Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss

Xue Yang[1,2,3], Junchi Yan[1,2,*], Qi Ming[4], Wentao Wang[1], Xiaopeng Zhang[3], Qi Tian[3]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[3]Huawei Inc. [4]School of Automation, Beijing Institute of Technology
yangxue-2019-sjtu@sjtu.edu.cn

Abstract

Boundary discontinuity and its inconsistency to the final detection metric have been the bottleneck for rotating detection regression loss design. In this paper, we propose a novel regression loss based on Gaussian Wasserstein distance as a fundamental approach to solve the problem. Specifically, the rotated bounding box is converted to a 2-D Gaussian distribution, which enables to approximate the indifferentiable rotational IoU induced loss by the Gaussian Wasserstein distance (GWD) which can be learned efficiently by gradient back-propagation. GWD can still be informative for learning even there is no overlapping between two rotating bounding boxes which is often the case for small object detection. Thanks to its three unique properties, GWD can also elegantly solve the boundary discontinuity and square-like problem regardless how the bounding box is defined. Experiments on five datasets using different detectors show the effectiveness of our approach. Codes are available at `https://github.com/yangxue0827/RotationDetection`.

1. Introduction

Arbitrary-oriented objects are ubiquitous for detection across visual datasets, such as aerial images [65, 2, 8, 69], scene text [82, 31, 20, 35, 27], faces [47] and 3D objects [78], retail scenes [5, 41], etc. Compared with the large literature on horizontal object detection [13, 44, 28, 29, 6], research in oriented object detection is relatively in its earlier stage, with many open problems to solve.

The dominant line of works [2, 8, 69, 67] take a regression methodology to predict the rotation angle, which has achieved state-of-the-art performance. However, compared with traditional horizontal detectors, the angle regression model will bring new issues, as summarized as follows: i) the inconsistency between metric and loss, ii) boundary discontinuity, and iii) square-like problem. In fact, these issues remain open without a unified solution, and they can largely hurt the final performance especially at the boundary position, as shown in the left of Fig. 1. In this paper, we use a two-dimensional Gaussian distribution to model an arbitrary-oriented bounding box for object detection, and approximate the indifferentiable rotational Intersection over Union (IoU) induced loss between two boxes by calculating their Gaussian Wasserstein Distance (GWD) [3].

Figure 1: Comparison of the detection results between Smooth L1 loss-based (left) and the proposed GWD-based (right) detector.

GWD elegantly aligns model learning with the final detection accuracy metric, which has been a bottleneck and not achieved in existing rotation detectors. Our GWD based detectors are immune from both boundary discontinuity and square-like problem, and this immunity is independent with how the bounding box protocol is defined, as shown on the right of Fig. 1. The highlights of this paper are four-folds:

i) We summarize three flaws in state-of-the-art rotation detectors, i.e. inconsistency between metric and loss, boundary discontinuity, and square-like problem, due to their regression based angle prediction nature.

ii) We propose to model the rotating bounding box distance by Gaussian Wasserstein Distance (GWD) which leads to an approximate and differentiable IoU induced loss. It resolves the loss inconsistency by aligning model learning with accuracy metric and thus naturally improves the model.

iii) Our GWD-based loss can elegantly resolve boundary

*Corresponding author is Junchi Yan.

discontinuity and square-like problem, regardless how the rotating bounding box is defined. In contrast, the design of most peer works [66, 64] are coupled with the parameterization of bounding box.

iv) Extensive experimental results on five public datasets and two popular detectors show the effectiveness of our approach. Source code will be made public available.

2. Related Work

In this paper, we mainly discuss the related work on rotating object detection. Readers are referred to [13, 44, 28, 29] for more comprehensive literature review on horizontal object detection.

Rotated object detection. As an emerging direction, advance in this area try to extend classical horizontal detectors to the rotation case by adopting the rotated bounding boxes. Compared with the few works [66] that treat the rotation detection tasks an angle classification problem, regression based detectors still dominate which have been applied in different applications. For aerial images, ICN [2], ROI-Transformer [8], SCRDet [69] and Gliding Vertex [62] are two-stage representative methods whose pipeline comprises of object localization and classification, while DRN [41], R^{3}Det [67] and RSDet [42] are single-stage methods. For scene text detection, RRPN [35] employ rotated RPN to generate rotated proposals and further perform rotated bounding box regression. TextBoxes++ [26] adopts vertex regression on SSD. RRD [27] further improves TextBoxes++ by decoupling classification and bounding box regression on rotation-invariant and rotation sensitive features, respectively. We discuss the specific challenges in existing regressors for rotation detection.

Boundary discontinuity and square-like problems. Due to the periodicity of angle parameters and the diversity of bounding box definitions, regression-based rotation detectors often suffer from boundary discontinuity and square-like problem. Many existing methods try to solve part of the above problems from different perspectives. For instance, SCRDet [69] and RSDet [42] propose IoU-smooth L1 loss and modulated loss to smooth the the boundary loss jump. CSL [66] transforms angular prediction from a regression problem to a classification one. DCL [64] further solves square-like object detection problem introduced by the long edge definition, which refers to rotation insensitivity issue for instances that are approximately in square shape, which will be detailed in Sec. 3.

Approximate differentiable rotating IoU loss. It has been shown in classic horizontal detectors that the use of IoU induced loss e.g. GIoU [45], DIoU [79] can ensure the consistency of the final detection metric and loss. However, these IoU loss cannot be applied directly in rotation detection because the rotating IoU is indifferentiable. Many efforts have been made to finding an approximate IoU loss for gradient computing. PIoU [5] is realized by simply counting the number of pixels. To tackle the uncertainty of convex caused by rotation, [78] proposes a projection operation to estimate the intersection area. SCRDet [69] combines IoU and smooth L1 loss to develop an IoU-smooth L1 loss, which partly circumvents the need for differentiable rotating IoU loss.

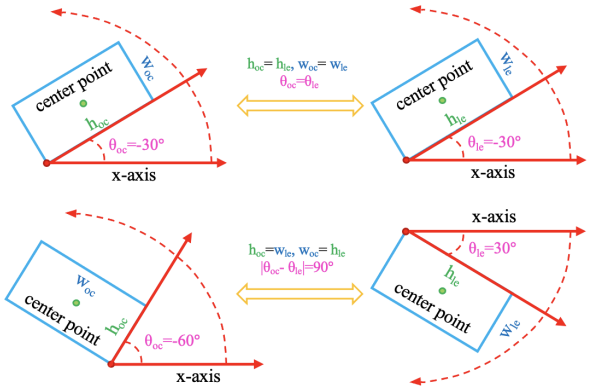


Figure 2: Two definitions of bounding boxes. **Left:** OpenCV Definition D_{oc}, **Right:** Long Edge Definition D_{le}.

So far, there exists no truly unified solution to all the above problems which are in fact interleaved to each other. Our method addresses all these issues in a unified manner. It is also decoupled from the specific definition of bounding box. All these merits make our approach elegant and effective.

3. Rotated Object Regression Detector Revisit

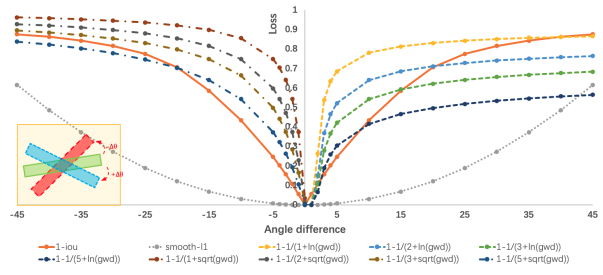
To motivate this work, in this section, we introduce and analyze some deficiencies in state-of-the-art rotating detectors, which are mostly based on angle regression.

3.1. Bounding Box Definition

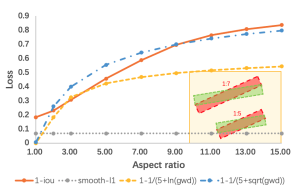
Fig. 2 gives two popular definitions for parameterizing rotating bounding box based angles: OpenCV protocol denoted by D_{oc}, and long edge definition by D_{le}. Note $\theta \in [-90^\circ, 0^\circ)$ of the former denotes the acute or right angle between the h_{oc} of bounding box and x-axis. In contrast, $\theta \in [-90^\circ, 90^\circ)$ of the latter definition is the angle between the long edge h_{le} of bounding box and x-axis. The two kinds of parameterization can be converted to each other:

$$D_{le}(h_{le}, w_{le}, \theta_{le}) = \begin{cases} D_{oc}(h_{oc}, w_{oc}, \theta_{oc}), & h_{oc} \geq w_{oc} \\ D_{oc}(w_{oc}, h_{oc}, \theta_{oc} + 90^\circ), & otherwise \end{cases}$$

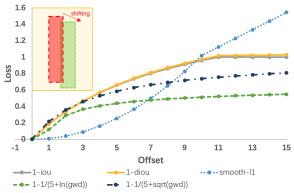
$$D_{oc}(h_{oc}, w_{oc}, \theta_{oc}) = \begin{cases} D_{le}(h_{le}, w_{le}, \theta_{le}), & \theta_{le} \in [-90^\circ, 0) \\ D_{le}(w_{le}, h_{le}, \theta_{le} - 90), & otherwise \end{cases}$$



(a) Angle difference case



(b) Aspect ratio case



(c) Center shifting case

Figure 3: Behavior comparison of different loss in different cases.

The main difference refers to the edge and angle (h, w, θ): when the same bounding box takes different representations by the two definitions, the order of the edges is exchanged and the angle difference is $90°$.

In many works, the pipeline design are tightly coupled with the choice of the bounding box definition to avoid specific problems: SCRDet [69], R^{3}Det [67] are based on D_{oc} to avoid the square-like problem, while CSL [66], DCL [64] resort to D_{le} to avoid the exchangeability of edges (EoE).

3.2. Inconsistency between Metric and Loss

Intersection over Union (IoU) has been the standard metric for both horizontal detection and rotation detection. However, there is an inconsistency between the metric and regression loss (e.g. l_n-norms), that is, a smaller training loss cannot guarantee a higher performance, which has been extensively discussed in horizontal detection [45, 79]. This misalignment becomes more prominent in rotating object detection due to the introduction of angle parameter in regression based models. To illustrate this, we use Fig. 3 to compare IoU induced loss and smooth L1 loss [13]:

Case 1: Fig. 3a depicts the relation between angle difference and loss functions. Though they all bear monotonicity, only smooth L1 curve is convex while the others are not.

Case 2: Fig. 3b shows the changes of the two loss functions under different aspect ratio conditions. It can be seen that the smooth L1 loss of the two bounding box are constant (mainly from the angle difference), but the IoU loss will change drastically as the aspect ratio varies.

Case 3: Fig. 3c explores the impact of center point shifting on different loss functions. Similarly, despite the same monotonicity, there is no high degree of consistency.

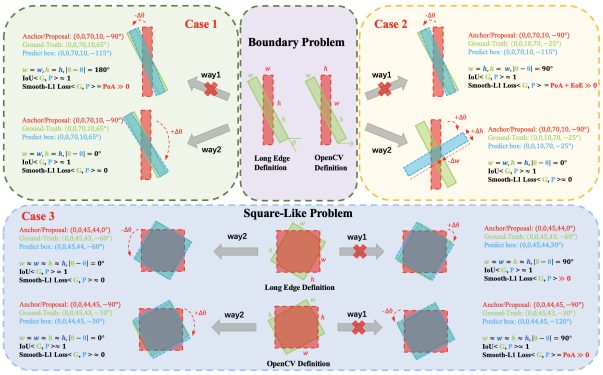


Figure 4: Boundary discontinuity under two bounding box definitions (top), and illustration of the square-like problem (bottom).

Seeing the above flaws of classic smooth L1 loss, IoU-induced loss has become recently popular for horizontal detection e.g. GIoU [45], DIoU [79]. It can help fill the gap between metric and regression loss for rotating object detection. However, different from horizontal detection, the IoU of two rotating boxes is indifferentiable for learning. In this paper, we propose a differentiable loss based on Wasserstein distance of two rotating boxes to replace the hard IoU loss. It is worth mentioning that the Wasserstein distance function has some unique properties to solve boundary discontinuity and square-like problem, which will be detailed later.

3.3. Boundary Discontinuity and Square-Like Problem

As a standing issue for regression-based rotation detectors, the boundary discontinuity [69, 66] in general refers to the sharp loss increase at the boundary induced by the angle and edge parameterization.

Specifically, **Case 1-2** in Fig. 4 summarize the boundary problem. Take **Case 2** as an example, we assume that there is a red anchor/proposal $(0, 0, 70, 10, -90°)$ and a green ground truth $(0, 0, 10, 70, -25°)$ at the boundary position[1], both of which are defined in OpenCV definition D_{oc}. The upper right corner of Fig. 4 shows two ways to regress from anchor/proposal to ground truth. The **way1** achieves the goal by only rotating anchor/proposal by an angle counterclockwise, but a very large smooth L1 loss occurs at this time due to the periodicity of angle (PoA) and the exchangeability of edges (EoE). As discussed in CSL [66], this is because the result of the prediction box $(0, 0, 70, 10, -115°)$

[1]The angle of the bounding box is close to the maximum and minimum values of the angle range. For more clearly visualization, the ground truth has been rendered with a larger angle in Fig. 4.

is outside the defined range. As a result, the model has to make predictions in other complex regression forms, such as rotating anchor/proposal by an large angle clockwise to the blue box while scaling w and h (**way2** in **Case 2**). A similar problem (only PoA) also occurs in the long edge definition D_{le}, as shown in **Case 1**.

In fact, when the predefined anchor/proposal and ground truth are not in the boundary position, **way1** will not produce a large loss. Therefore, there exists inconsistency between the boundary position and the non-boundary position regression, which makes the model very confused about in which way it should perform regression. Since non-boundary cases account for the majority, the regression results of models, especially those with weaker learning capacity, are fragile in boundary cases, as shown in the left of Fig. 1.

In addition, there is also a square-like object detection problem in the D_{le}-based method [64]. First of all, the D_{le} cannot uniquely define a square bounding box. For square-like objects[2], D_{le}-based method will encounter high IoU but high loss value similar to the boundary discontinuity, as shown by the upper part of **Case 3** in Fig. 4. In **way1**, the red anchor/proposal $(0, 0, 45, 44, 0°)$ rotates a small angle clockwise to get the blue prediction box. The IoU of ground truth $(0, 0, 45, 43, -60°)$ and the prediction box $(0, 0, 45, 44, 30°)$ is close to 1, but the regression loss is high due to the inconsistency of angle parameters. Therefore, the model will rotate a larger angle counterclockwise to make predictions, as described by **way2**. The reason for the square-like problem in D_{le}-based method is not the above-mentioned PoA and EoE, but the inconsistency of evaluation metric and loss. In contrast, the negative impact of EoE will be weakened when we use D_{oc}-based method to detect square-like objects, as shown in the comparison between **Case 2** and the lower part of **Case 3**. Therefore, there is no square-like problem in the D_{oc}-based method.

Recent methods start to address these issues. SCRDet [69] combines IoU and smooth L1 loss to propose a IoU-smooth L1 loss, which does not require the rotating IoU being differentiable. It also solves the problem of inconsistency between loss and metric by eliminating the discontinuity of loss at the boundary. However, SCRDet still needs to determine whether the predicted bounding box result conforms to the current bounding box definition method before calculating the IoU. In addition, the gradient direction of IoU-Smooth L1 Loss is still dominated by smooth L1 loss. RSDet [42] devises modulated loss to smooth the loss mutation at the boundary, but it needs to calculate the loss of as many parameter combinations as possible. CSL [66] transforms angular prediction from a regression problem to a classification problem. CSL needs to carefully design their method according to the bounding box definition (D_{le}), and is limited by the classification granularity with theoretical limitation for high-precision angle prediction. On the basis of CSL, DCL [64] further solves the problem of square-like object detection introduced by D_{le}.

[2]Many instances are in square shape. For instance, two categories of storage-tank (ST) and roundabout (RA) in DOTA dataset.

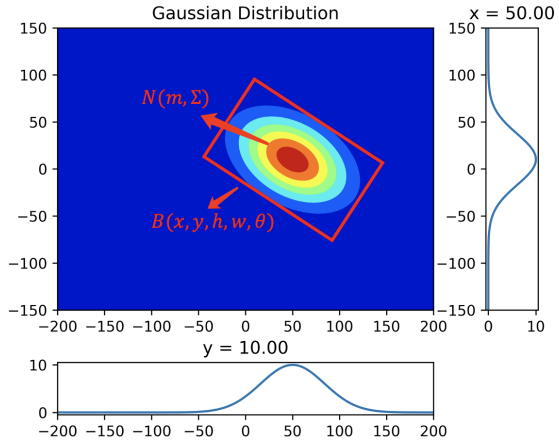


Figure 5: A schematic diagram of modeling a rotating bounding box by a two-dimensional Gaussian distribution.

4. The Proposed Method

In this section we introduce a new rotating object detector whose regression loss fulfills the following requirements:

Requirement 1: highly consistent with the IoU induced metrics (which also solves the square-like object problem);

Requirement 2: differentiable allowing for direct learning;

Requirement 3: smooth at angle boundary case.

4.1. Wasserstein Distance for Rotating Box

Most of the IoU-based loss can be considered as a distance function. Inspired by this, we propose a new regression loss based on Wasserstein distance. First, we convert a rotating bounding box $\mathcal{B}(x, y, h, w, \theta)$ into a 2-D Gaussian distribution $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ (see Fig. 5) by the following formula:

$$\begin{aligned}\mathbf{\Sigma}^{1/2} =& \mathbf{R}\mathbf{S}\mathbf{R}^{\top} \\ =& \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \frac{w}{2} & 0 \\ 0 & \frac{h}{2} \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \\ =& \begin{pmatrix} \frac{w}{2}\cos^2\theta + \frac{h}{2}\sin^2\theta & \frac{w-h}{2}\cos\theta\sin\theta \\ \frac{w-h}{2}\cos\theta\sin\theta & \frac{w}{2}\cos^2\theta + \frac{h}{2}\sin^2\theta \end{pmatrix} \\ \mathbf{m} =& (x, y) \end{aligned} \tag{1}$$

where $\mathbf{R}$ represents the rotation matrix, and S represents the diagonal matrix of eigenvalues.

The Wasserstein distance $\mathbf{W}$ between two probability measures μ and ν on $\mathbb{R}^n$ expressed as [3]:

$$\mathbf{W}(\mu;\nu) := \inf \mathbb{E}(\|\mathbf{X}-\mathbf{Y}\|_2^2)^{1/2} \tag{2}$$

where the inferior runs over all random vectors $(\mathbf{X},\mathbf{Y})$ of $\mathbb{R}^n \times \mathbb{R}^n$ with $\mathbf{X} \sim \mu$ and $\mathbf{Y} \sim \nu$. It turns out that we have $d := \mathbf{W}(\mathcal{N}(\mathbf{m}_1,\mathbf{\Sigma}_1);\mathcal{N}(\mathbf{m}_2,\mathbf{\Sigma}_2))$ and it writes as:

$$d^2 = \|\mathbf{m}_1-\mathbf{m}_2\|_2^2 + \mathbf{Tr}\left(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2-2(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}\right) \tag{3}$$

This formula has interested several works [14, 40, 22, 9]. Note in particular we have:

$$\mathbf{Tr}\left((\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}\right) = \mathbf{Tr}\left((\mathbf{\Sigma}_2^{1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{1/2})^{1/2}\right) \tag{4}$$

In the commutative case $\mathbf{\Sigma}_1\mathbf{\Sigma}_2 = \mathbf{\Sigma}_2\mathbf{\Sigma}_1$, Eq. 3 becomes:

$$\begin{aligned} d^2 =& \|\mathbf{m}_1-\mathbf{m}_2\|_2^2 + \|\mathbf{\Sigma}_1^{1/2}-\mathbf{\Sigma}_2^{1/2}\|_F^2 \\ =& (x_1-x_2)^2+(y_1-y_2)^2+\frac{(w_1-w_2)^2+(h_1-h_2)^2}{4} \\ =& l_2\text{-norm}\left(\left[x_1,y_1,\frac{w_1}{2},\frac{h_1}{2}\right]^\top,\left[x_2,y_2,\frac{w_2}{2},\frac{h_2}{2}\right]^\top\right) \end{aligned} \tag{5}$$

where $\|\|_F$ is the Frobenius norm. Note that both boxes are horizontal here, and Eq. 5 is approximately equivalent to the l_2-norm loss (note the additional denominator of 2 for w and h), which is consistent with the loss commonly used in horizontal detection. This also partly proves the correctness of using Wasserstein distance as the regression loss. See appendix for the detailed proof [3] of Eq. 3.

4.2. Gaussian Wasserstein Distance Regression Loss

Note that GWD alone can be sensitive to large errors. We perform a nonlinear transformation f and then convert GWD into an affinity measure $\frac{1}{\tau+f(d^2)}$ similar to IoU between two bounding boxes. Then we follow the standard IoU based loss form in detection literature [45, 79], as written by:

$$L_{gwd} = 1-\frac{1}{\tau+f(d^2)}, \quad \tau \geq 1 \tag{6}$$

where $f(\cdot)$ denotes a non-linear function to transform the Wasserstein distance d^2 to make the loss more smooth and expressive. The hyperparameter τ modulates the entire loss.

Fig. 3a plots the function curve under different different combinations of $f(\cdot)$ and τ. Compared with the smooth L1 loss, the curve of Eq. 6 is more consistent with the IoU loss curve. Furthermore, we can find in Fig. 3c that GWD still can measure the distance between two non-overlapping bounding boxes (IoU=0), which is exactly the problem that GIoU and DIoU try to solve in horizontal detection. However, they cannot be applied for rotating detection.

Obviously, GWD has met the first two requirements in terms of consistency and differentiability with IoU loss. To analyze **Requirement 3**, we first give basic properties of Eq. 1:

Property 1: $\mathbf{\Sigma}^{1/2}(w,h,\theta) = \mathbf{\Sigma}^{1/2}(h,w,\theta-\frac{\pi}{2})$;

Property 2: $\mathbf{\Sigma}^{1/2}(w,h,\theta) = \mathbf{\Sigma}^{1/2}(w,h,\theta-\pi)$;

Property 3: $\mathbf{\Sigma}^{1/2}(w,h,\theta) \approx \mathbf{\Sigma}^{1/2}(w,h,\theta-\frac{\pi}{2})$, if $w \approx h$.

From the two bounding box definitions recall that the conversion between two definitions is, the two sides are exchanged and the angle difference is 90°. Many methods are designated inherently according to the choice of definition in advance to solve some problems, such as D_{le} for EoE and D_{oc} for square-like problem. It is interesting to note that according to **Property 1**, definition D_{oc} and D_{le} are equivalent for the GWD-based loss, which makes our method free from the choice of box definitions. This does not mean that the final performance of the two definition methods will be the same. Different factors such as angle definition and angle regression range will still cause differences in model learning, but the GWD-based method does not need to bind a certain definition method to solve the problem.

GWD can also help resolve the boundary problem and square-like problem. The prediction box and ground truth in **way1** of **Case 1** in Fig. 4 satisfy the following relation: $x_p = x_{gt}, y_p = y_{gt}, w_p = h_{gt}, h_p = w_{gt}, \theta_p = \theta_{gt} - \frac{\pi}{2}$. According to **Property 1**, the Gaussian distribution corresponding to these two boxes are the same (in the sense of same mean $\mathbf{m}$ and covariance $\mathbf{\Sigma}$), so it naturally eliminates the ambiguity in box representation. Similarly, according to **Properties 2-3**, the ground truth and prediction box in **way1** of **Case 1** and **Case 3** in Fig. 4 are also the same or nearly the same (note the approximate equal symbol for $w \approx h$ for square-like boxes) Gaussian distribution. Through the above analysis, we know GWD meets **Requirement 3**.

Overall, GWD is a unified solution to all the requirements and its advantages in rotating detection can be summarized:

i) GWD makes the two bounding box definition methods equivalent, which enables our method to achieve significant improvement regardless how the bounding box is defined.

ii) GWD is a differentiable IoU loss approximation for rotating bounding box, which maintains a high consistency with the detection metric. GWD can also measure the distance between non-overlapping rotating bounding boxes and has properties similar to GIoU and DIoU for the horizontal case.

iii) GWD inherently avoids the interference of boundary discontinuity and square-like problem, so that the model can learn in more diverse forms of regression, eliminate the inconsistency of regression under boundary and non-boundary positions, and reduce the learning cost.

$1-\frac{1}{(\tau+f(d^2))}$	$\tau=1$	$\tau=2$	$\tau=3$	$\tau=5$	d^2
$f(\cdot)=sqrt$	68.56	**68.93**	68.37	67.77	49.11
$f(\cdot)=\log$	67.87	68.09	67.48	66.49	

Table 1: Ablation test of GWD-based regression loss form and hyperparameter on DOTA. The based detector is RetinaNet.

Method	Box Def.	Reg. Loss	Dataset	Data Aug.	mAP_{50}
RetinaNet	D_{oc}	Smooth L1	HRSC2016	R+F+G	84.28
	D_{oc}	GWD			**85.56** (+1.28)
	D_{oc}	Smooth L1	UCAS-AOD		94.56
	D_{oc}	GWD			**95.44** (+0.88)
	D_{oc}	Smooth L1	DOTA	F	65.73
	D_{oc}	GWD			**68.93** (+3.20)
	D_{le}	Smooth L1			64.17
	D_{le}	GWD			**66.31** (+2.14)
R^3Det	D_{oc}	Smooth L1			70.66
	D_{oc}	GWD			**71.56** (+0.90)

Table 2: Ablation study for GWD on three datasets. 'R', 'F' and 'G' indicate random rotation, flipping, and graying, respectively.

4.3. Overall Loss Function Design

In line with [66, 64, 67], we use the one-stage detector RetinaNet [29] as the baseline. Rotated rectangle is represented by five parameters (x, y, w, h, θ). In our experiments we mainly follow D_{oc}, and the regression equation is as follows:

$$\begin{aligned} t_x &= (x-x_a)/w_a, t_y = (y-y_a)/h_a \\ t_w &= \log(w/w_a), t_h = \log(h/h_a), t_\theta = \theta - \theta_a \\ t_x^* &= (x^*-x_a)/w_a, t_y^* = (y^*-y_a)/h_a \\ t_w^* &= \log(w^*/w_a), t_h^* = \log(h^*/h_a), t_\theta^* = \theta^* - \theta_a \end{aligned} \tag{7}$$

where x, y, w, h, θ denote the box's center coordinates, width, height and angle, respectively. Variables x, x_a, x^* are for the ground-truth box, anchor box, and predicted box, respectively (likewise for y, w, h, θ). The multi-task loss is:

$$L = \frac{\lambda_1}{N}\sum_{n=1}^{N} obj_n \cdot L_{gwd}(b_n, gt_n) + \frac{\lambda_2}{N}\sum_{n=1}^{N} L_{cls}(p_n, t_n) \tag{8}$$

where N indicates the number of anchors, obj_n is a binary value ($obj_n = 1$ for foreground and $obj_n = 0$ for background, no regression for background). b_n denotes the n-th predicted bounding box, gt_n is the n-th target ground-truth. t_n represents the label of n-th object, p_n is the n-th probability distribution of various classes calculated by sigmoid function. The hyper-parameter λ_1, λ_2 control the trade-off and are set to $\{1, 2\}$ by default. The classification loss L_{cls} is set as the focal loss [29].

5. Experiments

We use Tensorflow [1] for implementation on a server with Tesla V100 and 32G memory.

Method	Reg. Loss	Dataset	Data Aug.	Recall	Precision	Hmean
RetinaNet	Smooth L1	MLT	F	37.88	67.07	48.42
	GWD			44.01	71.83	**54.58** (+6.16)
	Smooth L1	ICDAR2015		71.55	68.10	69.78
	GWD			73.95	74.64	**74.29** (+4.51)
	Smooth L1		R+F	69.43	81.15	74.83
	GWD			72.17	80.59	**76.15** (+1.32)
R^3Det	Smooth L1		F	69.09	80.30	74.28
	GWD			70.00	82.15	**75.59** (+1.31)
	Smooth L1		R+F	71.69	79.80	75.53
	GWD			73.95	80.50	**77.09** (+1.56)

Table 3: Ablation study for GWD on two scene text datasets.

5.1. Datasets and Implementation Details

DOTA [57] is comprised of 2,806 large aerial images from different sensors and platforms. Objects in DOTA exhibit a wide variety of scales, orientations, and shapes. These images are then annotated by experts using 15 object categories. 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. The fully annotated DOTA benchmark contains 188,282 instances, each of which is labeled by an arbitrary quadrilateral. Half of the original images are randomly selected as the training set, 1/6 as the validation set, and 1/3 as the testing set. We divide the images into 600×600 subimages with an overlap of 150 pixels and scale it to 800×800. With all these processes, we obtain about 20,000 training and 7,000 validation patches.

UCAS-AOD [83] contains 1,510 aerial images of about $659 \times 1,280$ pixels, with 2 categories of 14,596 instances. In line with [2, 57], we sample 1,110 images for training and 400 for testing.

HRSC2016 [33] contains images from two scenarios including ships on sea and ships close inshore. The training, validation and test set include 436, 181 and 444 images, respectively.

ICDAR2015 [21] is commonly used for oriented scene text detection and spotting. This dataset includes 1,000 training images and 500 testing images.

ICDAR 2017 MLT [38] is a multi-lingual text dataset, which includes 7,200 training images, 1,800 validation images and 9,000 testing images. The dataset is composed of complete scene images in 9 languages, and text regions in this dataset can be in arbitrary orientations, being more diverse and challenging.

Experiments are initialized by ResNet50 [16] by default unless otherwise specified. We perform experiments on three aerial benchmarks and two scene text benchmarks to verify the generality of our techniques. Weight decay and momentum are set 0.0001 and 0.9, respectively. We employ

ID	MOETHOD	BACKBONE	SCHED.	DA	MS	MSC	SWA	ME	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	MAP$_{50}$
#1		R-50	20				✓		88.49	77.88	44.07	66.08	71.92	62.56	77.94	89.75	81.43	79.64	52.30	63.52	60.25	66.51	51.63	68.93
#2									88.60	78.59	44.10	67.24	70.77	62.54	79.78	88.86	81.92	80.46	57.44	64.02	62.64	66.52	55.29	69.92
#3		R-152	40	✓					89.06	83.48	49.84	65.34	74.64	67.63	82.39	88.39	84.19	84.80	63.74	61.32	66.47	70.94	67.52	73.32
#4				✓	✓				87.47	83.77	52.30	68.24	73.24	65.14	80.18	89.63	84.39	85.53	65.79	66.02	69.57	72.21	69.79	74.22
#5	RETINANET-GWD		60	✓					88.88	80.47	52.94	63.85	76.95	70.28	83.56	88.54	83.51	84.94	61.24	65.13	65.45	71.69	73.90	74.09
#6				✓	✓				87.12	81.64	54.79	68.74	76.17	68.39	83.93	89.06	84.51	85.99	63.33	66.68	72.60	70.63	74.17	75.18
#7				✓	✓	✓			86.14	81.59	55.33	75.57	74.20	67.34	81.75	87.48	82.80	85.46	69.47	67.20	70.97	70.91	74.07	75.35
#8				✓	✓		✓		87.63	84.32	54.83	69.99	76.17	70.12	83.13	88.96	83.19	86.06	67.72	66.17	73.47	74.57	72.80	75.94
#9				✓	✓	✓	✓		86.96	83.88	54.36	77.53	74.41	68.48	80.34	86.62	83.41	85.55	73.47	67.77	72.57	75.76	73.40	76.30
#10		–	–	✓	✓	✓	✓	✓	**89.06**	**84.32**	**55.33**	**77.53**	**76.95**	**70.28**	**83.95**	**89.75**	**84.51**	**86.06**	**73.47**	**67.77**	**72.60**	**75.76**	**74.17**	**77.43**
#11		R-101	30	✓					89.59	81.18	52.89	70.37	77.73	82.42	86.99	89.31	83.06	85.97	64.07	65.14	68.05	70.95	58.45	75.08
#12				✓	✓				89.64	81.70	52.52	72.96	76.02	82.60	87.17	89.57	81.25	86.09	62.24	65.74	68.05	74.96	64.38	75.66
#13				✓	✓		✓		89.66	82.11	52.74	71.64	75.95	83.09	86.97	89.28	85.04	86.17	65.52	63.29	72.18	74.88	63.17	76.11
#14				✓	✓	✓			89.56	81.23	53.38	79.38	75.12	82.14	86.86	88.87	81.21	86.28	65.36	65.06	72.88	73.04	62.97	76.22
#15				✓	✓	✓	✓		89.33	80.86	53.28	78.29	75.40	82.69	87.09	89.35	82.64	86.41	69.85	64.71	74.19	76.18	59.85	76.67
#16		R-152		✓					89.51	82.68	51.92	69.51	78.97	83.38	87.53	89.67	85.65	86.17	63.90	67.44	68.27	76.43	64.22	76.35
#17				✓	✓				89.55	82.28	52.39	68.30	77.86	83.40	87.48	89.56	84.27	86.14	65.38	63.25	71.33	72.36	69.21	76.18
#18				✓	✓	✓			89.62	82.27	52.35	77.30	76.95	82.53	87.20	89.08	84.58	86.21	65.21	64.46	74.99	76.30	65.19	76.95
#19		R-18	40	✓					86.63	80.12	51.98	49.67	75.73	77.54	86.10	90.05	83.22	82.31	56.05	58.86	63.30	69.06	55.07	71.05
#20				✓	✓				87.88	81.73	51.76	69.21	73.78	77.78	86.46	90.05	84.47	84.33	59.82	59.74	66.54	69.15	60.42	73.54
#21				✓	✓		✓		88.94	84.10	53.04	67.78	75.29	79.21	86.89	89.90	86.43	84.30	63.22	59.96	67.16	70.55	64.39	74.74
#22	R^{3}DET-GWD			✓	✓	✓			87.27	82.59	51.90	76.58	72.74	77.04	85.59	89.18	83.91	84.81	63.34	59.46	66.41	69.79	59.03	73.98
#23				✓	✓	✓	✓		88.38	84.75	52.63	77.35	74.29	78.53	86.32	89.12	85.73	85.13	67.84	59.48	66.88	71.59	62.58	75.37
#24		R-50	60	✓					88.82	82.94	55.63	72.75	78.52	83.10	87.46	90.21	86.36	85.44	64.70	61.41	73.46	76.94	57.38	76.34
#25				✓	✓				89.09	84.13	55.77	74.48	77.71	82.99	87.57	89.46	84.89	85.67	66.09	64.17	75.13	75.35	62.78	77.02
#26				✓	✓		✓		89.04	84.99	57.14	76.13	77.79	84.03	87.70	89.53	83.83	85.64	69.60	63.75	76.10	79.22	67.80	78.15
#27				✓	✓	✓			88.89	83.58	55.54	80.46	76.86	83.07	86.85	89.09	83.09	86.17	71.38	64.93	76.21	73.23	64.39	77.58
#28				✓	✓	✓	✓		88.43	84.33	56.91	82.19	76.69	83.23	86.78	88.90	83.93	85.73	72.07	65.67	76.76	78.37	65.31	78.35
#29		R-152		✓					88.74	82.63	54.88	70.11	78.87	84.59	87.37	89.81	84.79	86.47	66.58	64.11	75.31	78.43	70.87	77.57
#30				✓	✓				89.59	84.19	56.53	75.69	77.67	84.48	87.52	90.05	84.29	86.85	68.61	64.73	76.59	77.92	71.88	78.44
#31				✓	✓		✓		89.59	82.96	58.83	75.04	77.63	84.83	87.31	89.89	86.54	86.82	69.45	65.94	76.55	77.50	74.92	78.92
#32				✓	✓	✓			88.99	82.26	56.62	81.40	77.04	83.90	86.56	88.97	83.63	86.48	70.45	65.58	76.41	77.30	69.21	78.32
#33				✓	✓	✓	✓		89.28	83.70	59.26	79.85	76.42	83.87	86.53	89.06	85.53	86.50	73.04	67.56	76.92	77.09	71.58	79.08
#34		–	–	✓	✓	✓	✓	✓	**89.66**	**84.99**	**59.26**	**82.19**	**78.97**	**84.83**	**87.70**	**90.21**	**86.54**	**86.85**	**73.04**	**67.56**	**76.92**	**79.22**	**74.92**	**80.19**
#35	–	–	–	✓	✓	✓	✓	✓	**89.66**	**84.99**	**59.26**	**82.19**	**78.97**	**84.83**	**87.70**	**90.21**	**86.54**	**86.85**	**73.47**	**67.77**	**76.92**	**79.22**	**74.92**	**80.23**

Table 4: Ablation experiment of training strategies and tricks. R-101 denotes ResNet-101 (likewise for R-18, R-50, R-152). MS, MSC, SWA, and ME represent data augmentation, multi-scale training and testing, stochastic weights averaging, multi-scale image cropping, and model ensemble, respectively.

METHOD	REG. LOSS	AP$_{50}$	AP$_{60}$	AP$_{75}$	AP$_{85}$	AP$_{50:95}$
RETINANET	SMOOTH L1	84.28	74.74	48.42	12.56	47.76
	GWD	**85.56**	**84.04**	**60.31**	**17.14**	**52.89 +(5.13)**
R^{3}DET	SMOOTH L1	88.52	79.01	43.42	4.58	46.18
	GWD	**89.43**	**88.89**	**65.88**	**15.02**	**56.07 +(9.89)**

Table 5: High-precision detection experiment on HRSC206 data set. The image resolution is 512, and data augmentation is used.

MomentumOptimizer over 8 GPUs with a total of 8 images per mini-batch (1 image per GPU). All the used datasets are trained by 20 epochs in total, and learning rate is reduced tenfold at 12 epochs and 16 epochs, respectively. The initial learning rates for RetinaNet is 5e-4. The number of image iterations per epoch for DOTA, UCAS-AOD, HRSC2016, ICDAR2015, and MLT are 54k, 5k, 10k, 10k, 10k and 10k respectively, and increase exponentially if data augmentation and multi-scale training are used.

5.2. Ablation Study

Ablation test of GWD-based regression loss form and hyperparameter: Tab. 1 compares two different forms of GWD-based loss. The performance of directly using GWD (d^2) as the regression loss is extremely poor, only 49.11%, due to its rapid growth trend. In other words, the regression loss d^2 is too sensitive to large errors. In contrast, Eq. 6 achieves a significant improvement by fitting IoU loss. Eq. 6 introduces two new hyperparameters, the non-linear function $f(\cdot)$ to transform the Wasserstein distance, and the constant τ to modulate the entire loss. From Tab. 1, the overall performance of using $sqrt$ outperforms that using $\log$, about 0.98±0.3% higher. Indeed, in Fig. 3, the $sqrt$-based GWD curve is closer to the IoU loss curve than the $\log$-based GWD curve. For $f(\cdot) = sqrt$ with $\tau = 2$, the model achieves the best performance, about 68.93%. All the subsequent experiments follow this setting for hyperparameters unless otherwise specified.

Ablation test with different rotating box definitions: As mentioned above, definition D_{oc} and D_{le} are equivalent for the GWD-based loss according to **Property 1**, which makes our method free from the choice of box definitions. This does not mean that the final performance of the two definition methods will be the same, but that the GWD-based method does not need to bind a certain definition method to solve the boundary problem or square-like problem. Tab. 2 compares the performance of RetinaNet under different regression loss on DOTA, and both rotating box definitions: D_{le} and D_{oc} are tested. For the smooth L1 loss, the accuracy of D_{le}-based method is 1.56% lower than D_{le}-based, at 64.17% and 65.73%, respectively. GWD-based method does not need to be coupled with a certain definition to solve boundary problem or square-like problem, it has increased by 2.14% and 3.20% under above two definitions.

Ablation test across datasets and detectors: We use

Base Detector	Method	Box Def.	IML	BD		SLP	tranval/test									train/val		
				EoE	PoA		BR†	SV†	LV†	SH†	HA†	ST‡	RA‡	7-MAP$_{50}$	MAP$_{50}$	MAP$_{50}$	MAP$_{75}$	MAP$_{50:95}$
RetinaNet	-	D_{oc}	✓	✓	✓	×	42.17	65.93	51.11	72.61	53.24	78.38	62.00	60.78	65.73	64.70	32.31	34.50
	-	D_{le}	✓	✓	✓	✓	38.31	60.48	49.77	68.29	51.28	78.60	60.02	58.11	64.17	62.21	26.06	31.49
	IoU-Smooth L1 Loss	D_{oc}	✓	×	×	×	**44.32**	63.03	51.25	72.78	56.21	77.98	63.22	61.26	66.99	64.61	34.17	36.23
	CSL	D_{le}	✓	×	×	✓	42.25	68.28	54.51	72.85	53.10	75.59	58.99	60.80	67.38	64.40	32.58	35.04
	DCL (BCL)	D_{le}	✓	×	×	×	41.40	65.82	56.27	73.80	54.30	**79.02**	60.25	61.55	**67.39**	**65.93**	35.66	**36.71**
	GWD	D_{oc}	×	×	×	×	**44.07**	**71.92**	**62.56**	**77.94**	**60.25**	**79.64**	**63.52**	**65.70**	**68.93**	**65.44**	**38.68**	**38.71**
R^{3}Det	-	D_{oc}	✓	✓	✓	×	44.15	**75.09**	72.88	**86.04**	56.49	82.53	61.01	68.31	70.66		–	
	DCL (BCL)	D_{le}	✓	×	×	×	**46.84**	74.87	**74.96**	85.70	**57.72**	**84.06**	**63.77**	**69.70**	**71.21**			
	GWD	D_{oc}	×	×	×	×	**46.73**	**75.84**	**78.00**	**86.71**	**62.69**	**83.09**	**61.12**	**70.60**	**71.56**			

Table 6: Comparison between different solutions for inconsistency between metric and loss (IML), boundary discontinuity (BD) and square-like problem (SLP) on DOTA dataset. The ✓ indicates that the method has corresponding problem. † and ‡ represent the large aspect ratio object and the square-like object, respectively. The bold **red** and **blue** fonts indicate the top two performances respectively.

	Method	Backbone	MS	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	MAP$_{50}$
Two-stage methods	FR-O [57]	R-101		79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.4	52.52	46.69	44.80	46.30	52.93
	ICN [2]	R-101	✓	81.40	74.30	47.70	70.30	64.90	67.80	70.00	90.80	79.10	78.20	53.60	62.90	67.00	64.20	50.20	68.20
	KARNET [51]	R-50		89.33	83.55	44.79	71.61	63.05	67.06	69.53	90.47	79.46	77.84	51.04	60.97	65.38	69.46	49.53	68.87
	RADet [25]	RX-101		79.45	76.99	48.05	65.83	65.46	74.40	68.86	89.70	78.14	74.97	49.92	64.63	66.14	71.58	62.16	69.09
	RoI-Trans. [8]	R-101	✓	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
	CAD-Net [74]	R-101		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
	AOOD [85]	DPN-92	✓	89.99	81.25	44.50	73.20	68.90	60.33	66.86	90.89	80.99	86.23	64.98	63.88	65.24	68.36	62.13	71.18
	Cascade-FF [18]	R-152		89.9	80.4	51.7	77.4	68.2	75.2	75.6	90.8	78.8	84.4	62.3	64.6	57.7	69.4	50.1	71.8
	SCRDet [69]	R-101	✓	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	**87.94**	86.86	65.02	66.68	66.25	68.24	65.21	72.61
	SARD [55]	R-101		89.93	84.11	54.19	72.04	68.41	61.18	66.00	90.82	87.79	86.59	65.65	64.04	66.68	68.84	68.03	72.95
	GLS-Net [23]	R-101		88.65	77.40	51.20	71.03	73.30	72.16	84.68	90.87	80.43	85.38	58.33	62.27	67.58	70.69	60.42	72.96
	FADet [24]	R-101	✓	90.21	79.58	45.49	76.41	73.18	68.27	79.56	90.83	83.40	84.68	53.40	65.42	74.17	69.69	64.86	73.28
	MFIAR-Net [63]	R-152	✓	89.62	84.03	52.41	70.30	70.13	67.64	77.81	90.85	85.40	86.22	63.21	64.14	68.31	70.21	62.11	73.49
	Gliding Vertex [62]	R-101		89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
	SAR [34]	R-152		89.67	79.78	54.17	68.29	71.70	77.90	84.63	90.91	**88.22**	87.07	60.49	66.95	75.13	75.28	64.29	75.28
	Mask OBB [52]	RX-101	✓	89.56	85.95	54.21	72.90	76.52	74.16	85.63	89.85	83.81	86.48	54.89	69.64	73.94	69.06	63.32	75.33
	FFA [12]	R-101	✓	**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 [84]	RX-101		89.96	83.62	53.42	76.03	74.01	77.16	79.45	90.83	87.15	84.51	67.72	60.33	74.61	71.84	65.55	75.75
	F^3-Net [71]	R-152	✓	88.89	78.48	54.62	74.43	72.80	77.52	87.54	90.78	87.64	85.63	63.80	64.53	**78.06**	72.36	63.19	76.02
	CenterMap [54]	R-101	✓	89.83	84.41	54.60	70.25	77.66	78.32	87.19	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06	76.03
	CSL [66]	R-152	✓	**90.25**	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93	76.17
	MRDet [43]	R-101		89.49	84.29	55.40	66.68	76.27	82.13	87.86	90.81	86.92	85.00	52.34	65.98	76.22	76.78	67.49	76.24
	RSDet-II [42]	R-152	✓	89.93	84.45	53.77	74.35	71.52	78.31	78.12	**91.14**	87.35	86.93	65.64	65.17	75.35	79.74	63.31	76.34
	OPLD [48]	R-101	✓	89.37	**85.82**	54.10	79.58	75.00	75.13	86.92	90.88	86.42	86.62	62.46	68.41	73.98	68.11	63.69	76.43
	SCRDet++ [68]	R-101	✓	90.05	84.39	55.44	73.99	77.54	71.11	86.05	90.67	87.32	87.08	69.62	68.90	73.74	71.29	65.08	76.81
	HSP [61]	R-101	✓	90.39	**86.23**	56.12	**80.59**	77.52	73.26	83.78	90.80	87.19	85.67	69.08	**72.02**	76.98	72.50	67.96	78.01
	FR-Est [11]	R-101-DCN	✓	89.78	85.21	55.40	77.70	**80.26**	**83.78**	87.59	90.81	87.66	86.93	65.60	68.74	71.64	**79.99**	66.20	78.49
Single-stage methods	IENet [30]	R-101	✓	80.20	64.54	39.82	32.07	49.71	65.01	52.58	81.45	44.66	78.51	46.54	56.73	64.40	64.24	36.75	57.14
	TOSO [10]	R-101	✓	80.17	65.59	39.82	39.95	49.71	65.01	53.58	81.45	44.66	78.51	48.85	56.73	64.40	64.24	36.75	57.92
	PIoU [5]	DLA-34		80.9	69.7	24.1	60.2	38.3	64.4	64.8	**90.9**	77.2	70.4	46.5	37.1	57.1	61.9	64.0	60.5
	Axis Learning [58]	R-101		79.53	77.15	38.59	61.15	67.53	70.49	76.30	89.66	79.07	83.53	47.27	61.01	56.28	66.06	36.05	65.98
	A^{2}S-Det [59]	R-101		89.59	77.89	46.37	56.47	75.86	74.83	86.07	90.58	81.09	83.71	50.21	60.94	65.29	69.77	50.93	70.64
	O^2-DNet [56]	H-104	✓	89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03	71.04
	P-RSDet [81]	R-101	✓	88.58	77.83	50.44	69.29	71.10	75.79	78.66	90.88	80.10	81.71	57.92	63.03	66.30	69.77	63.13	72.30
	BBAVectors [72]	R-101	✓	88.35	79.96	50.69	62.18	78.43	78.98	87.94	90.85	83.58	84.35	54.13	60.24	65.22	64.28	55.70	72.32
	ROPDet [70]	R-101-DCN	✓	90.01	82.82	54.47	69.65	69.23	70.78	75.78	90.84	86.13	84.76	66.52	63.71	67.13	68.38	46.09	72.42
	HRP-Net [17]	HRNet-W48		89.33	81.64	48.33	75.21	71.39	74.82	77.62	90.86	81.23	81.96	62.93	62.17	66.27	66.98	62.13	72.83
	DRN [41]	H-104	✓	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23
	CFC-Net [36]	R-101	✓	89.08	80.41	52.41	70.02	76.28	78.11	87.21	90.89	84.47	85.64	60.51	61.52	67.82	68.02	50.09	73.50
	R^{4}Det [49]	R-152		88.96	85.42	52.91	73.84	74.86	81.52	80.29	90.79	86.95	85.25	64.05	60.93	69.00	70.55	67.76	75.84
	R^{3}Det [67]	R-152	✓	89.80	83.77	48.11	66.77	78.76	83.27	87.84	90.82	85.38	85.51	65.67	62.68	67.53	78.56	**72.62**	76.47
	PolarDet [77]	R-101	✓	89.65	87.07	48.14	70.97	78.53	80.34	87.45	90.76	85.63	86.87	61.64	70.32	71.92	73.09	67.15	76.64
	S^{2}A-Net-DAL [37]	R-50	✓	89.69	83.11	55.03	71.00	78.30	81.90	**88.46**	90.89	84.97	**87.46**	64.41	65.65	76.86	72.09	64.35	76.95
	R^{3}Det-DCL [64]	R-152	✓	89.26	83.60	53.54	72.76	79.04	82.56	87.31	90.67	86.59	86.98	67.49	66.88	73.29	70.56	69.99	77.37
	RDD [80]	R-101	✓	89.15	83.92	52.51	73.06	77.81	79.00	87.08	90.62	86.72	87.15	63.96	**70.29**	76.98	75.79	72.15	77.75
	S^{2}A-Net [15]	R-101	✓	89.28	84.11	**56.95**	79.21	**80.18**	82.93	**89.21**	90.86	84.66	**87.61**	**71.66**	68.23	**78.58**	78.20	65.55	**79.15**
	GWD (Ours)	R-152	✓	89.66	84.99	**59.26**	**82.19**	78.97	**84.83**	87.70	90.21	86.54	86.85	**73.47**	67.77	76.92	**79.22**	**74.92**	**80.23**

Table 7: AP on different objects and mAP on DOTA. R-101 denotes ResNet-101 (likewise for R-50, R-152), RX-101 and H-104 stands for ResNeXt101 [60] and Hourglass-104 [39]. Other backbone include DPN-92 [4], DLA-34 [73], DCN [7], HRNet-W48 [53], U-Net [46]. MS indicates that multi-scale training or testing is used.

two detectors on five datasets to verify the effectiveness of GWD. When RetinaNet is used as the base detector in Tab. 2, the GWD-based detector is improved by 1.28%, 0.88%, 3.20%, 2.14% under three different aerial image datasets of HRSC206, UCAS-AOD and DOTA, respectively. Note that to increase the reliability of the results from small dataset, the experiments of the first two datasets have involved additional data augmentation, including random graying and random rotation. The rotation detector R^{3}Net [67] achieves the state-of-the-art performance on large-scale DOTA. It

METHOD	BACKBONE	MAP_{50} (07)	MAP_{50} (12)
RC1 & RC2 [33]	VGG16	75.7	–
AXIS LEARNING [58]	R-101	78.15	–
TOSO [10]	R-101	79.29	–
R^2PN [76]	VGG16	79.6	–
RRD [27]	VGG16	84.3	–
ROI-TRANS. [8]	R-101	86.20	–
RSDET [42]	R-50	86.50	–
DRN [41]	H-104	–	92.70
CENTERMAP [54]	R-50	–	92.8
SBD [32]	R-50	–	93.70
GLIDING VERTEX [62]	R-101	88.20	–
OPLD [48]	R-101	88.44	–
BBAVECTORS [72]	R-101	88.6	–
S^2A-NET [15]	R-101	**90.17**	95.01
R^3DET [67]	R-101	89.26	96.01
R^3DET-DCL [64]	R-101	89.46	**96.41**
FPN-CSL [66]	R-101	89.62	96.10
DAL [37]	R-101	89.77	–
R^3DET-GWD (OURS)	R-101	**89.85**	**97.37**

Table 8: Detection accuracy on HRSC2016.

can be seen that GWD further improves the performance by 0.90%. Tab. 3 also gives ablation test on two scene text datasets. There are a large number of objects in the boundary position in scene text, so the GWD-based RetinaNet has obtained a notable gain – increased by 6.16% and 4.51% on the MLT and ICDAR2015 datasets, respectively. Even with the use of data augmentation or a stronger detector R^3Det, GWD can still obtain a stable gain, with an improvement range from 1.31% to 1.56%.

5.3. Training Strategies and Tricks

In order to further improve the performance of the model on DOTA, we verified many commonly used training strategies and tricks, including backbone, training schedule, data augmentation (DA), multi-scale training and testing (MS), stochastic weights averaging (SWA) [19, 75], multi-scale image cropping (MSC), model ensemble (ME), as shown in Tab. 4.

Backbone: Under the conditions of different detectors (RetinaNet and R^3Det), different training schedules (experimental groups {**#11**,**#16**}, {**#24**,**#29**}), and different tricks (experimental groups {**#26**,**#31**}, {**#28**,**#33**}), large backbone can bring stable performance improvement.

Multi-scale training and testing: Multi-scale training and testing is an effective means to improve the performance of aerial images with various object scales. In this paper, training and testing scale set to [450, 500, 640, 700, 800, 900, 1,000, 1,100, 1,200]. Experimental groups {**#3**,**#4**}, {**#5**,**#6**} and {**#11**,**#12**} show the its effectiveness, increased by 0.9%, 1.09%, and 0.58%, respectively.

Training schedule: When data augmentation and multi-scale training are added, it is necessary to appropriately lengthen the training time. From the experimental groups {**#3**,**#5**} and {**#16**,**#29**}, we can find that the performance respectively increases by 0.77% and 1.22% when the training schedule is increased from 40 or 30 epochs to 60 epochs.

Stochastic weights averaging (SWA): SWA technique has been proven to be an effective tool for improving object detection. In the light of [75], we train our detector for an extra 12 epochs using cyclical learning rates and then average these 12 checkpoints as the final detection model. It can be seen from experimental groups {**#1**, **#2**}, {**#20**, **#21**} and {**#25**, **#26**} in Tab. 4 that we get 0.99%, 1.20% and 1.13% improvement on the challenging DOTA benchmark.

Multi-scale image cropping: Large-scene object detection often requires image sliding window cropping before training. During testing, sliding window cropping testing is required before the results are merged. Two adjacent sub-images often have an overlapping area to ensure that the truncated object can appear in a certain sub-image completely. The cropping size needs to be moderate, too large is not conducive to the detection of small objects, and too small will cause large objects to be truncated with high probability. Multi-scale cropping is an effective detection technique that is beneficial to objects of various scales. In this paper, our multi-scale crop size and corresponding overlap size are [600, 800, 1,024, 1,300, 1,600] and [150, 200, 300, 300, 400], respectively. According to experimental groups {**#6**, **#7**} and {**#30**, **#32**}, the large object categories (e.g. GTF and SBF) that are often truncated have been significantly improved. Take group {**#6**, **#7**} as an example, GTF and SBF increased by 6.43% and 6.14%, respectively.

5.4. Further Comparison

High precision detection: The advantage of aligning detection metric and loss is that a higher precision prediction box can be learned. Object with large aspect ratios are more sensitive to detection accuracy, so we conduct high-precision detection experiments on the ship dataset HRSC2016. It can be seen in Tab. 5 that our GWD-based detector exhibits clear advantages under high IoU thresholds. Taking AP_{75} as an example, GWD has achieved improvement by 11.89% and 22.46% on the two detectors, respectively. We also compares the peer techniques, mainly including IoU-Smooth L1 Loss [69], CSL [66], and DCL [64] on DOTA validation set. As shown on the right of Tab. 6, the GWD-based method achieves the highest performance on mAP_{75} and $mAP_{50:95}$, at 38.68% and 38.71%.

Comparison of techniques to solve the regression issues: For the three issues of inconsistency between metric and loss, boundary discontinuity and square-like problem, Tab. 6 compares the five peer techniques, including IoU-Smooth L1 Loss, CSL, and DCL on DOTA test set. For fairness, these methods are all implemented on the same

baseline method, and are trained and tested under the same environment and hyperparameters.

In particular, we detail the accuracy of the seven categories, including large aspect ratio (e.g. BR, SV, LV, SH, HA) and square-like object (e.g. ST, RD), which contain many corner cases in the dataset. These categories are assumed can better reflect the real-world challenges and advantages of our method. Many methods that solve the boundary discontinuity have achieved significant improvements in the large aspect ratio object category, and the methods that take into account the square-like problem perform well in the square-like object, such as GWD, DCL and Modulated loss.

However, there is rarely a unified method to solve all problems, and most methods are proposed for part of problems. Among them, the most comprehensive method is IoU-Smooth L1 Loss. However, the gradient direction of IoU-Smooth L1 Loss is still dominated by smooth L1 loss, so the metric and loss cannot be regarded as truly consistent. Besides, IoU-Smooth L1 Loss needs to determine whether the prediction box is within the defined range before calculating IoU at the boundary position, Otherwise, it needs to convert to the same definition as ground truth. In contrast, due to the three unique properties of GWD, it need to make additional judgments to elegantly solve all problems. From Tab. 6, GWD outperforms on most categories. For the seven listed categories (7-mAP) and overall performance (mAP), GWD-based methods are also the best. Fig. 1 visualizes the comparison between Smooth L1 loss-based and GWD-based detector.

5.5. Comprehensive Overall Comparison

Results on DOTA: Due to the complexity of the aerial image and the large number of small, cluttered and rotated objects, DOTA is a very challenging dataset. We compare the proposed approach with other state-of-the-art methods on DOTA, as shown in Tab. 7. As far as I know, this is the most comprehensive statistical comparison of methods on the DOTA dataset. Since different methods use different image resolution, network structure, training strategies and various tricks, we cannot make absolutely fair comparisons. In terms of overall performance, our method has achieved the best performance so far, at around 80.23%.

Results on HRSC2016: The HRSC2016 contains lots of large aspect ratio ship instances with arbitrary orientation, which poses a huge challenge to the positioning accuracy of the detector. Experimental results at Tab. 8 shows that our model achieves state-of-the-art performances, about 89.85% and 97.37% in term of 2007 and 2012 evaluation metric.

6. Conclusion

This paper has presented a Gaussian Wasserstain distance based loss to model the deviation between two rotating bounding boxes for object detection. The designated loss directly aligns with the detection accuracy and the model can be efficiently learned via back-propagation. More importantly, thanks to its three unique properties, GWD can also elegantly solve the boundary discontinuity and square-like problem regardless how the bounding box is defined. Experimental results on extensive public benchmarks show the state-of-the-art performance of our detector.

Appendix

6.1. Proof of $d := \mathbf{W}(\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1); \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2))$

The entire proof process refers to this blog [3].

The Wasserstein coupling distance $\mathbf{W}$ between two probability measures μ and ν on $\mathbb{R}^n$ expressed as follows:

$$\mathbf{W}(\mu; \nu) := \inf \mathbb{E}(\|\mathbf{X} - \mathbf{Y}\|_2^2)^{1/2} \tag{9}$$

where the infimum runs over all random vectors $(\mathbf{X}, \mathbf{Y})$ of $\mathbb{R}^n \times \mathbb{R}^n$ with $\mathbf{X} \sim \mu$ and $\mathbf{Y} \sim \nu$. It turns out that we have the following formula for $d := \mathbf{W}(\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1); \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2))$:

$$d^2 = \|\mathbf{m}_1 - \mathbf{m}_2\|_2^2 + \mathbf{Tr}\left(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 - 2(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}\right) \tag{10}$$

This formula interested several works [14, 40, 22, 9]. Note in particular we have:

$$\mathbf{Tr}\left((\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}\right) = \mathbf{Tr}\left((\mathbf{\Sigma}_2^{1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{1/2})^{1/2}\right) \tag{11}$$

In the commutative case $\mathbf{\Sigma}_1\mathbf{\Sigma}_2 = \mathbf{\Sigma}_2\mathbf{\Sigma}_1$, Eq. 10 becomes:

$$\begin{aligned} d^2 =& \|\mathbf{m}_1 - \mathbf{m}_2\|_2^2 + \|\mathbf{\Sigma}_1^{1/2} - \mathbf{\Sigma}_2^{1/2}\|_F^2 \\ =& (x_1 - x_2)^2 + (y_1 - y_2)^2 + \frac{(w_1 - w_2)^2 + (h_1 - h_2)^2}{4} \\ =& l_2\text{-norm}\left(\left[x_1, y_1, \frac{w_1}{2}, \frac{h_1}{2}\right]^\top, \left[x_2, y_2, \frac{w_2}{2}, \frac{h_2}{2}\right]^\top\right) \end{aligned} \tag{12}$$

where $\|\|_F$ is the Frobenius norm. Note that both boxes are horizontal at this time, and Eq. 12 is approximately equivalent to the l_2-norm loss (note the additional denominator of 2 for w and h), which is consistent with the loss commonly used in horizontal detection. This also partly proves the correctness of using Wasserstein distance as the regression loss.

To prove Eq. 10, one can first reduce to the centered case $\mathbf{m}_1 = \mathbf{m}_2 = \mathbf{0}$. Next, if $(\mathbf{X}, \mathbf{Y})$ is a random vector (Gaussian or not) of $\mathbb{R}^n \times \mathbb{R}^n$ with covariance matrix

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{\Sigma}_2 \end{pmatrix} \tag{13}$$

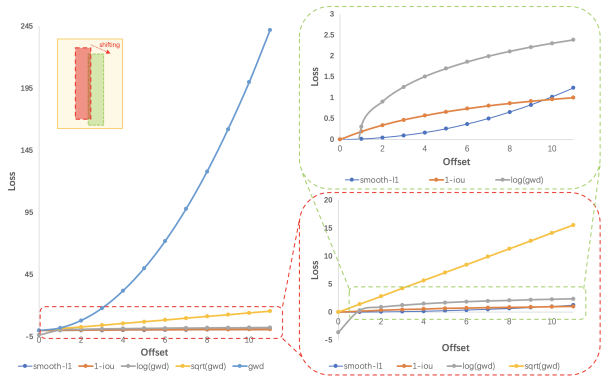


Figure 6: Different forms of GWD-based regression loss curve.

then the quantity

$$\mathbb{E}(\|\mathbf{X},\mathbf{Y}\|_2^2) = \mathbf{Tr}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 - 2\mathbf{C}) \tag{14}$$

depends only on $\mathbf{\Gamma}$. Also, when $\mu = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$ and $\nu = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2)$, one can restrict the infimum which defines W to run over Gaussian laws $\mathcal{N}(\mathbf{0}, \mathbf{\Gamma})$ on $\mathbb{R}^n \times \mathbb{R}^n$ with covariance matrix $\mathbf{\Gamma}$ structured as above. The sole constrain on $\mathbf{C}$ is the Schur complement constraint:

$$\mathbf{\Sigma}_1 - \mathbf{C}\mathbf{\Sigma}_2^{-1}\mathbf{C}^\top \succeq 0 \tag{15}$$

The minimization of the function

$$\mathbf{C} \rightarrowtail -2\mathbf{Tr}(\mathbf{C}) \tag{16}$$

under the constraint above leads to Eq. 10. A detailed proof is given by [14]. Alternatively, one may find an optimal transportation map as [22]. It turns out that $\mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$ is the image law of $\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ with the linear map

$$\mathbf{x} \rightarrowtail \mathbf{m}_2 + \mathbf{A}(\mathbf{x}\mathbf{m}_1) \tag{17}$$

where

$$\mathbf{A} = \mathbf{\Sigma}_1^{-1/2}(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}\mathbf{\Sigma}_1^{-1/2} = \mathbf{A}^\top \tag{18}$$

To check that this maps $\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ to $\mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$, say in the case $\mathbf{m}_1 = \mathbf{m}_2 = \mathbf{0}$ for simplicity, one may define the random column vectors $\mathbf{X} \sim \mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ and $\mathbf{Y} = \mathbf{A}\mathbf{X}$ and write

$$\begin{aligned}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top) =& \mathbf{A}\mathbb{E}(\mathbf{X}\mathbf{X}^\top)\mathbf{A}^\top \\ =& \mathbf{\Sigma}_1^{1/2}(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}\mathbf{\Sigma}_1^{1/2} \\ =& \mathbf{\Sigma}_2\end{aligned} \tag{19}$$

To check that the map is optimal, one may use,

$$\begin{aligned}\mathbb{E}(\|\mathbf{X}-\mathbf{Y}\|_2^2) =& \mathbb{E}(\|\mathbf{X}\|_2^2) + \mathbb{E}(\|\mathbf{Y}\|_2^2) - 2\mathbb{E}(<\mathbf{X},\mathbf{Y}>) \\ =& \mathbf{Tr}(\mathbf{\Sigma}_1) + \mathbf{Tr}(\mathbf{\Sigma}_2) - 2\mathbb{E}(<\mathbf{X},\mathbf{A}\mathbf{X}>) \\ =& \mathbf{Tr}(\mathbf{\Sigma}_1) + \mathbf{Tr}(\mathbf{\Sigma}_2) - 2\mathbf{Tr}(\mathbf{\Sigma}_1\mathbf{A})\end{aligned} \tag{20}$$

$1-\frac{1}{(\tau+f(d^2))}$	$\tau=1$	$\tau=2$	$\tau=3$	$\tau=5$	$f(d^2)$	d^2
$f(\cdot)=sqrt$	68.56	**68.93**	68.37	67.77	54.27	49.11
$f(\cdot)=\log$	67.87	68.09	67.48	66.49	**69.82**	

Table 9: Ablation test of GWD-based regression loss form and hyperparameter on DOTA. The based detector is RetinaNet.

and observe that by the cyclic property of the trace,

$$\mathbf{Tr}(\mathbf{\Sigma}_1\mathbf{A}) = \mathbf{Tr}((\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2})^{1/2}) \tag{21}$$

The generalizations to elliptic families of distributions and to infinite dimensional Hilbert spaces is probably easy. Some more "geometric" properties of Gaussians with respect to such distances where studied more recently by [50] and [50].

6.2. Improved GWD-based Regression Loss

In Tab. 9, we compare three different forms of GWD-based regression loss, including d^2, $1-\frac{1}{(\tau+f(d^2))}$ and $f(d^2)$. The performance of directly using GWD (d^2) as the regression loss is extremely poor, only 49.11%, due to its rapid growth trend (as shown on the left of Fig. 6). In other words, the regression loss d^2 is too sensitive to large errors. In contrast, $1-\frac{1}{(\tau+f(d^2))}$ achieves a significant improvement by fitting IoU loss. This loss form introduces two new hyperparameters, the non-linear function $f(\cdot)$ to transform the Wasserstein distance, and the constant τ to modulate the entire loss. From Tab. 9, the overall performance of using $sqrt$ outperforms that using $\log$, about 0.98±0.3% higher. For $f(\cdot) = sqrt$ with $\tau = 2$, the model achieves the best performance, about 68.93%. In order to further reduce the number of hyperparameters of the loss function, we directly use the GWD after nonlinear transformation ($f(d^2)$) as the regression loss. As shown in the red box in Fig. 6, $f(d^2)$ still has a nearly linear trend after transformation using the nonlinear function $sqrt$ and only achieves 54.27%. In comparison, the $\log$ function can better make the $f(d^2)$ change value close to IoU loss (see green box in Fig. 6) and achieve the highest performance, about 69.82%. In general, we do not need to strictly fit the IoU loss, and the regression loss should not be sensitive to large errors.

Acknowledgment

The author Xue Yang is supported by Wu Wen Jun Honorary Doctoral Scholarship, AI Institute, Shanghai Jiao Tong University.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghe-

mawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] Seyed Majid Azimi, Eleonora Vig, Reza Bahmanyar, Marco Körner, and Peter Reinartz. Towards multi-class object detection in unconstrained remote sensing imagery. In *Asian Conference on Computer Vision*, pages 150–165. Springer, 2018.

[3] Djalil Chafaï. Wasserstein distance between two gaussians. Website, 2010. https://djalil.chafai.net/blog/2010/04/30/wasserstein-distance-between-two-gaussians/.

[4] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Advances in neural information processing systems*, pages 4467–4475, 2017.

[5] Zhiming Chen, Kean Chen, Weiyao Lin, John See, Hui Yu, Yan Ke, and Cong Yang. Piou loss: Towards accurate oriented object detection in complex environments. *Proceedings of the European Conference on Computer Vision*, 2020.

[6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017.

[8] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 2849–2858, 2019.

[9] DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[10] Pengming Feng, Youtian Lin, Jian Guan, Guangjun He, Huifeng Shi, and Jonathon Chambers. Toso: Student'st distribution aided one-stage orientation target detection in remote sensing images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4057–4061. IEEE, 2020.

[11] Kun Fu, Zhonghan Chang, Yue Zhang, and Xian Sun. Point-based estimator for arbitrary-oriented object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[12] Kun Fu, Zhonghan Chang, Yue Zhang, Guangluan Xu, Keshu Zhang, and Xian Sun. 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, 2020.

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[14] Clark R Givens, Rae Michael Shortt, et al. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

[15] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *arXiv preprint arXiv:2008.09397*, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[17] Xu He, Shiping Ma, Linyuan He, and Le Ru. High-resolution polar network for object detection in remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 2020.

[18] Liping Hou, Ke Lu, Jian Xue, and Li Hao. Cascade detector with feature fusion for arbitrary-oriented objects in remote sensing images. In *2020 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2020.

[19] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[20] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R2cnn: rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017.

[21] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition*, pages 1156–1160. IEEE, 2015.

[22] Martin Knott and Cyril S Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

[23] Chengyuan Li, Bin Luo, Hailong Hong, Xin Su, Yajun Wang, Jun Liu, Chenjie Wang, Jing Zhang, and Linhai Wei. Object detection based on global-local saliency constraint in aerial images. *Remote Sensing*, 12(9):1435, 2020.

[24] Chengzheng Li, Chunyan Xu, Zhen Cui, Dan Wang, Tong Zhang, and Jian Yang. Feature-attentioned object detection in remote sensing imagery. In *2019 IEEE International Conference on Image Processing*, pages 3886–3890. IEEE, 2019.

[25] Yangyang Li, Qin Huang, Xuan Pei, Licheng Jiao, and Ronghua Shang. Radet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sensing*, 12(3):389, 2020.

[26] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690, 2018.

[27] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5909–5918, 2018.

[28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[30] Youtian Lin, Pengming Feng, and Jian Guan. Ienet: Interacting embranchment one stage anchor free detector for orientation aerial object detection. *arXiv preprint arXiv:1912.00969*, 2019.

[31] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5676–5685, 2018.

[32] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. Omnidirectional scene text detection with sequential-free box discretization. *arXiv preprint arXiv:1906.02371*, 2019.

[33] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods*, volume 2, pages 324–331, 2017.

[34] Junyan Lu, Tie Li, Jingyu Ma, Zhuqiang Li, and Hongguang Jia. Sar: Single-stage anchor-free rotating object detection. *IEEE Access*, 8:205902–205912, 2020.

[35] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122, 2018.

[36] Qi Ming, Lingjuan Miao, Zhiqiang Zhou, and Yunpeng Dong. Cfc-net: A critical feature capturing network for arbitrary-oriented object detection in remote sensing images. *arXiv preprint arXiv:2101.06849*, 2021.

[37] Qi Ming, Zhiqiang Zhou, Lingjuan Miao, Hongwei Zhang, and Linhao Li. Dynamic anchor learning for arbitrary-oriented object detection. *arXiv preprint arXiv:2012.04150*, 2020.

[38] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition*, volume 1, pages 1454–1459. IEEE, 2017.

[39] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[40] Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

[41] Xingjia Pan, Yuqiang Ren, Kekai Sheng, Weiming Dong, Haolei Yuan, Xiaowei Guo, Chongyang Ma, and Changsheng Xu. Dynamic refinement network for oriented and densely packed object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11207–11216, 2020.

[42] Wen Qian, Xue Yang, Silong Peng, Junchi Yan, and Yue Guo. Learning modulated loss for rotated object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[43] Ran Qin, Qingjie Liu, Guangshuai Gao, Di Huang, and Yunhong Wang. Mrdet: A multi-head network for accurate oriented object detection in aerial images. *arXiv preprint arXiv:2012.13135*, 2020.

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[45] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[47] Xuepeng Shi, Shiguang Shan, Meina Kan, Shuzhe Wu, and Xilin Chen. Real-time rotation-invariant face detection with progressive calibration networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2295–2303, 2018.

[48] Qing Song, Fan Yang, Lu Yang, Chun Liu, Mengjie Hu, and Lurui Xia. Learning point-guided localization for detection in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020.

[49] Peng Sun, Yongbin Zheng, Zongtan Zhou, Wanying Xu, and Qiang Ren. R4det: Refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images. *Image and Vision Computing*, 103:104036, 2020.

[50] Asuka Takatsu and Takumi Yokota. Cone structure of l2-wasserstein spaces. *Journal of Topology and Analysis*, 4(02):237–253, 2012.

[51] Tianhang Tang, Yiguang Liu, Yunan Zheng, Xianzhen Zhu, and Yangyu Zhao. Rotating objects detection in aerial images via attention denoising and angle loss refining. *DEStech Transactions on Computer Science and Engineering*, (cisnr), 2020.

[52] Jinwang Wang, Jian Ding, Haowen Guo, Wensheng Cheng, Ting Pan, and Wen Yang. Mask obb: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sensing*, 11(24):2930, 2019.

[53] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[54] Jinwang Wang, Wen Yang, Heng-Chao Li, Haijian Zhang, and Gui-Song Xia. Learning center probability map for detecting objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[55] Yashan Wang, Yue Zhang, Yi Zhang, Liangjin Zhao, Xian Sun, and Zhi Guo. Sard: Towards scale-aware rotated object detection in aerial imagery. *IEEE Access*, 7:173855–173865, 2019.

[56] Haoran Wei, Yue Zhang, Zhonghan Chang, Hao Li, Hongqi Wang, and Xian Sun. Oriented objects as pairs of middle lines. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:268–279, 2020.

[57] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018.

[58] Zhifeng Xiao, Linjun Qian, Weiping Shao, Xiaowei Tan, and Kai Wang. Axis learning for orientated objects detection in aerial images. *Remote Sensing*, 12(6):908, 2020.

[59] Zhifeng Xiao, Kai Wang, Qiao Wan, Xiaowei Tan, Chuan Xu, and Fanfan Xia. A2s-det: Efficiency anchor matching in aerial image oriented object detection. *Remote Sensing*, 13(1):73, 2021.

[60] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[61] Chunyan Xu, Chengzheng Li, Zhen Cui, Tong Zhang, and Jian Yang. Hierarchical semantic propagation for object detection in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 58(6):4353–4364, 2020.

[62] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[63] Feng Yang, Wentong Li, Haiwei Hu, Wanyi Li, and Peng Wang. Multi-scale feature integrated attention-based rotation network for object detection in vhr aerial images. *Sensors*, 20(6):1686, 2020.

[64] Xue Yang, Liping Hou, Yue Zhou, Wentao Wang, and Junchi Yan. Dense label encoding for boundary discontinuity free rotation detection. *arXiv preprint arXiv:2011.09670*, 2020.

[65] Xue Yang, Hao Sun, Kun Fu, Jirui Yang, Xian Sun, Menglong Yan, and Zhi Guo. 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, 2018.

[66] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *Proceedings of the European Conference on Computer Vision*, pages 677–694. Springer, 2020.

[67] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[68] Xue Yang, Junchi Yan, Xiaokang Yang, Jin Tang, Wenglong Liao, and Tao He. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *arXiv preprint arXiv:2004.13316*, 2020.

[69] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8232–8241, 2019.

[70] Zhixiang Yang, Kunkun He, Fuhao Zou, Wanhua Cao, Xiaoyun Jia, Kai Li, and Chuntao Jiang. Ropdet: real-time anchor-free detector based on point set representation for rotating object. *Journal of Real-Time Image Processing*, 17(6):2127–2138, 2020.

[71] Xinhai Ye, Fengchao Xiong, Jianfeng Lu, Jun Zhou, and Yuntao Qian. F3-net: Feature fusion and filtration network for object detection in optical remote sensing images. *Remote Sensing*, 12(24):4027, 2020.

[72] Jingru Yi, Pengxiang Wu, Bo Liu, Qiaoying Huang, Hui Qu, and Dimitris Metaxas. Oriented object detection in aerial images with box boundary-aware vectors. *arXiv preprint arXiv:2008.07043*, 2020.

[73] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2403–2412, 2018.

[74] Gongjie Zhang, Shijian Lu, and Wei Zhang. Cad-net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):10015–10024, 2019.

[75] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. Swa object detection. *arXiv preprint arXiv:2012.12645*, 2020.

[76] Zenghui Zhang, Weiwei Guo, Shengnan Zhu, and Wenxian Yu. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geoscience and Remote Sensing Letters*, 15(11):1745–1749, 2018.

[77] Pengbo Zhao, Zhenshen Qu, Yingjia Bu, Wenming Tan, Ye Ren, and Shiliang Pu. Polardet: A fast, more precise detector for rotated target in aerial images. *arXiv preprint arXiv:2010.08720*, 2020.

[78] Yu Zheng, Danyang Zhang, Sinan Xie, Jiwen Lu, and Jie Zhou. Rotation-robust intersection over union for 3d object detection. In *European Conference on Computer Vision*, pages 464–480. Springer, 2020.

[79] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12993–13000, 2020.

[80] Bo Zhong and Kai Ao. Single-stage rotation-decoupled detector for oriented object. *Remote Sensing*, 12(19):3262, 2020.

[81] Lin Zhou, Haoran Wei, Hao Li, Wenzhe Zhao, Yi Zhang, and Yue Zhang. Arbitrary-oriented object detection in remote sensing images based on polar coordinates. *IEEE Access*, 8:223373–223384, 2020.

[82] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and

accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017.

[83] Haigang Zhu, Xiaogang Chen, Weiqun Dai, Kun Fu, Qixiang Ye, and Jianbin Jiao. Orientation robust object detection in aerial images using deep convolutional neural network. In *2015 IEEE International Conference on Image Processing*, pages 3735–3739. IEEE, 2015.

[84] Yixing Zhu, Jun Du, and Xueqing Wu. Adaptive period embedding for representing oriented objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[85] Fuhao Zou, Wei Xiao, Wanting Ji, Kunkun He, Zhixiang Yang, Jingkuan Song, Helen Zhou, and Kai Li. Arbitrary-oriented object detection via dense feature fusion and attention model for remote sensing super-resolution image. *Neural Computing and Applications*, pages 1–14, 2020.