

# Learning Modulated Loss for Rotated Object Detection

Wen Qian<sup>1,2</sup>   Xue Yang<sup>3</sup>   Silong Peng<sup>2</sup>   Yue Guo<sup>2</sup>   Junchi Yan<sup>3</sup>

<sup>1</sup>University of Chinese Academy of Sciences

<sup>2</sup>Institute of Automation, Chinese Academy of Sciences

<sup>3</sup>Shanghai Jiao Tong University

qianwen2018@ia.ac.cn

yangxue-2019-sjtu@sjtu.edu.cn

## Abstract

Popular rotated detection methods usually use five parameters (coordinates of the central point, width, height, and rotation angle) to describe the rotated bounding box and  $\ell_1$  loss as the loss function. In this paper, we argue that the aforementioned integration can cause training instability and performance degeneration, due to the loss discontinuity resulted from the inherent periodicity of angles and the associated sudden exchange of width and height. This problem is further pronounced given the regression inconsistency among five parameters with different measurement units. We refer to the above issues as rotation sensitivity error (RSE) and propose a modulated rotation loss to dismiss the loss discontinuity. Our new loss is combined with the eight-parameter regression to further solve the problem of inconsistent parameter regression. Experiments show the state-of-art performances of our method on the public aerial image benchmark DOTA and UCAS-AOD. Its generalization abilities are also verified on ICDAR2015, HRSC2016, and Fddb. Qualitative improvements can be seen in Fig. 1, and the source code will be released with the publication of the paper.

## 1. Introduction

Object detection is an important and fundamental task in computer vision, and over the decades it has experienced a switch from traditional machine learning methods [1, 2, 3, 4] to deep learning methods. The powerful fitting ability of the convolutional neural network [5] makes a series of great breakthroughs in the field of object detection, and performances of the neural network in many subdivision fields have surpassed those of human beings. Object detection has been extensively applied in face detection [6, 7], automatic driving [8], optical text detection [9], and other fields.

Object detection can generally be divided into horizontal detection and rotation detection depending on directions



(a) RetinaNet-H [10] (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 loss discontinuity.

of the detected boxes. Specifically, horizontal detection, by which all the bounding boxes are set in the horizontal direction, are often more suitable for general natural scene images such as COCO [11] and Pascal VOC [12]. In contrast, in scene text, aerial imagery, face detection, and license plate detection, more accurate positioning is often needed and calls for effective rotation detectors. Until now, more and more rotated object detection datasets such as aerial dataset (DOTA [13], DIOR [14], HRSC2016 [15]), scene text dataset (ICDAR2015 [16], ICDAR2017 [17]), and face dataset (Fddb [18]) have appeared. The existing region-based rotated object detectors usually regress five parameters (coordinates of the central point, width, height, rotation angle) [10, 19, 20, 21] to describe rotated bounding boxes and use  $\ell_1$ -loss as loss functions. However, such methods bear two fundamental issues in practice:

Firstly, The loss discontinuity is caused by angle parameter. The loss value will jump 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 box and the detection box. The position of the reference rectangle has only been slightly changed, but its angle changes a lot due to parameterization. Moreover, the roles of height and width also exchange in a five-parameter system from OpenCV that makes

the case more degenerating.

Moreover, in 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. This issue may be mitigated by the eight-parameter system which use the coordinates of corners with the same units.

The above two issues, as referred collectively as rotation sensitivity error (RSE), which can lead to training instability (see Fig. 7) and resulting detection performance degeneration. In order to address the loss discontinuity, a modulated rotation loss  $\ell_{mr}$  is devised to carefully handle the boundary constraints for rotation, leading to a more smoothed loss curve during training. And then we resort to the eight-parameter regression model [22, 23, 24, 25] (As will be shown later in the paper, the discontinuity caused by rotation boundary inherently exist in eight-parameter system) to avoid regression parameter inconsistency. In the eight-parameter model, all the parameters are point coordinates of four corners in a bounding box such that the regression parameter consistency naturally holds.

In summary, we propose a rotation sensitive detector (RSDet) based on eight-parameter regression model and our modulated rotation loss  $\ell_{mr}$ , which can be trained end-to-end. Our RSDet model shows state-of-art performance on DOTA benchmark and its generalization capability and robustness are further verified on different datasets with different detectors. 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 and regression inconsistency.

ii) For the traditionally widely used five-parameter system (including the adoption in OpenCV), we formally identify the RSE in rotational object detectors. We then devise a special treatment to ensure the loss continuity. The new loss is termed by  $\ell_{mr}^{5p}$  (see Eq. 4).

iii) Even for the more recent eight-parameter system, we still identify the inherent discontinuity issue and develop a corresponding treatment to smooth the loss function. The new loss is termed by  $\ell_{mr}^{8p}$  (see Eq. 6), and the resulting detector with RetinaNet-H [26] as based model is termed rotation sensitive detector i.e. RSDet in this paper.

## 2. Related Work

**Horizontal Object Detectors** Visual object detection has been a hot topic over the decade. Since the seminal work R-CNN [27], there have been a series of improvements including Fast RCNN [28], Faster RCNN [29], and R-FCN [30], which fall the category of the two-stage methods. On the other hand, single-stage approaches have also been well

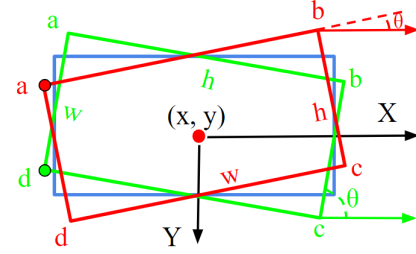
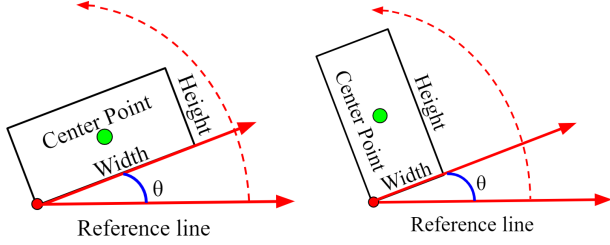


Figure 2: Illustration for the loss discontinuity: rectangles colored in blue, red, and green respectively represent a reference box, a ground truth box, and a predicted box. Assume that the reference box is rotated one degree clockwise to get the ground truth one and is rotated similarly counterclockwise to obtain the predicted one. Consequently, the above three boxes are described with five parameters: the reference box  $(0, 0, 10, 25, -90^\circ)$ , the ground truth box  $(0, 0, 25, 10, -1^\circ)$ , and the predicted box  $(0, 0, 10, 25, -89^\circ)$ . At this time,  $\ell_1$  loss is far more than 0.

developed which can be more efficient than the two-stage methods. Examples include Overfeat [31], YOLO [32], and SSD [33]. In particular, SSD [33] 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 [34], RetinaNet [35], and DSSD [36]. Recently, many cascaded or refined detectors are proposed. For example, Cascade RCNN [37], HTC [38], and FSCascade [39] 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 [40], FoveaBox [41], and RepPoints [26]. 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 bounding boxes along the horizontal direction, 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 [26].

**Rotated Object Detector** Rotated object detection has been widely used in natural scene text, aerial image, etc. And these detectors typically use rotated bounding boxes to describe positions of objects, which are more accurate than those using horizontal boxes. Represented by scene text, many excellent detectors have been proposed. For example, RRPN [21] uses rotating anchors to improve the qualities of region proposals. R<sup>2</sup>CNN [20] is a multi-tasking text detector that simultaneously detects rotated and horizontal



(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  $\theta$  ranges from -90 degree to 0 degree, but it should be distinguished from another definition [13], with 180 degree angular range, whose  $\theta$  is determined by the long side of the rectangle and x-axis.

bounding boxes. In TextBoxes++ [42], to accommodate the slenderness of the text, a long convolution kernel is used and the number of proposals is increased. EAST [23] proposes a simple yet powerful pipeline that yields fast and accurate text detection in natural scenes.

Moreover, object detection in aerial images is more difficult, and its main challenges are reflected in multiple categories, multiple scales, 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 [43] combines various modules such as image pyramid, feature pyramid network, and deformable inception sub-networks, and it achieves satisfactory performances on DOTA dataset. RoI Transformer [44] extracts rotation-invariant features for boosting subsequent classification and regression. SCRDet [45] proposes an IoU-smooth  $\ell_1$  loss to solve the sudden loss change caused by the angular periodicity so that it can better handle small, cluttered and rotated objects. R<sup>3</sup>Det [10] 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 consider the inherent loss discontinuity, as stated in Section 1, which we show can hurt learning stability and final detection performances in our experiments. However, no existing studies have addressed this fundamental problem that motivates our work.

### 3. Proposed Method

**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 identify the discontinuity of rotating angle and its resulting sudden exchange between width and height in the five-parameter system. Moreover, we quantitatively show the negative effect of the regression inconsistency in the

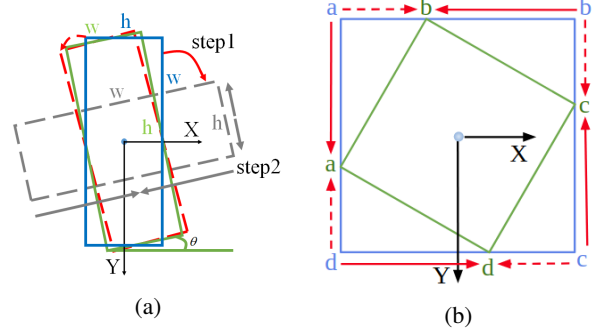


Figure 4: Boundary discontinuity analysis of five-parameter regression and eight-parameter regression. The red solid arrow indicates the actual regression process, and the red dotted arrow indicates the ideal regression process. (a) Five-parameter regression procedure including step 1 and step 2 under boundary conditions. (b) Eight-parameter regression procedure.

five-parameter system caused by the different measurement units. We call such issues collectively as rotation sensitivity error (RSE) and propose a modulated rotation loss for the five-parameter system to achieve more smooth learning. We further point out that even the improved eight-parameter system still suffers from loss discontinuity and then devise a corresponding modulated rotation loss for the eight-parameter system.

#### 3.1. 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  $\theta$ .

While the definition of eight parameters is more simple: four clockwise vertices (a,b,c,d) of the rotated bounding box are used to describe its position, as shown in Fig. 2. The eight-parameter regression methods have natural parameter consistency because it discards the inconsistency which is resulted from the non-coordinate parameter. At the same time, this kind of methods can describe quadrilateral, which can be used in more complex application scenarios.

#### 3.2. Rotation Sensitivity Error

As mentioned earlier, rotation sensitivity error is mainly caused by two reasons: i) The adoption of angle parameter and the resulting height-width exchange (in the popular five-parameter description in OpenCV) contribute to the sudden loss change (increase) in the boundary case. ii) Regression inconsistency of measure units exists in the five-

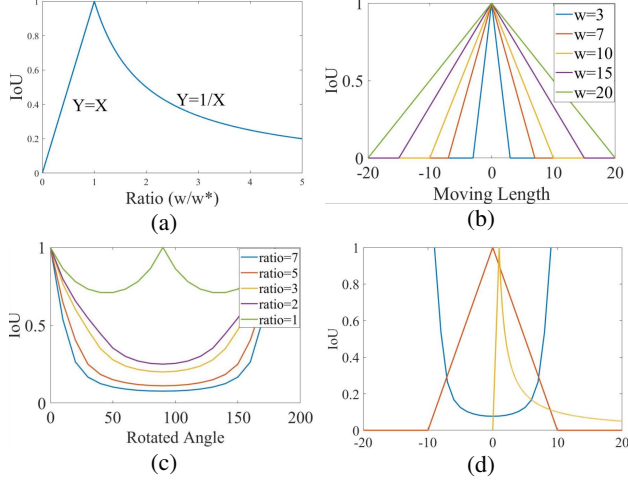
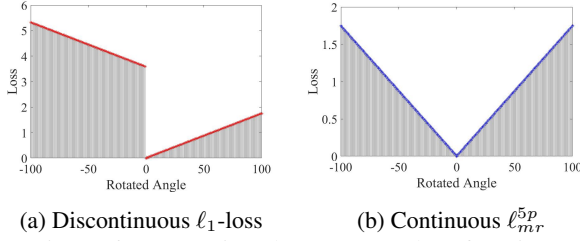


Figure 5: Inconsistency in five-parameter regression model. (a) Relation between angle parameter and IoU. Different colors denote different aspect ratios. (b) Relation between width (similar for height) and IoU. (c) Relation between center point and IoU. (d) Comparison of three relations.



(a) Discontinuous  $\ell_1$ -loss (b) Continuous  $\ell_{mr}^{5p}$   
Figure 6: Comparison between two loss functions.

parameter model.

**Loss Discontinuity.** The angle parameter causes the loss discontinuous. 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  $(0, 0, 100, 25, -90^\circ)$  to  $(0, 0, 100, 25, -100^\circ)$  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  $(0, 0, 25, 100, -10^\circ)$ . Despite the rotation is physically smooth, the loss will be quite large, which corresponds to the loss discontinuity. To avoid such a loss fluctuation, the reference box need to be rotated clockwise to obtain the gray box  $(0, 0, 100, 25, -10^\circ)$  in Fig. 4a (step 1), then width and height of the gray box will be scaled to obtain the final predicted box  $(0, 0, 25, 100, -10^\circ)$  (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 ad-

ressed 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. 5c. The relationship between the central point and IoU is a symmetric linear function, as illustrated in Fig. 5b. Completely different from other parameters, the relationship between the angle parameter and IoU is a multiple polynomial function (see Fig. 5a). 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 predicted box and the ground truth one is high enough.

### 3.3. Five-parameter Modulated Rotation Loss

The loss discontinuity only occurs in the boundary case, as shown in Fig. 6a. In this paper, we devise the following boundary constraints to modulate the loss as termed by modulated rotation loss  $\ell_{mr}$ :

$$\ell_{cp} = |x_1 - x_2| + |y_1 - y_2| \quad (1)$$

$$\ell_{mr}^{5p} = \min \begin{cases} \ell_{cp} + |w_1 - w_2| + |h_1 - h_2| + |\theta_1 - \theta_2| \\ \ell_{cp} + |w_1 - h_2| + |h_1 - w_2| + |90 - |\theta_1 - \theta_2|| \end{cases} \quad (2)$$

where  $\ell_{cp}$  is the central point loss. The first item in  $\ell_{mr}$  is  $\ell_1$ -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  $\ell_1$ -loss when it does not reach the range boundary of the angle parameter. However, this correction becomes normal when  $\ell_1$ -loss is abrupt. In other words, such correction can be seen as the symmetry of  $\ell_1$ -loss about the location of the mutation. Finally,  $\ell_{mr}$  takes the minimum of  $\ell_1$ -loss and the correction. The curve of  $\ell_{mr}$  is continuous, as sketched in Fig. 6b.

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

$$\nabla \ell_{cp} = |t_{x1} - t_{x2}| + |t_{y1} - t_{y2}| \quad (3)$$

$$\ell_{mr}^{5p} = \min \begin{cases} |t_{w1} - t_{w2}| + |t_{h1} - t_{h2}| + |t_{\theta1} - t_{\theta2}| + \nabla \ell_{cp} \\ |t_{w1} - t_{h2} - \log(r)| + |t_{h1} - t_{w2} + \log(r)| \\ + ||t_{\theta1} - t_{\theta2}| - \frac{\pi}{2}| + \nabla \ell_{cp} \end{cases} \quad (4)$$



where

$$\begin{aligned} t_x &= (x - x_a)/w_a, & t_y &= (y - y_a)/h_a \\ t_w &= \log(w/w_a), & t_h &= \log(h/h_a) \\ r &= \frac{w}{h}, & t_\theta &= \frac{\theta\pi}{180} \end{aligned} \quad (5)$$

where 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  $y$ ,  $w$ ,  $h$ , and  $\theta$ ).

### 3.4. Eight-parameter Modulated Rotation Loss

To avoid the inherent regression inconsistency, the eight-parameter representation has been recently developed [42, 46, 47]. Specifically, the eight-parameter regression-based detectors directly regress the four corners of the object, so the prediction is a quadrilateral. The key step to the quadrilateral regression is to sort the four corner points in advance, which can avoid a very large loss even if the pose prediction is correct. For vertex order, we adopt a cross-product based algorithm to obtain the sequence of four vertices, as detailed in Algorithm 1. Note that this algorithm is workable for convex quadrilaterals, and here we use the clockwise order for output without loss of generality. Algorithm 1 is similar to the counterpart proposed in Deep Matching Prior Network [46].

However, the loss discontinuity 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. 4b. 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 difficulty and the unsmooth regression.

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,  $\ell_{mr}^{8p}$  is expressed

---

#### Algorithm 1 Sequence ordering of quadrilateral corners.

---

**Input:** Four vertex of quadrilateral  $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4$

**Output:** Vertex in clockwise order:  $\mathbf{p}'_1, \mathbf{p}'_2, \mathbf{p}'_3, \mathbf{p}'_4$

```

1:  $S \leftarrow \{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4\}, \mathbf{p}'_2 = \mathbf{p}'_3 = \mathbf{p}'_4 = \mathbf{0}$ ;
2:  $\mathbf{p}'_1 \leftarrow FindLeftmostVertex(S)$ ;
3:  $S \leftarrow S - \{\mathbf{p}'_1\}$ 
4: for  $\mathbf{s}_1 \in S$  do
5:    $\mathbf{s}_2, \mathbf{s}_3 \in S - \{\mathbf{s}_1\}$ 
6:   if  $CrossProduct(\mathbf{s}_1 - \mathbf{p}'_1, \mathbf{s}_2 - \mathbf{p}'_1) \times$ 
      $CrossProduct(\mathbf{s}_1 - \mathbf{p}'_1, \mathbf{s}_3 - \mathbf{p}'_1) < 0$  then
7:      $\mathbf{p}'_3 = \mathbf{s}_1, S \leftarrow \{\mathbf{s}_2, \mathbf{s}_3\}$ ;
8:     break;
9:   end if
10: end for
11: for  $\mathbf{s}_1 \in S$  do
12:    $\mathbf{s}_1 = S - \{\mathbf{s}_1\}$ ;
13:   if  $CrossProduct(\mathbf{p}'_3 - \mathbf{p}'_1, \mathbf{s}_1 - \mathbf{p}'_1) > 0$  then
14:      $\mathbf{p}'_2 = \mathbf{s}_1, \mathbf{p}'_4 = \mathbf{s}_2$ ;
15:   else
16:      $\mathbf{p}'_2 = \mathbf{s}_2, \mathbf{p}'_4 = \mathbf{s}_1$ ;
17:   end if
18: end for
19: return  $\mathbf{p}'_1, \mathbf{p}'_2, \mathbf{p}'_3, \mathbf{p}'_4$ 

```

---

as follows:

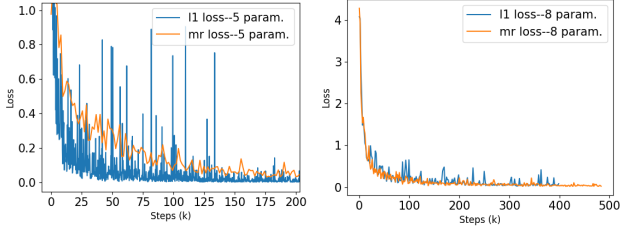
$$\ell_{mr}^{8p} = \min \begin{cases} \sum_{i=0}^3 (|x_{(i+3)\%4} - x_i^*| + |y_{(i+3)\%4} - y_i^*|) \\ \sum_{i=0}^3 (|x_i - x_i^*| + |y_i - y_i^*|) \\ \sum_{i=0}^3 (|x_{(i+1)\%4} - x_i^*| + |y_{(i+1)\%4} - y_i^*|) \end{cases} \quad (6)$$

where  $x_i$  and  $y_i$  respectively represent the coordinate offset between the  $i$ -th vertex of the predicted box and that of the reference box.  $x_i^*, y_i^*$  respectively represents the offset between the  $i$ -th vertex of the ground truth box and that of the reference box.

Through the eight-parameter regression and the definition of  $\ell_{mr}^{8p}$ , 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.

## 4. Experiments

Recall that the main contribution of this paper is to identify the problem of RSE and solve it through modulated rotation loss and eight-parameter regression. Experiments are implemented by Tensorflow [48] on a server with Ubuntu



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

Figure 7: Comparisons of loss curves during training with different loss functions.

16.04, NVIDIA GTX 2080, and 32G Memory. Aerial images (DOTA and UCAS-AOD), scene text images (ICDAR2015 and HRSC2016), and face benchmark (FDBB) are used for evaluation.

#### 4.1. Datasets and Implementation Details

**DOTA [13]:** 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  $800 \times 800$  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  $600 \times 600$  pixel sub-images with a 150-pixel overlap between two neighboring ones, and these sub-images are eventually scaled to  $800 \times 800$ .

**ICDAR2015 [16]:** ICDAR2015 is 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  $720 \times 1280$ , 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 [15]:** 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  $300 \times 300$  to  $1,500 \times 900$ . This dataset has 1061 images including 436 images for training, 181 images for validation, and 444 for testing.

**UCAS-AOD [49]:** 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 [44] and [43], we randomly select 1110 images for training and 400 ones for test.

**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,

Backbone	Loss	Regression	mAP
resnet-50	smooth- $\ell_1$	five-param.	62.14
resnet-50	$\ell_{mr}$	five-param.	64.49
resnet-50	smooth- $\ell_1$	eight-param.	65.59
resnet-50	$\ell_{mr}$	eight-param.	<b>66.77</b>

Table 1: Ablation experiments of  $\ell_{mr}$  and predefined eight-parameter regression on DOTA benchmark. RetinaNet-H [10] is used as the baseline.

Loss	Regression	mAP
smooth- $\ell_1$	five-param. $[-\frac{\pi}{2}, 0)$	62.14
smooth- $\ell_1$	five-param. $[-\pi, 0)$ [13]	62.39
smooth- $\ell_1$	five-param. $[-\frac{\pi}{2}, 0) + \tan$ [21, 52]	NAN
IoU-smooth- $\ell_1$ [45]	five-param. $[-\frac{\pi}{2}, 0)$	62.69
$\ell_{mr}$	five-param. $[-\frac{\pi}{2}, 0)$	64.49
smooth- $\ell_1$	eight-param.	65.59
$\ell_{mr}$	eight-param.	<b>66.77</b>

Table 2: Ablation study using the proposed techniques on DOTA. RetinaNet-H [10] is the base model.

which has been verified in work [10]. During training, we use RetinaNet-Res50, RetinaNet-Res101, and RetinaNet-Res152 [35] for experiments. Our network is initialized with the pre-trained ResNet50 [50] for object classification in ImageNet [51], 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. During the inference, rotating non-maximum suppression (R-NMS) is used for post-processing the final detection results.

#### 4.2. Ablation Study

**Modulated Rotation Loss and Eight-parameter Regression.** We use the ResNet50-based RetinaNet-H as our baseline to verify the effectiveness of modulated rotation loss  $\ell_{mr}$  and eight-parameter regression. We get a gain of 2.35% mAP, when the loss function is changed from the smooth- $\ell_1$  loss to  $\ell_{mr}$ , as shown in Tab. 1. Fig. 1 compares results before and after solving the RSE problem: objects in the images are all in the boundary case where the loss function is not continuous. A lot of inaccurate results (see red circles in Fig. 1a) are predicted in the baseline method, but these do not occur after using  $\ell_{mr}$  (see the same location in Fig. 1b). Similarly, an improvement of 3.45% mAP is obtained after using the eight-parameter regression. Finally, we achieve 66.77% mAP after combining these two techniques. This set of ablation experiments prove that  $\ell_{mr}$  and eight-parameter regression are effective for improving the rotated object detector. More importantly, the number of parameters and calculations added by these two techniques

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
FR-O [13]	79.1	69.1	17.2	63.5	34.2	37.2	36.2	89.2	69.6	59.0	49.4	52.5	46.7	44.8	46.3	52.9
R <sup>2</sup> CNN [20]	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 [21]	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
RetinaNet-H+ResNet50 [10]	88.9	74.5	40.1	58.0	63.1	50.6	63.6	90.9	77.9	76.4	48.3	55.9	50.7	60.2	34.2	62.2
RetinaNet-R+ResNet50 [10]	88.9	67.7	33.6	56.8	66.1	73.3	75.2	90.9	74.0	75.1	43.8	56.7	51.1	55.7	21.5	62.0
ICN [43]	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 Transformer [44]	88.6	78.5	43.4	<b>75.9</b>	68.8	73.7	83.6	90.7	77.3	81.5	58.4	53.5	62.8	58.9	47.7	69.6
SCRDet [45]	90.0	80.7	52.1	68.4	68.4	60.3	72.4	90.9	88.0	<b>86.9</b>	65.0	66.7	66.3	68.2	65.2	72.6
R <sup>3</sup> Det+ResNet152 [10]	89.2	80.8	51.1	65.6	<b>70.7</b>	76.0	<b>78.3</b>	90.8	84.9	84.4	<b>65.1</b>	57.2	<b>68.1</b>	69.0	60.9	72.8
RSDet+ResNet50 (ours)	89.3	82.7	47.7	63.9	66.8	62.0	67.3	90.8	85.3	82.4	62.3	62.4	65.7	68.6	64.6	70.8
RSDet+ResNet101 (ours)	89.8	82.9	48.6	65.2	69.5	70.1	70.2	90.5	85.6	83.4	62.5	63.9	65.6	67.2	<b>68.0</b>	72.2
RSDet+ResNet152 (ours)	<b>90.2</b>	<b>83.5</b>	53.6	70.1	64.6	79.4	67.3	91.0	<b>88.3</b>	82.5	64.1	<b>68.7</b>	62.8	69.5	66.9	73.5
RSDet+ResNet152+Refine (ours)	90.1	82.0	<b>53.8</b>	68.5	70.2	<b>78.7</b>	73.6	<b>91.2</b>	87.1	84.7	64.3	68.2	66.1	<b>69.3</b>	63.7	<b>74.1</b>

Table 3: Detection accuracy (AP for each category and overall mAP) 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. For RetinaNet, 'H' and 'R' denote horizontal anchors and the rotated anchors, respectively.

Backbone	Data Augmentation	Balance	mAP
resnet-50			66.77
resnet-50	✓		70.79
resnet-50	✓	✓	71.22
resnet-101	✓	✓	72.16
resnet-152	✓	✓	<b>73.51</b>

Table 4: Ablation experiments of backbone, data augmentation and balance on DOTA. RSDet is the base model.

Loss	Regression	ICDAR2015	HRSC2016
smooth- $\ell_1$	five-param.	76.8	82.4
$\ell_{mr}$	five-param.	79.6	83.6
smooth- $\ell_1$	eight-param.	81.2	85.4
$\ell_{mr}$	eight-param.	<b>83.2</b>	<b>86.5</b>

Table 5: Performances of  $\ell_{mr}$  and eight-parameter regression on ICDAR2015 and HRSC2016. RetinaNet-H [10] is the base model, and ResNet152 is the backbone.

are almost negligible.

**Training Stability.** In Section 3, we have analyzed that the loss discontinuity and the regression inconsistency greatly affect the training stability and the detection performance in detail. Although the detection performance using our techniques has been verified through mAPs, 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 ( $\ell^{5p}$ ), RetinaNet-H ( $\ell^{5p}_{mr}$ ), RetinaNet-H ( $\ell^{8p}$ ), and RSDet ( $\ell^{8p}_{mr}$ ), as shown in Fig. 7. We can see that training convergences become more stable after using modulated rotation losses.

**Comparison with Similar Methods.** Although we formally introduce the concept of RSE for the first time, it is worth noting that some previous articles have also mentioned similar problems. In [13], a 180-degree angle definition is used to eliminate the loss burst caused by the exchangeability of height and width. While related works [21, 52] use periodic trigonometric functions (such as  $\tan$ ) to eliminate the effects of the angular periodicity. SCRDet [45] proposes IoU-smooth- $\ell_1$  loss to solve the boundary discontinuity. However, these methods are limited and do not completely solve the RSE problem. Tab. 2 compares our proposed method with other methods mentioned above. Our approach still yields the most promising results.

**Backbone, Data Augmentation, and Data Balance.** Data augmentation is effective to improve detection per-

formance. 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 ones in each category by copying, which brings a 0.43% boost, and the most prominent contribution is from a small number of samples such as helicopter and swimming pool. We also explore the impact of different backbones on the detector and 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 Tab. 4 for detailed results.

**Regression Refinement.** R<sup>3</sup>Det proposes to increase the accuracy of the regression box by adding a refinement stage. The idea of cascading is also used by other detection methods, such as RefineDet [39] and Cascade RCNN [37]. Therefore, in order to further improve the performance of RSDet, we also add a refinement stage. The foreground and background thresholds for the refinement stage are 0.6 and 0.5, respectively. Tab. 3 shows that the refinement stage helps improve the performance from 73.51% to 74.13%, especially for objects with large aspect ratios such as ships, vehicles, and harbors.



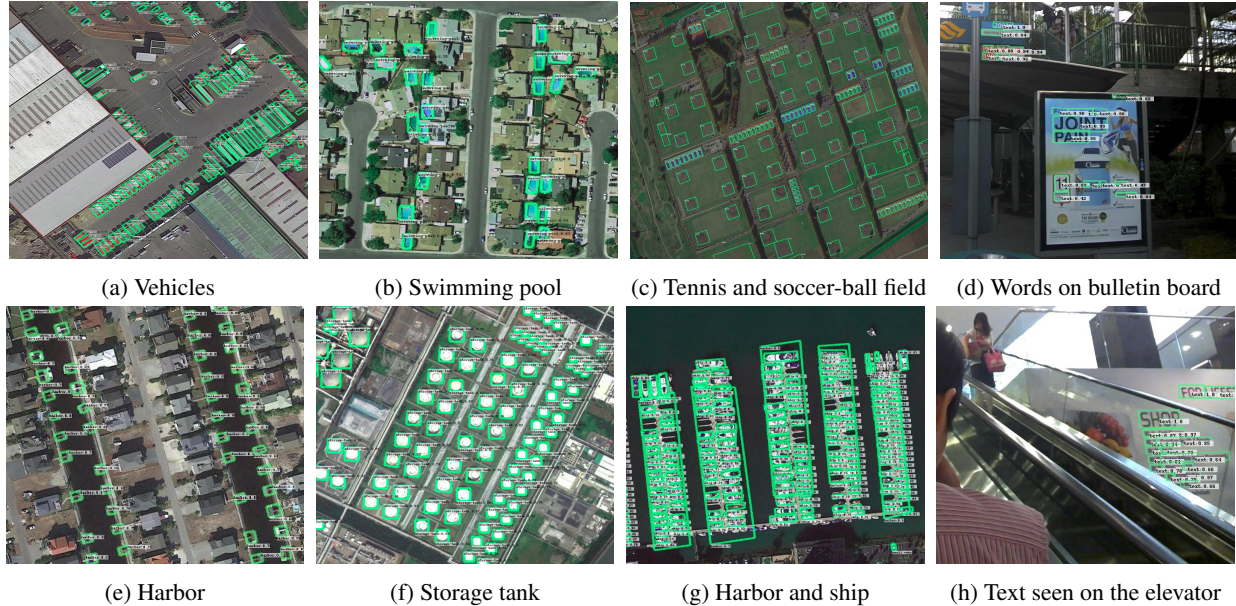


Figure 8: Detection results on DOTA and ICDAR15.

Method	Plane	Car	mAP
YOLOv2 [32]	96.60	79.20	87.90
R-DFPN [19]	95.90	82.50	89.20
DRBox [54]	94.90	85.00	89.95
S <sup>2</sup> ARN [52]	97.60	92.20	94.90
RetinaNet-H [10]	97.34	93.60	95.47
ICN [43]	-	-	95.67
FADet [55]	98.69	92.72	95.71
R <sup>3</sup> Det [10]	98.20	94.14	96.17
Ours (RSDet)	<b>98.04</b>	<b>94.97</b>	<b>96.50</b>

Table 6: Performance evaluation on UCAS-AOD dataset.

**Using Two-stage Detectors as Base Model.** Extra experiments are performed based on the Rotating Faster RCNN. Unlike RetinaNet, Faster RCNN is a two-stage detector. We take a rotating Faster RCNN as the baseline, then add the  $\ell_{mr}$  and eight-parameter regression method for the ablation experiments. The performance improvement of these two techniques are 1.6% and 2.84%, respectively.

**Performances on Other Datasets.** We further do experiments on ICDAR2015, and HRSC2016 as shown in Tab. 5. For ICDAR2015, there are rich existing methods such as R<sup>2</sup>CNN, Deep direct regression [25] and FOTS [53], 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  $\ell_{mr}$  and eight-parameter regression method. Positive results are obtained for all validation experiments on both datasets. Our detector performs competitively which shows the generalization of our method on scene text data. Besides, our method has also been verified on HRSC2016, and the experimental results are also comparable to state-of-art.

### 4.3. Overall Evaluation

The results on DOTA are shown in Table 3. The compared methods include i) traditional deep learning methods, such as Faster RCNN [29] and RetinaNet [35]; ii) scene text detection methods, like R<sup>2</sup>CNN [20] and RPN [21]; iii) recently published methods for multi-category rotation detectors, includes ICN [43], RoI Transformer [44], SCRDet [45] and R<sup>3</sup>Det [10]. The results of DOTA reported here are all obtained by submitting predictions to official DOTA evaluation server. None of the compared methods pays attention to the problem of rotation sensitivity error. To make the comparison fair, clean and direct, we do not use multi-scale training and testing, oversized backbones, and model integration, which are often used on DOTA’s leaderboard methods. For the overall mAP, our method’s performance is 1.3% higher than the existing best method (R<sup>3</sup>Det+ResNet152 [10]). Tab. 6 gives the comparison on UCAS-AOD dataset, where our method achieves 96.50% for OBB task which outperforms all the published methods. 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 on aerial images and natural images are shown in Fig. 8.

## 5. 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 loss discontinuity and the five-parameter regression inconsistency. We propose a novel modulated rotation loss  $\ell_{mr}$  to address the



loss discontinuity and optimize the regression inconsistency with the eight-parameter regression. As a result, the new detector termed as RSDet can be trained end-to-end. Extensive experiments demonstrate that RSDet achieves the state-of-art performance on the DOTA benchmark and is also proven good generalization and robustness on different datasets and multiple detectors.

## References

- [1] P. Viola, M. Jones, *et al.*, “Rapid object detection using a boosted cascade of simple features,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, no. 511-518, p. 3, 2001. **1**
- [2] R. Lienhart and J. Maydt, “An extended set of haar-like features for rapid object detection,” in *IEEE International Conference on Image Processing*, vol. 1, pp. I-I, IEEE, 2002. **1**
- [3] J. Shotton, A. Blake, and R. Cipolla, “Multiscale categorical object recognition using contour fragments,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 7, pp. 1270–1281, 2008. **1**
- [4] D. G. Lowe *et al.*, “Object recognition from local scale-invariant features,” in *The IEEE International Conference on Computer Vision*, vol. 99, pp. 1150–1157, 1999. **1**
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012. **1**
- [6] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, “Faceboxes: A cpu real-time face detector with high accuracy,” in *IEEE International Joint Conference on Biometrics*, pp. 1–9, IEEE, 2017. **1**
- [7] K. Zhang, Z. Zhang, Z. Li, and Q. Yu, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016. **1**
- [8] G. Bagschik, T. Menzel, and M. Maurer, “Ontology based scene creation for the development of automated vehicles,” in *2018 IEEE Intelligent Vehicles Symposium*, pp. 1813–1820, IEEE, 2018. **1**
- [9] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, “Textboxes: A fast text detector with a single deep neural network,” in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. **1**
- [10] X. Yang, Q. Liu, J. Yan, and A. Li, “R3det: Refined single-stage detector with feature refinement for rotating object,” *arXiv preprint arXiv:1908.05612*, 2019. **1, 3, 6, 7, 8**
- [11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Proceedings of the European Conference on Computer Vision*, pp. 740–755, Springer, 2014. **1**
- [12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010. **1**
- [13] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, “Dota: A large-scale dataset for object detection in aerial images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018. **1, 3, 6, 7**
- [14] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, “Object detection in optical remote sensing images: A survey and a new benchmark,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019. **1**
- [15] L. ZK, W. LB, Y. YP, *et al.*, “A high resolution optical satellite image dataset for ship recognition and some new base-lines,” **1, 6**
- [16] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, *et al.*, “Icdar 2015 competition on robust reading,” in *2015 13th International Conference on Document Analysis and Recognition*, pp. 1156–1160, IEEE, 2015. **1, 6**
- [17] R. Gomez, B. Shi, L. Gomez, L. Numann, A. Veit, J. Matas, S. Belongie, and D. Karatzas, “Icdar2017 robust reading challenge on coco-text,” in *2017 14th IAPR International Conference on Document Analysis and Recognition*, vol. 1, pp. 1435–1443, IEEE, 2017. **1**
- [18] V. Jain and E. Learned-Miller, “Fddb: A benchmark for face detection in unconstrained settings,” 2010. **1**
- [19] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, “Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks,” *Remote Sensing*, vol. 10, no. 1, p. 132, 2018. **1, 8**
- [20] Y. Jiang, X. Zhu, X. Wang, S. Yang, and Z. Luo, “R2cnn: Rotational region cnn for orientation robust scene text detection,” 2017. **1, 2, 7, 8**
- [21] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, “Arbitrary-oriented scene text detection via rotation proposals,” *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018. **1, 2, 6, 7, 8**
- [22] M. Liao, Z. Zhu, B. Shi, G.-s. Xia, and X. Bai, “Rotation-sensitive regression for oriented scene text detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5909–5918, 2018. **2**
- [23] X. Zhou, Y. Cong, W. He, Y. Wang, and J. Liang, “East: An efficient and accurate scene text detector,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. **2, 3**
- [24] C. Zhang, B. Liang, Z. Huang, M. En, J. Han, E. Ding, and X. Ding, “Look more than once: An accurate detector for text of arbitrary shapes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 10552–10561, 2019. **2**
- [25] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, “Deep direct regression for multi-oriented scene text detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 745–753, 2017. **2, 8**

- [26] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *The IEEE International Conference on Computer Vision*, October 2019. 2
- [27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014. 2
- [28] R. Girshick, "Fast r-cnn," in *The IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015. 2
- [29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015. 2, 8
- [30] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, pp. 379–387, 2016. 2
- [31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013. 2
- [32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016. 2, 8
- [33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision*, pp. 21–37, Springer, 2016. 2
- [34] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017. 2
- [35] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, no. 99, pp. 2999–3007, 2017. 2, 6, 8
- [36] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 2
- [37] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, 2018. 2, 7
- [38] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, *et al.*, "Hybrid task cascade for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4974–4983, 2019. 2
- [39] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4203–4212, 2018. 2, 7
- [40] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *The IEEE International Conference on Computer Vision*, October 2019. 2
- [41] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "Foveabox: Beyond anchor-based object detector," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019. 2
- [42] M. Liao, B. Shi, and X. Bai, "Textboxes++: A single-shot oriented scene text detector," *IEEE transactions on image processing*, vol. 27, no. 8, pp. 3676–3690, 2018. 3, 5
- [43] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Asian Conference on Computer Vision*, pp. 150–165, Springer, 2018. 3, 6, 7, 8
- [44] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, June 2019. 3, 6, 7, 8
- [45] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, "Scrnet: Towards more robust detection for small, cluttered and rotated objects," in *The IEEE International Conference on Computer Vision*, October 2019. 3, 6, 7, 8
- [46] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1962–1969, 2017. 5
- [47] Y. Liu, S. Zhang, L. Jin, L. Xie, Y. Wu, and Z. Wang, "Omnidirectional scene text detection with sequential-free box discretization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 5
- [48] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: a system for large-scale machine learning," in *OSDI*, vol. 16, pp. 265–283, 2016. 5
- [49] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *IEEE International Conference on Image Processing*, pp. 3735–3739, IEEE, 2015. 6
- [50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 6
- [51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009. 6
- [52] S. Bao, X. Zhong, R. Zhu, X. Zhang, Z. Li, and M. Li, "Single shot anchor refinement network for oriented object detection in optical remote sensing imagery," *IEEE Access*, vol. 7, pp. 87150–87161, 2019. 6, 7, 8

- [53] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "Fots: Fast oriented text spotting with a unified network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5676–5685, 2018. 8
- [54] L. Liu, Z. Pan, and B. Lei, "Learning a rotation invariant detector with rotatable bounding box," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 8
- [55] C. Li, C. Xu, Z. Cui, D. Wang, T. Zhang, and J. Yang, "Feature-attentioned object detection in remote sensing imagery," in *IEEE International Conference on Image Processing*, pp. 3886–3890, IEEE, 2019. 8