow c$ to describe the red rectangle and green rectangle respectively. If the loss is calculated directly based on the corresponding order of the points, the loss is huge when the ground-truth

[*]Corresponding author is Silong Peng.

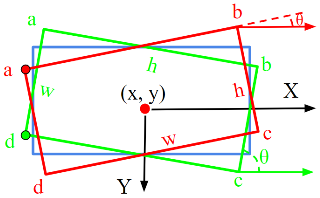


Figure 2: Loss discontinuity: rectangles in blue, red, and green respectively denote reference box, ground truth, and prediction. Here the reference box is rotated one degree clockwise to get the ground truth and is rotated similarly counterclockwise to obtain the prediction. Then the three boxes are described with five parameters: reference (0, 0, 10, 25, -90°), ground truth (0, 0, 25, 10, -1°), and prediction (0, 0, 10, 25, -89°). Here ℓ_1 loss is far more than 0.

and prediction largely overlap with each other.

Loss discontinuity exists in both the five-parameter methods and eight-parameter methods. We call this phenomenon as rotation sensitivity error (RSE), which can lead to training instability (see Fig. 7). In order to address the discontinuity of loss, a modulated rotation loss ℓ_{mr} is devised to carefully handle the boundary constraints for rotation, leading to a smoother loss curve during training. In other words, we add a correction term to the original loss and take the minimum value of the original loss and correction term. This correction is particularly large than ℓ_1-loss when it does not reach the range boundary of the angle. However, this correction becomes normal when ℓ_1-loss is abrupt. In other words, such correction can be seen as the symmetry of ℓ_1-loss about the location of the mutation. Finally, ℓ_{mr} takes the minimum of ℓ_1-loss and the correction, and the curve of ℓ_{mr} is continuous.

To verify the generalization performance of ℓ_{mr}, we design different experimental frameworks based on one-stage and two-stage methods and collectively referred to these frameworks as RSDet. RSDet shows state-of-art performance on the DOTA benchmark and UCAS-AOD benchmark, and our techniques are all orthogonal to existing methods. **The contributions of this paper are:**

i) We formally formulate the important while relatively ignored rotation sensitivity error (RSE) for region-based rotation detectors, which refers to the loss discontinuity.

ii) For the traditionally widely used five-parameter system and eight-parameter system, we devise a special treatment to ensure the loss continuity. The new loss is termed by ℓ_{mr}.

iii) Based on ℓ_{mr}, we respectively extend it to the one-stage and two-stage detection frameworks, which show state-of-the-art performance on DOTA and UCAS-AOD benchmarks.

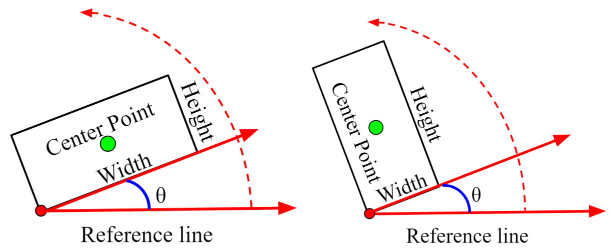
Related Work

Horizontal Object Detectors Visual object detection has been a hot topic over the decades. Since the seminal work R-CNN (Girshick et al. 2014), there have been a series of improvements including Fast RCNN (Girshick 2015), Faster RCNN (Ren et al. 2017), and R-FCN (Dai et al. 2016), which fall the category of the two-stage methods. On the other hand, single-stage approaches have also been well developed which can be more efficient than the two-stage methods. Examples include Overfeat (Sermanet et al. 2014), YOLO (Redmon et al. 2016), and SSD (Liu et al. 2016). In particular, SSD (Liu et al. 2016) combines advantages of Faster RCNN and YOLO to achieve the trade-off between speed and accuracy. Subsequently, multi-scale feature fusion techniques are widely adopted in both single-stage methods and two-stage ones, such as FPN (Lin et al. 2017a), RetinaNet (Lin et al. 2017b), and DSSD (Fu et al. 2017). Recently, many cascaded or refined detectors are proposed. For example, Cascade RCNN (Cai and Vasconcelos 2018), HTC (Chen et al. 2019), and FSCascade (Zhang et al. 2018) perform multiple classifications and regressions in the second stage, leading to notable accuracy improvements in both localization and classification. Besides, the anchor free methods have become a new research focus, including FCOS (Tian et al. 2019), FoveaBox (Kong et al. 2019), and RepPoints (Yang et al. 2019d). Structures of these detectors are simplified by discarding anchors, so anchor-free methods have opened up a new direction for object detection.

However, the above detectors only generate horizontal bounding boxes, which limits their applicability in many real-world scenarios. In fact, in scene texts and aerial images, objects tend to be densely arranged and have large aspect ratios, which requires more accurate localization. Therefore, rotated object detection has become a prominent direction in recent studies (Yang et al. 2019d).

Rotated Object Detectors Rotated object detection has been widely used in natural scene text (Jiang et al. 2017; Ma et al. 2018), aerial imagery (Fu et al. 2018; Yang et al. 2018a, 2019a), etc. And these detectors typically use rotated bounding boxes to describe positions of objects, which are more accurate than those horizontal boxes. Represented by scene text, many excellent detectors have been proposed. For example, RRPN (Ma et al. 2018) uses rotating anchors to improve the qualities of region proposals. R^{2}CNN (Jiang et al. 2017) is a multi-tasking text detector that simultaneously detects rotated and horizontal bounding boxes. In TextBoxes++ (Liao, Shi, and Bai 2018), to accommodate the slenderness of the text, a long convolution kernel is used and the number of proposals is increased.

Moreover, object detection in aerial images is more difficult, and its main challenges are reflected in complex backgrounds, dense arrangements, and a high proportion of small objects. Many scholars have also applied general object detection algorithms to aerial images, and many robust rotated detectors have emerged in aerial images. For example, ICN (Azimi et al. 2018) combines various modules such as image pyramid, feature pyramid network, and deformable inception sub-networks, and it achieves satisfactory performances on DOTA benchmark. RoI Transformer (Ding et al. 2019) extracts rotation-invariant features for boosting subsequent classification and regression. SCRDet (Yang et al. 2019c) proposes an IoU-smooth ℓ_1 loss to solve the sudden loss change caused by the angular periodicity so that it can bet-



(a) Width is longer than height. (b) Height is longer than width.

Figure 3: The five-parameter definition in OpenCV exchanges the width and the height in the boundary condition for rotation. The angle parameter θ ranges from -90 degree to 0 degree, but it should be distinguished from another definition (Xia et al. 2018), with 180 degree angular range, whose θ is determined by long side of rectangle and x-axis.

ter handle small, cluttered, and rotated objects. R^{3}Det (Yang et al. 2019b) proposes an end-to-end refined single-stage rotated object detector for fast and accurate object localization by solving the feature misalignment problem.

All the above mentioned rotated object detectors do not take the inherent discontinuity of loss into account, which we show can damage learning stability and final detection performances in our experiments. However, no existing studies have addressed this fundamental problem that motivates our work.

Methodology

Overview. In this section, we firstly present two mainstream protocols for bounding box parameterization *i.e.* the five-parameter and eight-parameter models. Then we formally determine the loss discontinuity in the five-parameter and eight-parameter methods. We call such issues collectively as rotation sensitivity error (RSE) and propose a modulated rotation loss to achieve more smooth learning.

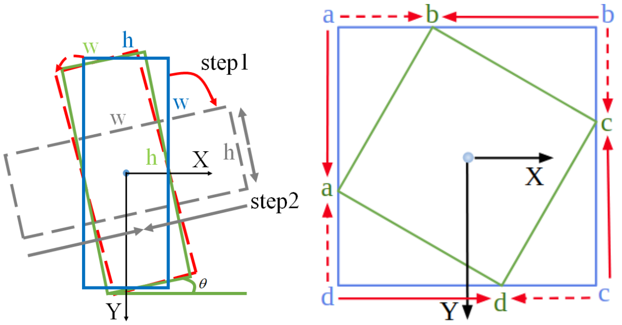
Parameterization of Rotated Bounding Box

Without loss of generality, our five-parameter definition is in line with that in OpenCV, as shown in Fig. 3: a) define the reference line along the horizontal direction on which the vertex with the smallest vertical coordinate is located. b) rotate the reference line counterclockwise, the first rectangular side being touched by the reference line is defined as width w regardless of its length compared with the other side – height h. c) the central point coordinate is (x, y) and the rotation angle is θ.

While the definition of eight-parameter is more simple: starting from the lower left corner, four clockwise vertices (a, b, c, d) of the rotated bounding box are used to describe its location, as shown in Fig. 2. In fact, such a parameterization protocol is convenient for quadrilateral description, which is friendly in more complex application scenarios.

Rotation Sensitivity Error

As mentioned earlier, rotation sensitivity error (RSE) exists in five-parameter and eight-parameter methods.



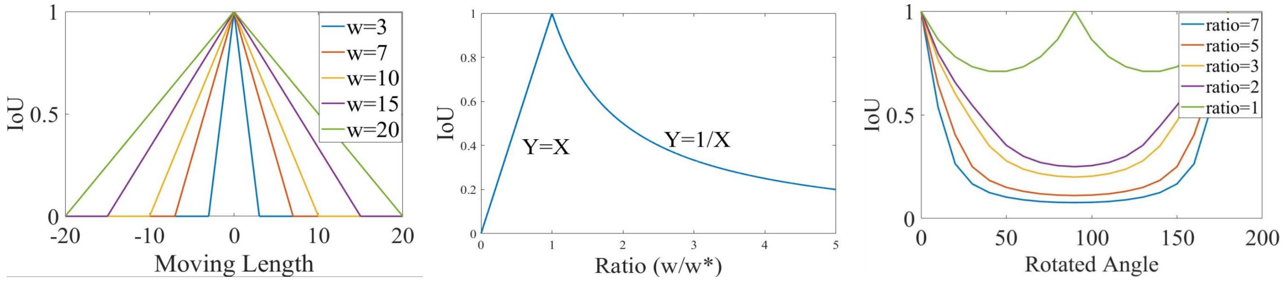
(a) 5-parameter regression w/ two steps in boundary condition (b) Eight-parameter regression procedure

Figure 4: Boundary discontinuity analysis of five-parameter regression and eight-parameter regression. The red solid arrow indicates the actual regression process, and the red dotted arrow indicates the ideal regression process.

RSE in Five-parameter Methods. Here, RSE is mainly caused by two reasons: i) The adoption of the angle parameter and the exchange between width and height contribute to the sudden loss change (increase) in the boundary case. ii) Regression inconsistency of measure units exists in the five-parameter model.

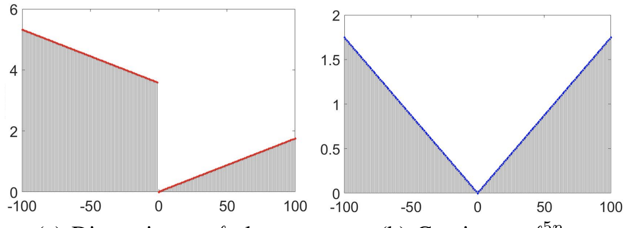
Loss Discontinuity. The angle parameter causes the loss discontinuity. To obtain the predicted box that coincides with the ground truth box, the horizontal reference box is rotated counterclockwise, as shown in Fig. 4. In this figure, the coordinates of the predicted box are transformed from those of the reference box $(0, 0, 100, 25, -90°)$ to $(0, 0, 100, 25, -100°)$ in the normal coordinate system. However, the angle of the predicted box is out of the defined range, and the coordinates of the ground truth box are $(0, 0, 25, 100, -10°)$. Despite the rotation is physically smooth, the loss will be quite large, which corresponds to the discontinuity of loss. To avoid such a loss fluctuation, the reference box need to be rotated clockwise to obtain the gray box $(0, 0, 100, 25, -10°)$ in Fig. 4(a) (step 1), then width and height of the gray box will be scaled to obtain the final predicted box $(0, 0, 25, 100, -10°)$ (step 2). At this time, although the loss value is close to zero, the detector experiences a complex regression. This requires relatively high robustness, which increases the training difficulty. More importantly, an explicit and specific way is lacked to achieve a smooth regression, which will be addressed in the subsequent part of the paper.

Regression Inconsistency. Different measurement units of five parameters make regression inconsistent. However, the impact of such artifacts is still unclear and has been rarely studied in the literature. Relationships among all the parameters and IoU are empirically studied in Fig. 5. Specifically, the relationship between IoU and width (height) is a combination of a linear function and inverse proportion function, as illustrated in Fig. 5(a). The relationship between the central point and IoU is a symmetric linear function, as illustrated in Fig. 5(b). Completely different from other parameters, the relationship between the angle parameter and



(a) Relation between center point and IoU (b) Relation between width and IoU (c) Relation between angle and IoU

Figure 5: Inconsistency in five-parameter regression model. The relationship between height and IoU is similar to relationship between width and IoU.



(a) Discontinuous ℓ_1-loss (b) Continuous ℓ_{mr}^{5p}

Figure 6: Comparison between two loss functions.

IoU is a multiple polynomial function (see Fig. 5(c)). Such regression inconsistency is highly likely to deteriorate the training convergence and the detection performance. Note that we use IoU as the standard measurement is because the final detection performance depends on whether IoU between the prediction and ground truth is high enough.

RSE in Eight-parameter Methods To avoid the inherent regression inconsistency, the eight-parameter representation has been developed (Liao, Shi, and Bai 2018; Liu and Jin 2017; Liu et al. 2019). Specifically, the eight-parameter regression-based detectors directly regress the four corners of the object, so the prediction is a quadrilateral.

However, the discontinuity of loss still exists in the eight-parameter regression model. For example, we can suppose that a ground truth box can be described with the corner sequence $a \rightarrow b \rightarrow c \rightarrow d$ (see red box in Fig. 2). However, the corner sequence becomes $d \rightarrow a \rightarrow b \rightarrow c$ (see green box in Fig. 2) when the ground truth box is slightly rotated by a small angle. Therefore, consider the situation of an eight-parameter regression in the boundary case, as shown in Fig. 4(b). The actual regression process from the blue reference box to the green ground truth box is $\{(a \rightarrow a), (b \rightarrow b), (c \rightarrow c), (d \rightarrow d)\}$, but apparently the ideal regression process should be $\{(a \rightarrow b), (b \rightarrow c), (c \rightarrow d), (d \rightarrow a)\}$. This situation also causes the model training more difficulty and the unsmooth of regression.

The Proposed Modulated Rotation Loss

In this part, we propose ℓ_{mr} to solve RSE. The pseudo equation of ℓ_{mr} can be described as:

$$\ell_{mr} = \min \begin{cases} \ell_1(para.) \\ \ell_1(modulated - para.). \end{cases} \quad (1)$$

Five-parameter Modulated Rotation Loss The RSE only occurs in the boundary case (see Fig. 6(a)). In this paper, we devise the following boundary constraints to modulate the loss as termed by modulated rotation loss ℓ_{mr}:

$$\ell_{cp} = |x_1 - x_2| + |y_1 - y_2|, \quad (2)$$

$$\ell_{mr}^{5p} = \min \begin{cases} \ell_{cp} + |w_1 - w_2| + |h_1 - h_2| + |\theta_1 - \theta_2| \\ \ell_{cp} + |w_1 - h_2| + |h_1 - w_2| \\ \qquad +|90 - |\theta_1 - \theta_2||, \end{cases} \quad (3)$$

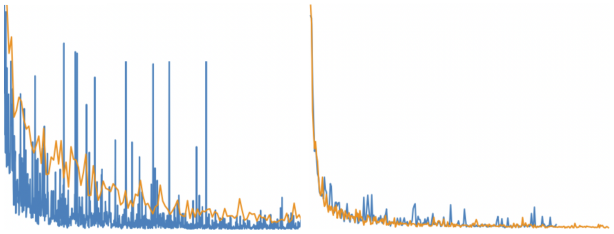
where ℓ_{cp} is the central point loss. The first item in ℓ_{mr} is ℓ_1-loss. The second item is a correction used to make the loss continuous by eliminating the angular periodicity and the exchangeability of height and width. This correction is particularly larger than ℓ_1-loss when it does not reach the range boundary of the angle parameter. However, this correction becomes normal when ℓ_1-loss is abrupt. In other words, such correction can be seen as the symmetry of ℓ_1-loss about the location of the mutation. Finally, ℓ_{mr} takes the minimum of ℓ_1-loss and the correction. The curve of ℓ_{mr} is continuous, as sketched in Fig. 6(b).

In practice, relative values of bounding box regression are usually used to avoid errors caused by objects on different scales. Therefore, ℓ_{mr} in this paper is expressed as follows:

$$\nabla \ell_{cp} = |t_{x1} - t_{x2}| + |t_{y1} - t_{y2}|, \quad (4)$$

$$\ell_{mr}^{5p} = \min \begin{cases} \nabla \ell_{cp} + |t_{w1} - t_{w2}| \\ +|t_{h1} - t_{h2}| + |t_{\theta 1} - t_{\theta 2}| \\ \nabla \ell_{cp} + |t_{w1} - t_{h2} - \log(r)| \\ +|t_{h1} - t_{w2} + \log(r)| + ||t_{\theta 1} - t_{\theta 2}| - \frac{\pi}{2}|, \end{cases} \quad (5)$$

where $t_x = (x - x_a)/w_a, t_y = (y - y_a)/h_a$, $t_w = \log(w/w_a), t_h = \log(h/h_a), r = \frac{h}{w}, t_\theta = \frac{\theta\pi}{180}$. Here the measurement unit of the angle parameter is radian, r represents the aspect ratio. x and x_a are respectively the predicted box and the anchor box (likewise for y, w, h, θ).



(a) Loss curves (five-param.) (b) Loss curves (eight-param.)

Figure 7: Comparisons of loss curves during training: where blue for L1-loss and yellow for our loss.

Eight-parameter Modulated Rotation Loss Here we devise the eight-parameter version of our modulated rotation loss which consists of three components: i) move the four vertices of the predicted box clockwise by one place; ii) keep the order of the vertices of the predicted box unchanged; iii) move the four vertices of the predicted box counterclockwise by one place; iv) take the minimum value in the above three cases. Therefore, ℓ_{mr}^{8p} is expressed as follows:

$$\ell_{mr}^{8p} = \min \begin{cases} \sum_{i=0}^{3} \left(\frac{|x_{(i+3)\%4} - x_i^*|}{w_a} + \frac{|y_{(i+3)\%4} - y_i^*|}{h_a} \right) \\ \sum_{i=0}^{3} \left(\frac{|x_i - x_i^*|}{w_a} + \frac{|y_i - y_i^*|}{h_a} \right) \\ \sum_{i=0}^{3} \left(\frac{|x_{(i+1)\%4} - x_i^*|}{w_a} + \frac{|y_{(i+1)\%4} - y_i^*|}{h_a} \right) \end{cases} \tag{6}$$

where x_i and y_i denote the i-th vertex coordinates of the predicted box and the reference box. x_i^*, y_i^* respectively denote the i-th vertex coordinates of the ground truth box and the reference box.

Through the eight-parameter regression and the definition of ℓ_{mr}^{8p}, the problems of the regression inconsistency and the loss discontinuity in rotation detection are eliminated. Extensive experiments show that our method is more stable for training (see Fig. 7) and outperforms other methods.

Experiments

Recall that the main contribution of this paper is to identify the problem of RSE and solve it through modulated rotation loss. Experiments are implemented by Tensorflow (Abadi et al. 2016) on a server with Ubuntu 16.04, NVIDIA GTX 2080 Ti, and 11G Memory.

Datasets and Implementation Details

DOTA (Xia et al. 2018): The main experiments are carried out around DOTA which has a total of 2,806 aerial images and 15 categories. The size of images in DOTA ranges from 800×800 pixels to $4,000 \times 4,000$ pixels. The proportions of the training set, the validation set, and the test set are respectively 1/2, 1/6, and 1/3. There are 188,282 instances for training and validation, and they are labeled with a clockwise quadrilateral. In this paper, we use the 1.0 version of annotations for rotated object detection. Due to the large size of a single aerial image, we divide the image into 600×600 pixel sub-images with a 150-pixel overlap between two neighboring ones, and these sub-images are eventually scaled to 800×800.

Backbone	Loss	Regression	mAP
resnet-50	smooth-ℓ_1	five-param.	62.14
resnet-50	ℓ_{mr}	five-param.	64.49
resnet-50	smooth-ℓ_1	eight-param.	65.59
resnet-50	ℓ_{mr}	eight-param.	**66.77**

Table 1: Ablation experiments of ℓ_{mr} and predefined eight-parameter regression on DOTA benchmark.

ICDAR2015 (Karatzas et al. 2015): a scene text dataset that includes a total of 1,500 images, 1000 of which are used for training and the remaining for testing. The size of the images in this dataset is 720×1280, and the source of the images is street view. The annotation of the text in an image is four clockwise point coordinates of a quadrangle.

HRSC2016 (Liu et al. 2017): HRSC2016 is a dataset for ship detection which range of aspect ratio and that of arbitrary orientation are large. This dataset contains two scenarios: ship on sea and ship close inshore. The size of each image ranges from 300×300 to $1,500 \times 900$. This dataset has 1,061 images including 436 images for training, 181 images for validation, and 444 for testing.

UCAS-AOD (Zhu et al. 2015): UCAS-AOD is a remote sensing dataset which contains two categories: car and plane. UCAS-AOD contains 1510 aerial images, each of which has approximately $659 \times 1,280$ pixels. In line with (Ding et al. 2019) and (Azimi et al. 2018), we randomly select 1110 images for training and 400 ones for testing.

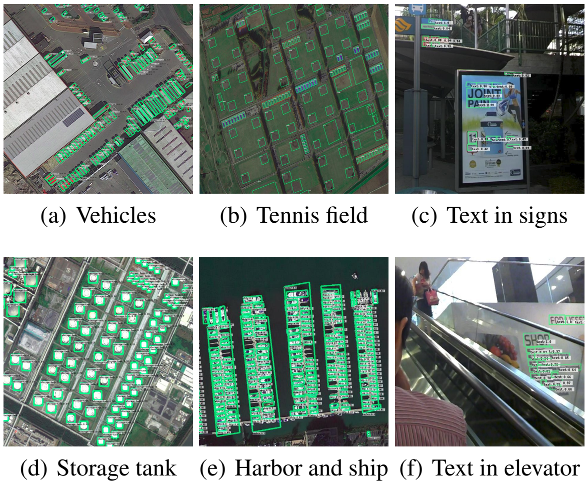
Baselines and Training Details. To make the experimental results more reliable, the baseline we chose is a multi-class rotated object detector based on RetinaNet, which has been verified in work (Yang et al. 2019b). During training, we use RetinaNet-Res50, RetinaNet-Res101, and RetinaNet-Res152 (Lin et al. 2017b) for experiments. Our network is initialized with the pre-trained ResNet50 (He et al. 2016) for object classification in ImageNet (Deng et al. 2009), and the pre-trained models are officially published by TensorFlow. Besides, weight decay and momentum are correspondingly 1e-4 and 0.9. The training epoch is 30 in total, and the number of iterations per epoch depends on the number of samples in the dataset. The initial learning rate is 5e-4, and the learning rate changes from 5e-5 at epoch 18 to 5e-6 at epoch 24. In the first quarter of the training epochs, we adopt the warm-up strategy to find a suitable learning rate. In inference, rotating non-maximum suppression (R-NMS) is used for post-processing the final detection results.

Ablation Study

Modulated Rotation Loss in 5-parameter and 8-parameter settings. We use the ResNet50-based RetinaNet-H as our baseline to verify the effectiveness of modulated

Loss	Regression	mAP
smooth-ℓ_1	five-param. $[-\frac{\pi}{2},0)$	62.14
smooth-ℓ_1(Xia et al. 2018)	five-param. $[-\pi,0)$	62.39
IoU-smooth-ℓ_1 (Yang et al. 2019c)	five-param. $[-\frac{\pi}{2},0)$	62.69
ℓ_{mr}	five-param. $[-\frac{\pi}{2},0)$	64.49
smooth-ℓ_1	eight-param.	65.59
ℓ_{mr}	eight-param.	**66.77**

Table 2: Ablation study using the proposed techniques on DOTA. RetinaNet-H(Yang et al. 2019b) is the base model.


(a) Vehicles (b) Tennis field (c) Text in signs

(d) Storage tank (e) Harbor and ship (f) Text in elevator

Figure 8: Detection results on DOTA and ICDAR15.

Backbone	Data Aug	Balance	mAP
resnet-50			66.77
resnet-50	✓		70.79
resnet-50	✓	✓	71.22
resnet-101	✓	✓	72.16
resnet-152	✓	✓	**73.51**

Table 3: Ablation experiments of backbone, data augmentation and balance on DOTA. RSDet is the base model.

Loss	Regression	ICDAR2015	HRSC2016
smooth-ℓ_1	five-param.	76.8	82.4
ℓ_{mr}	five-param.	79.6	83.6
smooth-ℓ_1	eight-param.	81.2	85.4
ℓ_{mr}	eight-param.	**83.2**	**86.5**

Table 4: Performances of ℓ_{mr} and eight-para. regression on ICDAR2015 and HRSC2016. Use RetinaNet-H (Yang et al. 2019b) as base model, and ResNet152 backbone.

rotation loss ℓ_{mr}. We get a gain of 2.35% mAP, when the loss function is changed from the smooth-ℓ_1 loss to ℓ_{mr}, as shown in Table. 1. Fig. 1 compares results before and after solving the RSE problem: some objects in the images are all in the boundary case where the loss function is not continuous. A lot of inaccurate results (see red squares in Fig. 1(a)) are predicted in the baseline method, but these do not occur after using ℓ_{mr} (see the same location in Fig. 1(b)). Similarly, an improvement of 1.18% mAP is obtained after using the eight-parameter ℓ_{mr}. This set of ablation experiments prove that ℓ_{mr} is effective for improving the rotated object detector. Note the number of parameters and calculations added by these two techniques are almost negligible.

Training Stability. We have analyzed that the discontinuity of loss greatly affects the training stability and the detection performance in detail. Although the detection performance using our techniques has been verified through mAP, we have not proven the stability improvement of model training brought by our techniques. To this end, we plot the training loss curves using models including RetinaNet-H (ℓ^{5p}), RSDet-I (ℓ^{5p}_{mr}), RetinaNet-H (ℓ^{8p}), and RSDet-I (ℓ^{8p}_{mr}), as shown in Fig. 7. The training process converges more stable with modulated rotation losses.

Comparison with Similar Methods. Although we introduce the concept of RSE for the first time, it is worth noting that some previous articles have also mentioned similar problems. In (Xia et al. 2018), a 180-degree definition is used to eliminate the loss burst caused by the exchangeability of height and width. While related works (Ma et al. 2018; Bao et al. 2019) use periodic trigonometric functions to eliminate the effects of the angular periodicity. SCRDet (Yang et al. 2019c) proposes IoU-smooth-ℓ_1 loss to solve the boundary discontinuity. However, these methods are limited and do not completely solve the RSE problem. Our approach yields the most promising results while compare with these methods as shown in Table. 2.

Backbone, Data Augmentation, and Data Balance. Data augmentation is effective to improve performance. Operations of augmentations we use include random horizontal flipping, random vertical flipping, random image graying, and random rotation. Consequently, the baseline performance increased by 4.22% to 70.79% on DOTA. Data imbalance is severe in the DOTA. For instance, there are 76,833 ship instances in the dataset, but there are only 962 ground track fields. We extend samples fewer than 10,000 to 10,000 in each category by random copying, which brings a 0.43% boost, and the most prominent contribution is from a small number of samples e.g. helicopter and swimming pool. We also explore the impact of different backbones and tend to conclude that larger backbones bring more performance gains. Performances of the detectors based on ResNet50, ResNet101, and ResNet152 are respectively 71.22%, 72.16% and 73.51%. Refer to Table. 3 for details.

Using Two-stage Detectors as Base Model. To prove that ℓ_{mr} can cooperate well with other existing frameworks, we apply ℓ_{mr} to the two-stage methods including RSDet-II and Faster RCNN. By comparison, we find that ℓ_{mr} can achieve consistent improvement on RSDet-II and Faster RCNN. Note that Faster RCNN are based on five-parameter and eight-parameter methods, and RSDet-II is only conducted on the basis of five parameters. Taking Faster RCNN for example, it outperforms baseline by 1.6% and 2.84% in mAP after cooperating with ℓ_{mr} in five-para. and eight-para. respectively. The results of RSDet-II can also be seen in Ta-

Method	BB	Reg.	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Two-stage																		
IENet	R-101	6p	80.2	64.5	39.8	32.0	49.7	65.0	52.6	81.5	44.7	78.5	46.5	56.7	64.4	64.2	36.8	57.1
R-DFPN	R-101	5p	80.9	65.8	33.8	58.9	55.8	50.9	54.8	90.3	66.3	68.7	48.7	51.8	55.1	51.3	35.9	57.9
R^{2}CNN	R-101	5p	80.9	65.7	35.3	67.4	59.9	50.9	55.8	90.7	66.9	72.4	55.1	52.2	55.1	53.4	48.2	60.7
RRPN	R-101	5p	88.5	71.2	31.7	59.3	51.9	56.2	57.3	90.8	72.8	67.4	56.7	52.8	53.1	51.9	53.6	61.0
ICN	R-101	5p	81.4	74.3	47.7	70.3	64.9	67.8	70.0	90.8	79.1	78.2	53.6	62.9	67.0	64.2	50.2	68.2
RoI Trans.	R-101	5p	88.6	78.5	43.4	75.9	68.8	73.7	83.6	90.7	77.3	81.5	58.4	53.5	62.8	58.9	47.7	69.6
CAD-Net	R-101	5p	87.8	82.4	49.4	73.5	71.1	63.5	76.7	90.9	79.2	73.3	48.4	60.9	62.0	67.0	62.2	69.9
SCRDet	R-101	5p	89.9	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
Gliding Ver.	R-101	9p	89.6	85.0	52.3	**77.3**	73.0	73.1	86.8	90.7	79.0	86.8	59.6	70.9	72.9	70.9	57.3	75.0
FFA	R-101	5p	90.1	82.7	54.2	75.2	71.0	**79.9**	83.5	90.7	83.9	84.6	61.2	68.0	70.7	76.0	63.7	75.7
APE	R-101	5p	89.9	83.6	53.4	76.0	74.0	77.2	79.5	90.8	87.2	84.5	**67.7**	60.3	74.6	71.8	65.6	75.8
CenterMap	R-101	5p	89.8	84.4	54.6	70.3	**77.7**	78.3	**87.2**	90.7	84.9	85.3	56.5	69.2	74.1	71.6	66.1	76.0
RSDet-II	R-152	8p	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**
One-stage																		
P-RSDet	R-101	3p	89.0	73.7	47.3	72.0	70.6	73.7	72.8	90.8	80.1	81.3	59.5	57.9	60.8	65.2	52.6	69.8
O^2-DNet	H-104	10p	89.3	82.1	47.3	61.2	71.3	74.0	78.6	90.8	82.2	81.4	60.9	60.2	58.2	66.9	61.0	71.0
DRN	H-104	5p	89.7	82.3	47.2	64.1	76.2	74.4	85.8	90.6	86.2	84.9	57.7	61.9	69.3	69.6	58.5	73.2
R^{3}Det	R-152	5p	89.5	81.2	50.5	66.1	70.9	78.7	78.2	90.8	85.3	84.2	61.8	63.8	68.2	69.8	**67.2**	73.7
RSDet-I	R-152	8p	**90.0**	83.9	**54.7**	69.9	70.6	79.6	75.4	**91.2**	**88.0**	85.6	65.2	69.2	67.0	70.2	64.6	**75.0**

Table 5: Detection accuracy on different objects and overall performances with the state-of-the-art methods on DOTA. The short names for categories are defined as (abbreviation-full name): PL-Plane, BD-Baseball diamond, BR-Bridge, GTF-Ground field track, SV-Small vehicle, LV-Large vehicle, SH-Ship, TC-Tennis court, BC-Basketball court, ST-Storage tank, SBF-Soccer-ball field, RA-Roundabout, HA-Harbor, SP-Swimming pool, and HC-Helicopter. BB means Backbone.

ble. 5, which reach state-of-the-art on DOTA benchmark.

Performances on Other Datasets. We further do experiments on ICDAR2015, and HRSC2016 as shown in Table. 4. For ICDAR2015, there are rich existing methods such as R^{2}CNN, Deep direct regression (He et al. 2017) and FOTS (Liu et al. 2018), and the current state-of-art has reached 91.67%. They all have a lot of text-based tricks, but we find that they are also not aware of the rotation sensitivity error. Therefore, we conduct some verification experiments based on ℓ_{mr}. Positive results are obtained for all validations on both datasets. Our detector performs competitively which shows its generalization on scene text data. Besides, our method has also been verified on HRSC2016, and the experimental results are also comparable to state-of-art.

Overall Evaluation

The results on DOTA are shown in Table 5. The compared methods include i) one-stage methods: P-RSDet (Zhou et al. 2020) , O^2-DNet (Wei et al. 2020) , DRN (Pan et al. 2020), R^{3}Det (Yang et al. 2019b) ii) two stage methods: R^{2}CNN (Jiang et al. 2017), Gliding Vertex (Xu et al. 2020), FFA (Fu et al. 2020), APE (Zhu, Du, and Wu 2020), CenterMap OBB (Wang et al. 2020). The results of DOTA are all obtained by submitting predictions to official evaluation server. We apply ℓ_{mr} to the one-stage and two-stage methods respectively, and call them RSDet-I and RSDet-II respectively. In the one-stage methods, RSDet-I obtained 75.03% mAP, which is 1.29% higher than the existing best method (R^{3}Det). And in the two-stage methods, RSDet-II obtained 76.34% mAP, which is 0.31% higher than CenterMap OBB, 1.28% higher than Gliding Vertex. Table. 6 gives the comparison on UCAS-AOD dataset, where our method achieves 96.50% for OBB task which outperforms all the published methods.

Method	Plane	Car	mAP
YOLOv2 (Redmon et al. 2016)	96.60	79.20	87.90
R-DFPN (Yang et al. 2018b)	95.90	82.50	89.20
DRBox (Liu, Pan, and Lei 2017)	94.90	85.00	89.95
S^{2}ARN (Bao et al. 2019)	97.60	92.20	94.90
RetinaNet-H (Yang et al. 2019b)	97.34	93.60	95.47
ICN (Azimi et al. 2018)	-	-	95.67
FADet (Li et al. 2019a)	98.69	92.72	95.71
R^{3}Det (Yang et al. 2019b)	98.20	94.14	96.17
Ours (RSDet)	**98.04**	**94.97**	**96.50**

Table 6: Performance evaluation on UCAS-AOD dataset.

Moreover, the amount of parameters and calculations added by our techniques are almost negligible, and they can be applied to all region-based rotation detection algorithms. Visualization results are shown in Fig. 8.

Conclusion

In this paper, the issue of rotation sensitivity error (RSE) is formally identified and formulated for region-based rotated object detectors. RSE mainly refers to the discontinuity of loss which caused by the contradiction between the definition of the rotated bounding box and the loss function. We propose a novel modulated rotation loss ℓ_{mr} to address the discontinuity of loss in five-parameter and eight-parameter methods. To prove the effectiveness of modulated loss, we conduct experiments based on one-stage and two-stage methods respectively. Extensive experiments demonstrate that RSDet achieves state-of-art performance on the DOTA benchmark and is also proven good generalization and robustness on different datasets and multiple detectors.

Acknowledgements

This research was partially supported by the Commercialization of Research Findings Fund of Inner Mongolia Autonomous Region (2020CG0075), China Major State Research Development Program (2018AAA0100704), NSFC (61972250, U20B2068), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, 265–283.

Azimi, S. M.; Vig, E.; Bahmanyar, R.; Körner, M.; and Reinartz, P. 2018. Towards multi-class object detection in unconstrained remote sensing imagery. In *Asian Conference on Computer Vision*, 150–165. Springer.

Bao, S.; Zhong, X.; Zhu, R.; Zhang, X.; Li, Z.; and Li, M. 2019. Single Shot Anchor Refinement Network for Oriented Object Detection in Optical Remote Sensing Imagery. *IEEE Access* 7: 87150–87161.

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6154–6162.

Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. 2019. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4974–4983.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, 379–387.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Ding, J.; Xue, N.; Long, Y.; Xia, G.; and Lu, Q. 2019. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2849–2858. Computer Vision Foundation / IEEE. doi: 10.1109/CVPR.2019.00296.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88(2): 303–338.

Fu, C.; Liu, W.; Ranga, A.; Tyagi, A.; and Berg, A. C. 2017. DSSD : Deconvolutional Single Shot Detector. *CoRR* abs/1701.06659.

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.

Fu, K.; Li, Y.; Sun, H.; Yang, X.; Xu, G.; Li, Y.; and Sun, X. 2018. A ship rotation detection model in remote sensing images based on feature fusion pyramid network and deep reinforcement learning. *Remote Sensing* 10(12): 1922.

Girshick, R. 2015. Fast r-cnn. In *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*, volume 1, 1435–1443. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, W.; Zhang, X.-Y.; Yin, F.; and Liu, C.-L. 2017. Deep direct regression for multi-oriented scene text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 745–753.

Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; and Luo, Z. 2017. R2CNN: Rotational Region CNN for Orientation Robust Scene Text Detection. *CoRR* abs/1706.09579. URL http://arxiv.org/abs/1706.09579.

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*, 1156–1160. IEEE.

Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; and Shi, J. 2019. FoveaBox: Beyond Anchor-based Object Detector. *CoRR* abs/1904.03797. URL http://arxiv.org/abs/1904.03797.

Li, C.; Xu, C.; Cui, Z.; Wang, D.; Zhang, T.; and Yang, J. 2019a. Feature-Attentioned Object Detection in Remote Sensing Imagery. In *IEEE International Conference on Image Processing*, 3886–3890. IEEE.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2019b. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. *CoRR* abs/1909.00133.

Liao, M.; Shi, B.; and Bai, X. 2018. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing* 27(8): 3676–3690.

Liao, M.; Zhu, Z.; Shi, B.; Xia, G.-s.; and Bai, X. 2018. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5909–5918.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T. Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017b. Focal Loss for Dense Object Detection. *IEEE transactions on pattern analysis and machine intelligence* PP(99): 2999–3007.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755. Springer.

Liu, L.; Pan, Z.; and Lei, B. 2017. Learning a Rotation Invariant Detector with Rotatable Bounding Box. *CoRR* abs/1711.09405. URL http://arxiv.org/abs/1711.09405.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, 21–37. Springer.

Liu, X.; Liang, D.; Yan, S.; Chen, D.; Qiao, Y.; and Yan, J. 2018. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5676–5685.

Liu, Y.; and Jin, L. 2017. Deep matching prior network: Toward tighter multi-oriented text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1962–1969.

Liu, Y.; Zhang, S.; Jin, L.; Xie, L.; Wu, Y.; and Wang, Z. 2019. Omnidirectional Scene Text Detection with Sequential-free Box Discretization. In Kraus, S., ed., *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 3052–3058. ijcai.org. doi:10.24963/ijcai.2019/423.

Liu, Z.; Yuan, L.; Weng, L.; and Yang, Y. 2017. A High Resolution Optical Satellite Image Dataset for Ship Recognition and Some New Baselines 324–331.

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.

Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; and Xu, C. 2020. Dynamic Refinement Network for Oriented and Densely Packed Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11204–11213. IEEE. doi:10.1109/CVPR42600.2020.01122.

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.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(6): 1137–1149. doi:10.1109/TPAMI.2016.2577031.

Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; and LeCun, Y. 2014. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations*.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 9626–9635. IEEE. doi:10.1109/ICCV.2019.00972.

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* .

Wei, H.; Zhang, Y.; Chang, Z.; Li, H.; Wang, H.; and Sun, X. 2020. Oriented objects as pairs of middle lines. *ISPRS Journal of Photogrammetry and Remote Sensing* 169: 268–279.

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*, 3974–3983.

Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; and Bai, X. 2020. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP(99): 1–1.

Yang, J.; Ji, L.; Geng, X.; Yang, X.; and Zhao, Y. 2019a. Building detection in high spatial resolution remote sensing imagery with the U-Rotation Detection Network. *International Journal of Remote Sensing* 40(15): 6036–6058.

Yang, X.; Fu, K.; Sun, H.; Sun, X.; Yan, M.; Diao, W.; and Guo, Z. 2018a. Object Detection with Head Direction in Remote Sensing Images Based on Rotational Region CNN. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, 2507–2510. IEEE.

Yang, X.; Liu, Q.; Yan, J.; and Li, A. 2019b. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. *arXiv preprint arXiv:1908.05612* .

Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; and Guo, Z. 2018b. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sensing* 10(1): 132.

Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; and Fu, K. 2019c. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In *2019 IEEE/CVF International Conference on Computer Vision*, 8231–8240. IEEE. doi:10.1109/ICCV.2019.00832.

Yang, Z.; Liu, S.; Hu, H.; Wang, L.; and Lin, S. 2019d. RepPoints: Point Set Representation for Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision*, 9656–9665. IEEE. doi:10.1109/ICCV.2019.00975.

Zhang, C.; Liang, B.; Huang, Z.; En, M.; Han, J.; Ding, E.; and Ding, X. 2019. Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 10552–10561.

Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; and Li, S. Z. 2018. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4203–4212.

Zhou, L.; Wei, H.; Li, H.; Zhang, Y.; Sun, X.; and Zhao, W. 2020. Objects detection for remote sensing images based on polar coordinates. *arXiv preprint arXiv:2001.02988* .

Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; and Liang, J. 2017. EAST: An Efficient and Accurate Scene Text Detector. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2642–2651. IEEE Computer Society. doi:10.1109/CVPR.2017.283.

Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; and Jiao, J. 2015. Orientation robust object detection in aerial images using deep convolutional neural network. In *IEEE International Conference on Image Processing*, 3735–3739. IEEE.

Zhu, Y.; Du, J.; and Wu, X. 2020. Adaptive period embedding for representing oriented objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing* 58(10): 7247–7257.