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]

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, and 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 (Yang et al., 2018a; Jiao et al., 2018; Yang et al., 2018b; 2019), scene text (Zhou et al., 2017; Liu et al., 2018; Jiang et al., 2017; Ma et al., 2018; Liao et al., 2018b), faces (Shi et al., 2018) and 3D objects (Zheng et al., 2020a), retail scenes (Chen et al., 2020; Pan et al., 2020), etc. Compared with the large literature on horizontal object detection (Girshick, 2015; Ren et al., 2015; Dai et al., 2016; Lin et al., 2017a;b), research in oriented object detection is relatively in its earlier stage, with many open problems to solve.

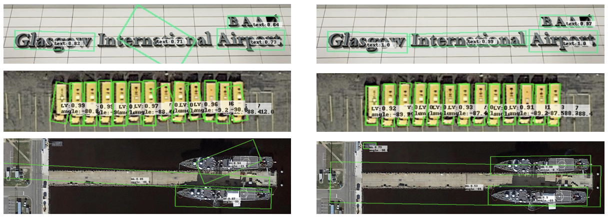


Figure 1. Comparison of the detection results between Smooth L1 loss-based (left) and the proposed GWD-based (right) detector.

The dominant line of works (Azimi et al., 2018; Ding et al., 2019; Yang et al., 2019; 2021b) 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) (Chafaï, 2010).

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

[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. Correspondence to: Junchi Yan <yangjunchi@sjtu.edu.cn>, Xue Yang <yangxue-2019-sjtu@sjtu.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

with accuracy metric and thus naturally improves the model.

iii) Our GWD-based loss can elegantly resolve boundary discontinuity and square-like problem, regardless how the rotating bounding box is defined. In contrast, the design of most peer works (Yang & Yan, 2020; Yang et al., 2021a) 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 (Girshick, 2015; Ren et al., 2015; Lin et al., 2017a;b) 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 (Yang & Yan, 2020) 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 (Azimi et al., 2018), ROI-Transformer (Ding et al., 2019), SCRDet (Yang et al., 2019) and Gliding Vertex (Xu et al., 2020) are two-stage representative methods whose pipeline comprises of object localization and classification, while DRN (Pan et al., 2020), R^{3}Det (Yang et al., 2021b) and RSDet (Qian et al., 2021) are single-stage methods. For scene text detection, RRPN (Ma et al., 2018) employ rotated RPN to generate rotated proposals and further perform rotated bounding box regression. TextBoxes++ (Liao et al., 2018a) adopts vertex regression on SSD. RRD (Liao et al., 2018b) 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 (Yang et al., 2019) and RSDet (Qian et al., 2021) propose IoU-smooth L1 loss and modulated loss to smooth the the boundary loss jump. CSL (Yang & Yan, 2020) transforms angular prediction from a regression problem to a classification one. DCL (Yang et al., 2021a) 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. Instance segmentation-based methods are practical, and relevant methods (e.g. Mask OBB (Wang et al., 2019)) have been published. However, this approach has its limitations. First, using rotated boxes as binary masks will introduce background area, which will reduce the classification accuracy of pixels and affect the accuracy of the final prediction box. Secondly, for the top-down methods (e.g. Mask RCNN (He et al., 2017)), dense scenes will limit the detection of horizontal boxes because of the excessive suppression of dense horizontal overlapping bounding boxes due to NMS, thereby affecting subsequent segmentation. Aerial images often show large scenes with a large number of dense and small objects, which is not suitable for the bottom-up methods, such as SOLO (Wang et al., 2020b) and CondInst (Tian et al., 2020), which assign different instances to different channels. This is the main reason why regression-based rotation detection algorithms still dominate in the field of aerial imagery.

Approximate differentiable rotating IoU loss. It has been shown in classic horizontal detectors that the use of IoU induced loss e.g. GIoU (Rezatofighi et al., 2019), DIoU (Zheng et al., 2020b) 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. The approximate IoU loss proposed by PolarMask (Xie et al., 2020) is also an effective design idea. However, its calculation is discrete, which means that there is a theoretical calculation error and the number of discrete point samples greatly affects the final calculation accuracy. PIoU (Chen et al., 2020) is realized by simply counting the number of pixels. To tackle the uncertainty of convex caused by rotation, (Zheng et al., 2020a) proposes a projection operation to estimate the intersection area. SCRDet (Yang et al., 2019) combines IoU and smooth L1 loss to develop an IoU-smooth L1 loss, which partly circumvents the need for differentiable rotating IoU loss.

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 Specific Detector Design

Fig. 2 gives two popular definitions for parameterizing rotating bounding box based angles: OpenCV protocol

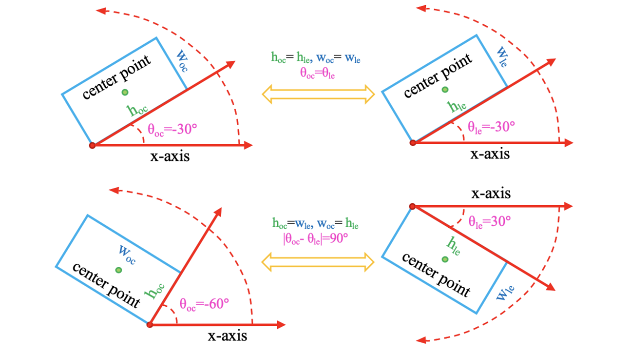


Figure 2. Two definitions of bounding boxes. **Left:** OpenCV Definition D_{oc}, **Right:** Long Edge Definition D_{le}.

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:

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

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

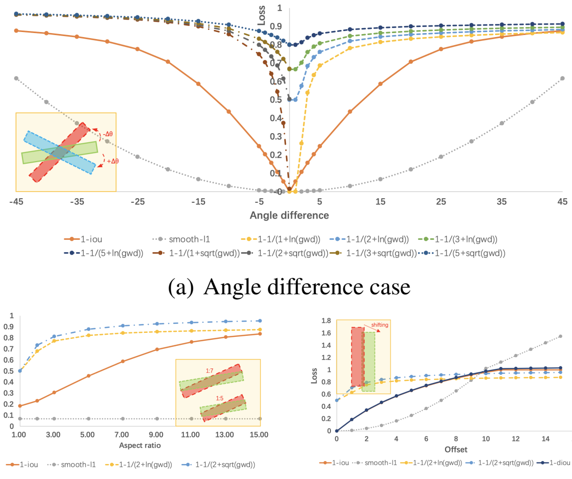
The main difference refers to the edge and angle (h, w, θ): when the same bounding box takes different representations by the two definitions, the order of the edges is exchanged and the angle difference is 90°.

In many works, the pipeline design are tightly coupled with the choice of the bounding box definition to avoid specific problems: SCRDet (Yang et al., 2019), R^{3}Det (Yang et al., 2021b) are based on D_{oc} to avoid the square-like problem, while CSL (Yang & Yan, 2020), DCL (Yang et al., 2021a) 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 (Rezatofighi et al., 2019; Zheng et al., 2020b). 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 (Girshick, 2015):

Case 1: Fig. 3(a) 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.



(a) Angle difference case

(b) Aspect ratio case

(c) Center shifting case

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

Case 2: Fig. 3(b) 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. 3(c) explores the impact of center point shifting on different loss functions. Similarly, despite the same monotonicity, there is no high degree of consistency.

Seeing the above flaws of classic smooth L1 loss, IoU-induced loss has become recently popular for horizontal detection e.g. GIoU (Rezatofighi et al., 2019), DIoU (Zheng et al., 2020b). 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 (Yang et al., 2019; Yang & Yan, 2020) 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 discontinuity. Take **Case 2** as an example, we assume that there is a red anchor/proposal $(0, 0, 70, 10, -90^\circ)$ and a green ground truth $(0, 0, 10, 70, -25^\circ)$ at the boundary position[1], both of which are defined in OpenCV definition

[1] The angle of the bounding box is close to the maximum and minimum values of the angle range. For more clearly visualization,

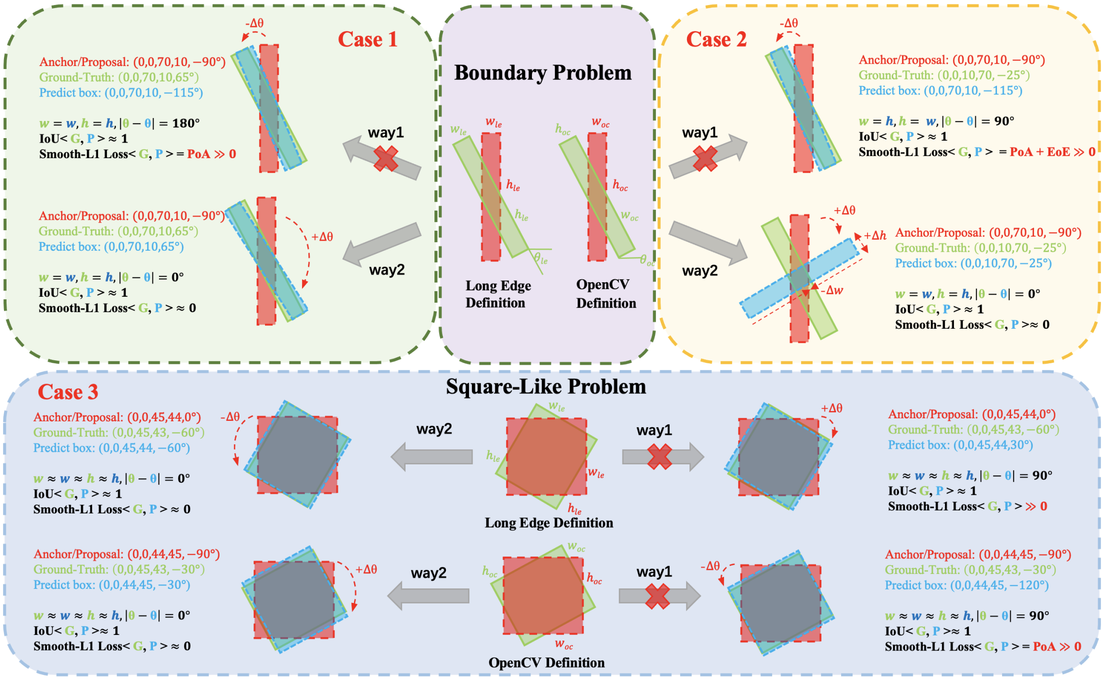

Case 1
-Δθ
Anchor/Proposal: (0,0,70,10, −90°)
Ground-Truth: (0,0,70,10,65°)
Predict box: (0,0,70,10, −115°)
$w = w, h = h, |\theta - \theta| = 180°$
IoU< G, P > ≈ 1
Smooth-L1 Loss< G, P > = PoA ≫ 0
way1
Anchor/Proposal: (0,0,70,10, −90°)
Ground-Truth: (0,0,70,10,65°)
Predict box: (0,0,70,10,65°)
+Δθ
$w = w, h = h, |\theta - \theta| = 0°$
IoU< G, P > ≈ 1
Smooth-L1 Loss< G, P > ≈ 0
way2
Boundary Problem
w_{le}
w_{le}
h_{le}
h_{le}
θ_{le}
h_{oc}
w_{oc}
h_{oc}
w_{oc}
θ_{oc}
Long Edge Definition
OpenCV Definition
Case 2
-Δθ
Anchor/Proposal: (0,0,70,10, −90°)
Ground-Truth: (0,0,10,70, −25°)
Predict box: (0,0,70,10, −115°)
way1
$w = h, h = w, |\theta - \theta| = 90°$
IoU< G, P > ≈ 1
Smooth-L1 Loss< G, P > = PoA + EoE ≫ 0
+Δθ
+Δh
-Δw
way2
Anchor/Proposal: (0,0,70,10, −90°)
Ground-Truth: (0,0,10,70, −25°)
Predict box: (0,0,10,70, −25°)
$w = w, h = h, |\theta - \theta| = 0°$
IoU< G, P > ≈ 1
Smooth-L1 Loss< G, P >≈ 0
Case 3
Square-Like Problem
Anchor/Proposal: (0,0,45,44,0°)
Ground-Truth: (0,0,45,43, −60°)
Predict box: (0,0,45,44, −60°)
-Δθ
$w \approx w \approx h \approx h, |\theta - \theta| = 0°$
IoU< G, P > ≈ 1
Smooth-L1 Loss< G, P > ≈ 0
way2
w_{le}
h_{le}
w_{le}
h_{le}
Long Edge Definition
way1
+Δθ
Anchor/Proposal: (0,0,45,44,0°)
Ground-Truth: (0,0,45,43, −60°)
Predict box: (0,0,45,44,30°)
$w \approx w \approx h \approx h, |\theta - \theta| = 90°$
IoU< G, P > ≈ 1
Smooth-L1 Loss< G, P > ≫ 0
Anchor/Proposal: (0,0,44,45, −90°)
Ground-Truth: (0,0,45,43, −30°)
Predict box: (0,0,44,45, −30°)
+Δθ
$w \approx w \approx h \approx h, |\theta - \theta| = 0°$
IoU< G, P >≈ 1
Smooth-L1 Loss< G, P > ≈ 0
way2
w_{oc}
h_{oc}
h_{oc}
w_{oc}
OpenCV Definition
way1
-Δθ
Anchor/Proposal: (0,0,44,45, −90°)
Ground-Truth: (0,0,45,43, −30°)
Predict box: (0,0,44,45, −120°)
$w \approx w \approx h \approx h, |\theta - \theta| = 90°$
IoU< G, P >≈ 1
Smooth-L1 Loss< G, P > = PoA ≫ 0

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

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 (Yang & Yan, 2020), this is because the result of the prediction box $(0, 0, 70, 10, -115°)$ 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 (Yang et al., 2021a). 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 (Yang et al., 2019) 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 (Qian et al., 2021) devises modulated loss

the ground truth has been rendered with a larger angle in Fig. 4.

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

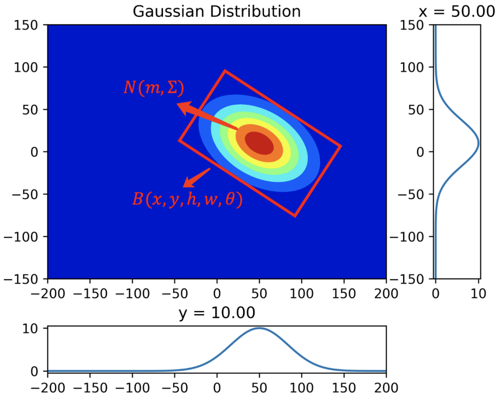


Figure 5. A schematic diagram of modeling a rotating bounding box by a two-dimensional Gaussian distribution.

to smooth the loss mutation at the boundary, but it needs to calculate the loss of as many parameter combinations as possible. CSL (Yang & Yan, 2020) 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 (Yang et al., 2021a) further solves the problem of square-like object detection introduced by D_{le}.

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 Bounding 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:

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

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

The Wasserstein distance $\mathbf{W}$ between two probability measures μ and ν on $\mathbb{R}^n$ expressed as (Chafaï, 2010):

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

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 (Givens et al., 1984; Olkin & Pukelsheim, 1982; Knott & Smith, 1984; Dowson & Landau, 1982). 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 (horizontal detection task) $\mathbf{\Sigma}_1\mathbf{\Sigma}_2 = \mathbf{\Sigma}_2\mathbf{\Sigma}_1$, Eq. 3 becomes:

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

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 (Chafaï, 2010) 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 (Rezatofighi et al., 2019; Zheng et al., 2020b), 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. 3(a) 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. 3(c) 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 boundary discontinuity and square-like problem.

GWD can also help resolve the boundary discontinuity 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.

4.3. Overall Loss Function Design

In line with (Yang & Yan, 2020; Yang et al., 2021a;b), we use the one-stage detector RetinaNet (Lin et al., 2017b) 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 $\{2,1\}$ by default. The classification loss L_{cls} is set as the focal loss (Lin et al., 2017b).

5. Experiments

We use Tensorflow (Abadi et al., 2016) for implementation on a server with Tesla V100 and 32G memory.

5.1. Datasets and Implementation Details

DOTA (Xia et al., 2018) 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. Then, 188,282 instances are annotated by experts using 15 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

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

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

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

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^{3}Det	D_{oc}	Smooth L1			70.66
	D_{oc}	GWD			**71.56** **(+0.90)**

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

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^{3}Det	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 4. Ablation study for training strategies and tricks on DOTA.

Backbone	Schedule	MS	MSC	SWA	ME	RetinaNet-GWD	R^{3}Det-GWD
R-101	30	✓				–	75.66
R-152	30	✓				–	76.18
R-152	40	✓				74.22	–
R-152	60					74.09	77.57
R-152	60	✓				75.18	78.44
R-152	60	✓	✓			75.35	78.32
R-152	60	✓		✓		75.94	78.92
R-152	60	✓	✓	✓		76.30	79.08
R-152	60	✓	✓	✓	✓	77.43	80.19

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

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)**

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 (Zhu et al., 2015) contains 1,510 aerial images of about $659 \times 1,280$ pixels, with 2 categories of 14,596 instances. In line with (Azimi et al., 2018; Xia et al., 2018), we sample 1,110 images for training and 400 for testing.

HRSC2016 (Liu et al., 2017) contains images from two scenarios including ships on sea and ships close inshore. All images are collected from six famous harbors. The training, validation and test set include 436, 181 and 444 images, respectively.

ICDAR2015 (Karatzas et al., 2015) is commonly used for oriented scene text detection and spotting. This dataset includes 1,000 training images and 500 testing images.

ICDAR 2017 MLT (Nayef et al., 2017) 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 (He et al., 2016) 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 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. 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: 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 discontinuity 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 two detectors on five datasets to verify the effectiveness of GWD.

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.

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
	Modulated Loss	D_{oc}	✓	×	×	×	42.92	67.92	52.91	72.67	53.64	**80.22**	58.21	61.21	66.05	63.50	33.32	34.61
	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	67.18	**38.41**	**38.46**
	DCL (BCL)	D_{le}	✓	×	×	×	**46.84**	74.87	**74.96**	85.70	**57.72**	**84.06**	**63.77**	**69.70**	**71.21**	**67.45**	35.44	37.54
	GWD	D_{oc}	×	×	×	×	**46.73**	**75.84**	**78.00**	**86.71**	**62.69**	**83.09**	**61.12**	**70.60**	**71.56**	**69.28**	**43.35**	**41.56**

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 represent ResNeXt101 (Xie et al., 2017) and Hourglass-104 (Newell et al., 2016). MS indicates that multi-scale training or testing is used.

	Method	Backbone	MS	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP$_{50}$
Two-stage methods	ICN (Azimi et al., 2018)	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
	RoI-Trans. (Ding et al., 2019)	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 (Zhang et al., 2019)	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
	SCRDet (Yang et al., 2019)	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
	FADet (Li et al., 2019)	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
	Gliding Vertex (Xu et al., 2020)	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
	Mask OBB (Wang et al., 2019)	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 (Fu et al., 2020)	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 (Zhu et al., 2020)	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
	CenterMap (Wang et al., 2020a)	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 (Yang & Yan, 2020)	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
	RSDet-II (Qian et al., 2021)	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
	SCRDet++ (Yang et al., 2020)	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
Single-stage methods	PIoU (Chen et al., 2020)	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
	O^2-DNet (Wei et al., 2020)	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 (Zhou et al., 2020)	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 (Yi et al., 2020)	R-101	✓	88.35	79.96	50.69	62.18	78.43	78.98	87.94	90.85	83.58	84.35	54.13	60.24	65.22	64.28	55.70	72.32
	DRN (Pan et al., 2020)	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
	R^{3}Det (Yang et al., 2021b)	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 (Zhao et al., 2020)	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 (Ming et al., 2020)	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 (Yang et al., 2021a)	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 (Zhong & Ao, 2020)	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 (Han et al., 2021)	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 8. Detection accuracy on HRSC2016.

Method	Backbone	mAP$_{50}$ (07)	mAP$_{50}$ (12)
RoI-Trans. (Ding et al., 2019)	R-101	86.20	–
RSDet (Qian et al., 2021)	R-50	86.50	–
DRN (Pan et al., 2020)	H-104	–	92.70
CenterMap (Wang et al., 2020a)	R-50	–	92.8
SBD (Liu et al., 2019)	R-50	–	93.70
Gliding Vertex (Xu et al., 2020)	R-101	88.20	–
OPLD (Song et al., 2020)	R-101	88.44	–
BBAVectors (Yi et al., 2020)	R-101	88.6	–
S^{2}A-Net (Han et al., 2021)	R-101	**90.17**	95.01
R^{3}Det (Yang et al., 2021b)	R-101	89.26	96.01
R^{3}Det-DCL (Yang et al., 2021a)	R-101	89.46	**96.41**
FPN-CSL (Yang & Yan, 2020)	R-101	89.62	96.10
DAL (Ming et al., 2020)	R-101	89.77	–
R^{3}Det-GWD (Ours)	R-101	**89.85**	**97.37**

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}Det (Yang et al., 2021b) achieves the state-of-the-art performance on large-scale DOTA. It 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^{3}Det, GWD can still obtain a stable gain, with an improvement range from 1.31% to 1.56%.

Ablation experiment of 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, multi-scale training and testing (MS), stochastic weights averaging (SWA) (Izmailov et al., 2018; Zhang et al., 2020), multi-scale image cropping (MSC), model ensemble (ME), etc. Tab. 4 demonstrates the improvement effects of various techniques on the RetinaNet-GWD and R^{3}Det-GWD, and finally achieved top performances of 77.43% and 80.19%.

5.3. 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 (Yang et al., 2019), Modulated loss (Qian et al., 2021), CSL (Yang & Yan, 2020), and DCL (Yang et al., 2021a) on DOTA validation set. As shown on the right of Tab. 6, the GWD-based method achieves the highest performance on mAP_{75} and $\text{mAP}_{50:95}$, 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, Modulated loss, CSL, and DCL on DOTA. 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.

5.4. 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. 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.

Acknowledgments

This work was partly supported by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102) and NSFC (72061127003, U20B2068). Xue Yang is partly supported by Wu Wen Jun Honorary Doctoral Scholarship, AI Institute, Shanghai Jiao Tong University.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283, 2016.

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

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

Chen, Z., Chen, K., Lin, W., See, J., Yu, H., Ke, Y., and Yang, C. Piou loss: Towards accurate oriented object detection in complex environments. *Proceedings of the European Conference on Computer Vision*, 2020.

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

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

Dowson, D. and Landau, B. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Fu, K., Chang, Z., Zhang, Y., Xu, G., Zhang, K., and Sun, X. 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.

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

Givens, C. R., Shortt, R. M., et al. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

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

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Jiang, Y., Zhu, X., Wang, X., Yang, S., Li, W., Wang, H., Fu, P., and Luo, Z. R2cnn: rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017.

Jiao, J., Zhang, Y., Sun, H., Yang, X., Gao, X., Hong, W., Fu, K., and Sun, X. A densely connected end-to-end neural network for multiscale and multiscene sar ship detection. *IEEE Access*, 6:20881–20892, 2018.

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

Knott, M. and Smith, C. S. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

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

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

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

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

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

Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., and Yan, J. 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.

Liu, Y., Zhang, S., Jin, L., Xie, L., Wu, Y., and Wang, Z. Omnidirectional scene text detection with sequential-free box discretization. *arXiv preprint arXiv:1906.02371*, 2019.

Liu, Z., Yuan, L., Weng, L., and Yang, Y. 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, pp. 324–331, 2017.

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

Ming, Q., Zhou, Z., Miao, L., Zhang, H., and Li, L. Dynamic anchor learning for arbitrary-oriented object detection. *arXiv preprint arXiv:2012.04150*, 2020.

Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., 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, pp. 1454–1459. IEEE, 2017.

Newell, A., Yang, K., and Deng, J. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pp. 483–499. Springer, 2016.

Olkin, I. and Pukelsheim, F. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

Pan, X., Ren, Y., Sheng, K., Dong, W., Yuan, H., Guo, X., Ma, C., and Xu, C. Dynamic refinement network for oriented and densely packed object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11207–11216, 2020.

Qian, W., Yang, X., Peng, S., Yan, J., and Guo, Y. Learning modulated loss for rotated object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. 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*, pp. 658–666, 2019.

Shi, X., Shan, S., Kan, M., Wu, S., and Chen, X. Real-time rotation-invariant face detection with progressive calibration networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2295–2303, 2018.

Song, Q., Yang, F., Yang, L., Liu, C., Hu, M., and Xia, L. Learning point-guided localization for detection in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020.

Tian, Z., Shen, C., and Chen, H. Conditional convolutions for instance segmentation. Springer, 2020.

Wang, J., Ding, J., Guo, H., Cheng, W., Pan, T., and Yang, W. 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.

Wang, J., Yang, W., Li, H.-C., Zhang, H., and Xia, G.-S. Learning center probability map for detecting objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020a.

Wang, X., Kong, T., Shen, C., Jiang, Y., and Li, L. Solo: Segmenting objects by locations. In *Proceedings of the European Conference on Computer Vision*, pp. 649–665. Springer, 2020b.

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

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. 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.

Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., and Luo, P. Polarmask: Single shot instance segmentation with polar representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12193–12202, 2020.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, 2017.

Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.-S., and Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

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

Yang, X., Sun, H., Fu, K., Yang, J., Sun, X., Yan, M., and Guo, Z. 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, 2018a.

Yang, X., Sun, H., Sun, X., Yan, M., Guo, Z., and Fu, K. Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network. *IEEE Access*, 6:50839–50849, 2018b.

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

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

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

Yang, X., Yan, J., Feng, Z., and He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021b.

Yi, J., Wu, P., Liu, B., Huang, Q., Qu, H., and Metaxas, D. Oriented object detection in aerial images with box boundary-aware vectors. *arXiv preprint arXiv:2008.07043*, 2020.

Zhang, G., Lu, S., and Zhang, W. 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.

Zhang, H., Wang, Y., Dayoub, F., and Sünderhauf, N. Swa object detection. *arXiv preprint arXiv:2012.12645*, 2020.

Zhao, P., Qu, Z., Bu, Y., Tan, W., Ren, Y., and Pu, S. Polardet: A fast, more precise detector for rotated target in aerial images. *arXiv preprint arXiv:2010.08720*, 2020.

Zheng, Y., Zhang, D., Xie, S., Lu, J., and Zhou, J. Rotation-robust intersection over union for 3d object detection. In *European Conference on Computer Vision*, pp. 464–480. Springer, 2020a.

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

Zhong, B. and Ao, K. Single-stage rotation-decoupled detector for oriented object. *Remote Sensing*, 12(19):3262, 2020.

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

Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., and Liang, J. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 5551–5560, 2017.

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

Zhu, Y., Du, J., and Wu, X. Adaptive period embedding for representing oriented objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.