
A Normalized Gaussian Wasserstein Distance for Tiny Object Detection

Jinwang Wang
 Electronic Information School
 Wuhan University
 jwwangchn@whu.edu.cn

Chang Xu
 Electronic Information School
 Wuhan University
 xuchangeis@whu.edu.cn

Wen Yang*
 Electronic Information School
 Wuhan University
 yangwen@whu.edu.cn

Lei Yu
 Electronic Information School
 Wuhan University
 ly.wd@whu.edu.cn

Abstract

Detecting tiny objects is a very challenging problem since a tiny object only contains a few pixels in size. We demonstrate that state-of-the-art detectors do not produce satisfactory results on tiny objects due to the lack of appearance information. Our key observation is that Intersection over Union (IoU) based metrics such as IoU itself and its extensions are very sensitive to the location deviation of the tiny objects, and drastically deteriorate the detection performance when used in anchor-based detectors. To alleviate this, we propose a new evaluation metric using Wasserstein distance for tiny object detection. Specifically, we first model the bounding boxes as 2D Gaussian distributions and then propose a new metric dubbed Normalized Wasserstein Distance (NWD) to compute the similarity between them by their corresponding Gaussian distributions. The proposed NWD metric can be easily embedded into the assignment, non-maximum suppression, and loss function of any anchor-based detector to replace the commonly used IoU metric. We evaluate our metric on a new dataset for tiny object detection (AI-TOD) in which the average object size is much smaller than existing object detection datasets. Extensive experiments show that, when equipped with NWD metric, our approach yields performance that is 6.7 AP points higher than a standard fine-tuning baseline, and 6.0 AP points higher than state-of-the-art competitors.

1 Introduction

Tiny objects are ubiquitous in many real world applications including driving assistance, large-scale surveillance, and maritime rescue. Even though object detection has achieved significant progress due to the development of deep neural networks [21, 15, 27], most of them are devoted to detecting objects with normal size. While tiny objects (less than 16×16 pixels in the AI-TOD dataset [29]) often exhibit with extremely limited appearance information, which increases difficulty in learning discriminative features, leading to enormous failure cases when detecting tiny objects [25, 29, 35].

Recent advances for tiny object detection (TOD) mainly focus on improving the feature discrimination [14, 37, 20, 12, 1, 19]. Some efforts have been devoted to normalizing the scale of input images to enhance the resolution of small objects and corresponding features [24, 25]. While the Generative Adversarial Network (GAN) is proposed to directly generate super-resolved representations for small

*Corresponding author

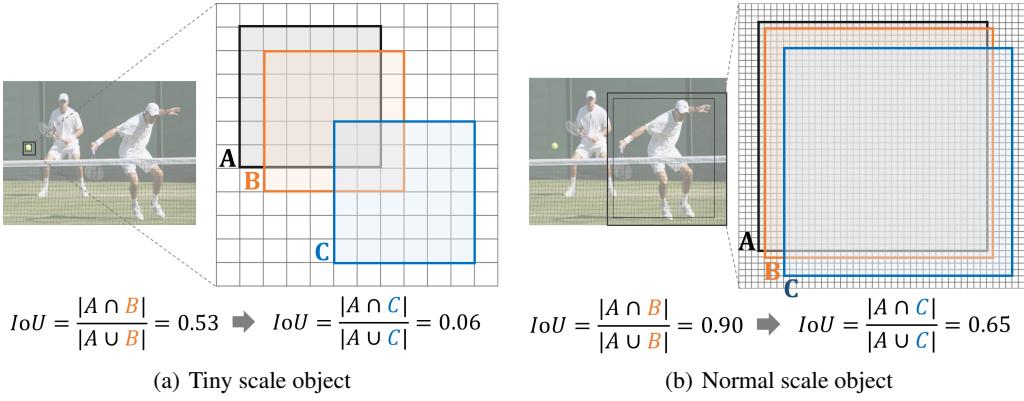


Figure 1: The sensitivity analysis of IoU on tiny and normal scale objects. Note that each grid denotes a pixel, box A denotes the ground truth bounding box, box B, C denote the predicted bounding box with 1 pixel and 4 pixels diagonal deviation respectively.

objects [12, 1, 19]. Besides, the Feature Pyramid Network (FPN) is proposed to learn multi-scale features to achieve scale-invariant detectors [14, 37, 20]. Indeed, existing approaches have improved TOD performance to some extent, but the precision boost is commonly achieved with additional cost.

In addition to learning discriminative features, the quality of the training sample selection plays an important role for anchor-based tiny object detectors [36] where the assignment of positive and negative (*pos/neg*) labels is essential. However, for tiny object, the properties of few pixels will increase the difficulty of training sample selection. As shown in Fig. 1, we can observe that the sensitivity of IoU to objects with different scales is of great variance. Specifically, for the tiny object with 6×6 pixels, a minor location deviation will lead to notable IoU drop (from 0.53 to 0.06), resulting in inaccurate label assignment. However, for the normal object with 36×36 pixels, the IoU changes slightly (from 0.90 to 0.65) with the same location deviation. In addition, Fig. 2 shows four IoU-Deviation curves with different object scales, the curve declines faster as the object size becomes smaller. It is worth noting that, the sensitivity of IoU results from the particularity that the location of bounding box can only change discretely.

This phenomenon implies that IoU metric is no longer invariant to object scale with discretized location deviations and finally leads to the following two flaws in label assignment. Specifically, IoU thresholds (θ_p, θ_n) are used to assign *pos/neg* training samples in anchor-based detectors, and (0.7, 0.3) are used in Region Proposal Network (RPN) [7]. Firstly, the sensitivity of IoU on tiny object makes a minor location deviation flip the anchor label, leading to *pos/neg* sample features' similarity and the network's difficulty in convergence. Secondly, we find that the average number of positive samples assigned to each ground-truth (*gt*) in AI-TOD dataset [29] is less than one using IoU metric since the IoU between some *gt* and any anchor is lower than minimum positive threshold. Therefore, there will be insufficient supervision information for training tiny object detectors. Although dynamic assignment strategies such as ATSS [36] can adaptively attain IoU thresholds for assigning *pos/neg* labels according to the statistical characteristics of objects, the sensitivity of IoU makes it difficult to find a good threshold and provide high-quality *pos/neg* samples for tiny object detectors.

Observing that IoU is not a good metric for tiny objects, in this paper, we propose a new metric to measure the similarity of bounding boxes by Wasserstein distance to replace standard IoU. Specifically, we firstly model the bounding boxes as 2-D Gaussian distributions, and then use our proposed Normalized Wasserstein Distance (NWD) to measure the similarity of derived Gaussian distributions. The major advantage of Wasserstein distance is that it can measure the distribution similarity even if there is no overlap or the overlap is negligible. In addition, the NWD is insensitive to objects with different scales and thus more suitable for measuring the similarity between tiny objects.

NWD can be applied to both single-stage and multi-stage anchor-based detectors. Besides, NWD can not only replace IoU in label assignment, but also replace IoU in Non-maximum Suppression (NMS) and regression loss function. Extensive experiments on a new TOD dataset AI-TOD [29] demonstrate that our proposed NWD can consistently improve the detection performance for all the detectors experimented. The contributions of this paper are summarized as follows.

- We analyze the sensitivity of IoU to location deviations of tiny objects, and propose NWD as a better metric for measuring the similarity between two bounding boxes.
- We design a powerful tiny object detector by applying NWD to label assignment, NMS and loss function in anchor-based detectors.
- Our proposed NWD can significantly improve TOD performance of the popular anchor-based detectors, and it achieves performance improvement from 11.1% to 17.6% on Faster R-CNN on AI-TOD dataset.

2 Related Work

2.1 Tiny Object Detection

Most of the previous small/tiny object detection methods can be roughly divided into three categories: multi-scale feature learning, designing better training strategy and GAN-based detection [28].

Multi-scale Feature Learning: A simple and classic way is to resize input images into different scales and to train different detectors, each of which can achieve best performance in a certain range of scales. To reduce the computation cost, some works [18, 14, 37] try to construct feature-level pyramid of different scales. For instance, SSD [18] detects objects from feature maps of different resolutions. Feature Pyramid Network (FPN) [14] constructs a top-down structure with lateral connections to combine feature information of different scales for improving object detection performance. After that, lots of methods are proposed to further improve FPN performance, including PANet [17], BiFPN [26], Recursive-FPN [20]. Besides, TridentNet [13] constructs a parallel multi-branch architecture with different receptive fields to generate scale-specific feature maps.

Designing Better Training Strategy: Inspired by the observation that it is difficult to detect tiny objects and large objects simultaneously, Singh *et al.* propose SNIP [24] and SNIPER [25] to selectively train objects within a certain scale range. Besides, Kim *et al.* [10] introduce Scale-Aware Network (SAN) and map the features extracted from different spaces onto a scale-invariant subspace, making detectors more robust to scale variation.

GAN-based Detectors: Perceptual GAN [12] is the first to attempt to apply GAN to small object detection, it improves small object detection through narrowing representation difference of small objects from the large ones. Besides, Bai *et al.* [1] propose a MT-GAN to train the image-level super-resolution model for enhancing the features of small RoIs. Furthermore, the work in [19] proposes a feature-level super-resolution approach to improve small object detection performance for proposal based detectors.

2.2 Evaluation Metric in Object Detection

IoU is the mostly widely used metric for measuring the similarity between bounding boxes. However, IoU can only work when the bounding boxes have overlap. To handle this problem, generalized IoU (GIoU) [22] is proposed by adding a penalty term of the smallest box converting bounding boxes. Nevertheless, GIoU will degrade to IoU when one bounding box contains another. Thus, DIoU [38] and CIoU [38] are proposed to overcome the limitations of IoU and GIoU by taking three geometric properties into account, *i.e.*, overlap area, central point distance and aspect ratio. GIoU, CIoU and DIoU are mainly applied in NMS and loss function to replace IoU for improving general object detection performance, but the application in label assignment is rarely discussed. In co-current work, Yang *et al.* [32] also propose a Gaussian Wasserstein Distance (GWD) loss for oriented object detection by measuring the positional relationship of oriented bounding boxes. However, the motivation of GWD is to solve the boundary discontinuity and square-like problem in oriented object detection. Our motivation is to alleviate the sensitivity of IoU for location deviations of tiny objects and our proposed method can replace IoU in all parts of anchor-based object detectors.

2.3 Label Assignment Strategies

It is a challenging task to assign high-quality anchors to *gt* boxes of tiny objects. A simple way is to lower the IoU threshold when selecting positive samples. Although it can make tiny objects match more anchors, the overall quality of training samples will deteriorate. Besides, many recent

works try to make the label assignment process more adaptive, aiming to improve the detection performance [6]. For instance, Zhang *et al.* [36] propose an Adaptive Training Sample Selection (ATSS) to automatically compute the *pos/neg* threshold for each *gt* by statistic value of IoU from a set of anchors. Kang *et al.* [9] introduce Probabilistic Anchor Assignment (PAA) by assuming that the distribution of joint loss for *pos/neg* samples follows the Gaussian distribution. In addition, Optimal Transport Assignment (OTA) [6] formulates the label assignment process as an Optimal Transport problem from a global perspective. However, these methods all use IoU metric to measure the similarity between two bounding boxes, and mainly focus on the threshold setting in the label assignment which are not suitable for TOD. In contrast, our method mainly focuses on designing a better evaluation metric which can be used to replace IoU metric in tiny object detectors.

3 Methodology

Inspired by the fact that IoU is actually the Jaccard similarity coefficient for computing similarity of two limited sample sets, we design a better metric for tiny objects based on Wasserstein Distance since it can consistently reflect the distance between distributions even if they have no overlap. Therefore, the new metric has better properties than IoU in measuring similarity between tiny objects. The details are as follows.

3.1 Gaussian Distribution Modeling for Bounding Box

For tiny objects, there tend to be some background pixels in their bounding boxes since most *real objects* are not strict rectangles. In these bounding boxes, foreground pixels and background pixels are concentrated on the center and boundary of the bounding boxes, respectively [30]. To better describe the weights of different pixels in bounding boxes, the bounding box can be modeled into two dimension (2D) Gaussian distribution, where the center pixel of bounding box has the highest weight and importance of the pixel decreases from the center to the boundary. Specifically, for horizontal bounding box $R = (cx, cy, w, h)$, where (cx, cy) , w and h denote the center coordinates, width and height, respectively. The equation of its inscribed ellipse can be represented as

$$\frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(y - \mu_y)^2}{\sigma_y^2} = 1, \quad (1)$$

where (μ_x, μ_y) is the center coordinates of the ellipse, σ_x, σ_y are the lengths of semi-axes along x and y axes. Accordingly, $\mu_x = cx, \mu_y = cy, \sigma_x = \frac{w}{2}, \sigma_y = \frac{h}{2}$.

The probability density function of a 2D Gaussian distribution is given by:

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))}{2\pi|\boldsymbol{\Sigma}|^{\frac{1}{2}}}, \quad (2)$$

where \mathbf{x} , $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the coordinate (x, y) , the mean vector and the co-variance matrix of Gaussian distribution. When

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = 1, \quad (3)$$

the ellipse in Eq. 1 will be a density contour of the 2D Gaussian distribution. Therefore, the horizontal bounding box $R = (cx, cy, w, h)$ can be modeled into a 2D Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{bmatrix} c_x \\ c_y \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix}. \quad (4)$$

Furthermore, the similarity between bounding box A and B can be converted to the distribution distance between two Gaussian distributions.

3.2 Normalized Gaussian Wasserstein Distance

We use the Wasserstein distance which comes from Optimal Transport theory to compute distribution distance. For two 2D Gaussian distributions $\mu_1 = \mathcal{N}(\mathbf{m}_1, \boldsymbol{\Sigma}_1)$ and $\mu_2 = \mathcal{N}(\mathbf{m}_2, \boldsymbol{\Sigma}_2)$, the 2nd order Wasserstein distance between μ_1 and μ_2 is defined as:

$$W_2^2(\mu_1, \mu_2) = \|\mathbf{m}_1 - \mathbf{m}_2\|_2^2 + \text{Tr} \left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2 \left(\boldsymbol{\Sigma}_2^{1/2} \boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2^{1/2} \right)^{1/2} \right), \quad (5)$$

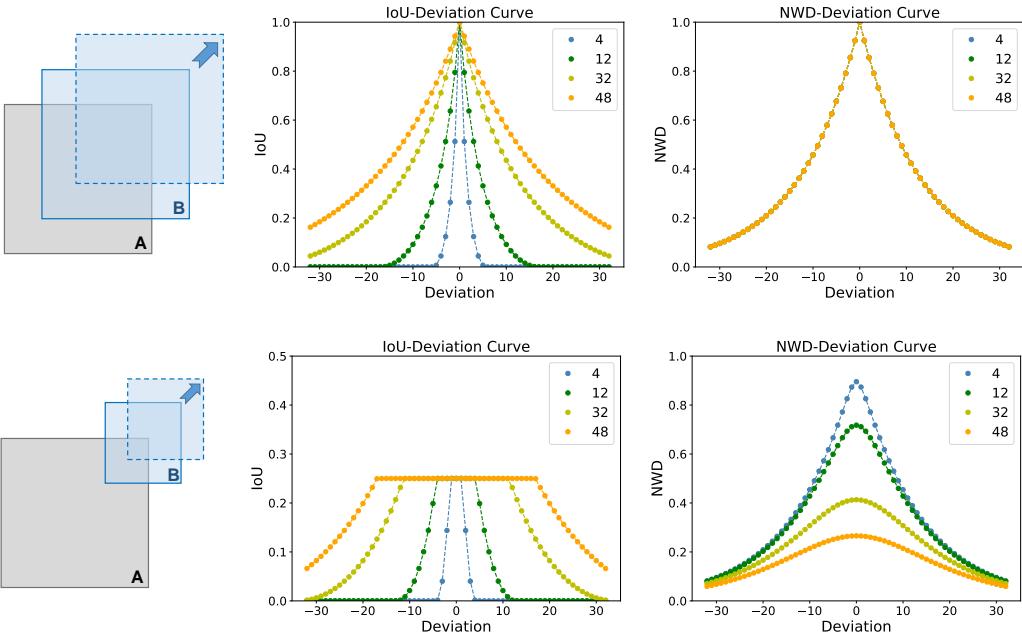


Figure 2: A comparison between IoU-Deviation Curve and NWD-Deviation Curve in two different scenarios. The abscissa value denotes the number of pixels deviation between the center points of A and B , the ordinate value denotes the corresponding metric value. Note that the location of bounding box can only change discretely, the Value-Deviation curve is presented in the form of scatter diagram.

and it can be simplified as:

$$W_2^2(\mu_1, \mu_2) = \|\mathbf{m}_1 - \mathbf{m}_2\|_2^2 + \left\| \Sigma_1^{1/2} - \Sigma_2^{1/2} \right\|_F^2, \quad (6)$$

where $\|\cdot\|_F$ is the Frobenius norm.

Furthermore, for Gaussian distributions \mathcal{N}_a and \mathcal{N}_b which are modeled from bounding boxes $A = (cx_a, cy_a, w_a, h_a)$ and $B = (cx_b, cy_b, w_b, h_b)$, Eq. 6 can be further simplified as:

$$W_2^2(\mathcal{N}_a, \mathcal{N}_b) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^T, \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2} \right]^T \right) \right\|_2^2. \quad (7)$$

However, $W_2^2(\mathcal{N}_a, \mathcal{N}_b)$ is a distance metric, and cannot be directly used as similarity metric (*i.e.*, a value between 0 and 1 as IoU). Therefore, we use its exponential form normalization and obtain the new metric dubbed Normalized Wasserstein Distance (NWD):

$$NWD(\mathcal{N}_a, \mathcal{N}_b) = \exp \left(-\frac{\sqrt{W_2^2(\mathcal{N}_a, \mathcal{N}_b)}}{C} \right), \quad (8)$$

where C is a constant closely related to the dataset. In the following experiments, we empirically set C to the average absolute size of AI-TOD and achieve the best performance. Moreover, we observe that C is robust in a certain range, details will be shown in supplementary materials.

Compared with IoU, NWD has the following advantages for detecting tiny objects: (1) scale invariance, (2) smoothness to location deviation, (3) the capability of measuring the similarity between non-overlapping or mutually inclusive bounding boxes. As shown in Fig. 2, without losing generality, we discuss the change of metric value in the following two scenarios. In the first row of Fig. 2, we keep box A and B the same scale and move away B along the diagonal of A . It can be seen that the four curves of NWD completely coincide, which indicates that NWD is insensitive to the scale variance of boxes. Moreover, we can observe that IoU is too sensitive to minor location deviation, but the NWD change resulting from location deviation is more smooth. The smoothness to location deviation indicates a possibility of a better distinction between pos/neg samples than IoU under the

same threshold. In the second row of Fig. 2, we set the side length of B to half of the side length of A and move away B along the diagonal of A . Compared with IoU, the curve of NWD is much more smooth and it can consistently reflect the similarity between A and B even if $|A \cap B| = A$ or B and $|A \cap B| = 0$.

3.3 NWD-based Detectors

The proposed NWD can be easily integrated into any anchor-based detector to replace IoU. Without loss of generality, the representative anchor-based Faster R-CNN is adopted to describe the usage of NWD. Specifically, all the modifications are conducted in the three parts which originally employ IoU, including *pos/neg* label assignment, NMS and regression loss function. The details are as follows.

NWD-based Label Assignment. Faster R-CNN [21] consists of two networks: RPN for generating region proposals and R-CNN [7] for detecting objects based on these proposals. The RPN and R-CNN both include label assignment process. For the RPN, anchors of different scales and ratios are firstly generated, and then binary labels are assigned to the anchors for training the classification and regression head. For the R-CNN, the label assignment process is similar with the RPN, and the difference is that the input of R-CNN is the output of RPN. In order to overcome the aforementioned shortcomings of IoU in tiny object detection, we design NWD-based label assignment strategy, which utilizes NWD to assign labels. Specifically, for training RPN, the positive label will be assigned to two kinds of anchors: (1) the anchor with the highest NWD value with a *gt* box and the NWD value is larger than θ_n or (2) the anchor that has the NWD value higher than the positive threshold θ_p with any *gt*. Accordingly, the negative label will be assigned to the anchor if its NWD value is lower than the negative threshold θ_n with all *gt* boxes. In addition, the anchors that are neither assigned positive labels nor negative labels do not participate in the training process. Note that, in order to apply NWD to anchor-based detectors directly, θ_p and θ_n as the original detectors are used in the experiments.

NWD-based NMS. NMS is an integral part of the object detection pipeline to suppress the redundant prediction bounding boxes, in which the IoU metric is applied. First, it sorts all prediction boxes based on their scores. The prediction box M with the highest score is selected and all other prediction boxes with a significant overlap (using a pre-defined threshold N_t) with M are suppressed. This process is recursively applied on the remaining boxes. However, the sensitivity of IoU to tiny object will make the IoU values lower than N_t for lots of prediction boxes, which further leads to false positive predictions. To handle this problem, we suggest that NWD is a better criterion for NMS in tiny object detection since NWD overcomes the scale sensitivity problem. Moreover, the NWD-based NMS is flexible to be integrated into any tiny object detector with only a few codes.

NWD-based Regression Loss. IoU-Loss [34] is introduced to eliminate the performance gap between training and testing [22]. However, IoU-Loss cannot provide gradient for optimizing network in the following two cases: (1) there is no overlap between the predicted bounding box P and the ground-truth box G (*i.e.*, $|P \cap G| = 0$) or (2) box P contains box G completely or vice versa (*i.e.*, $|P \cap G| = P$ or G). In addition, these two cases are very common for tiny objects. Specifically, on one hand, the deviation of a few pixels in P will cause no overlap between P and G , on the other hand, the tiny object is easy to be false predicted, leading to $|P \cap G| = P$ or G . Therefore, IoU-Loss is not suitable for tiny object detector. Although CIoU and DIoU can handle above two situations, they are sensitive to the location deviation of the tiny objects since they are both based on IoU. To handle above problems, we design the NWD metric as loss function by:

$$\mathcal{L}_{NWD} = 1 - NWD(\mathcal{N}_p, \mathcal{N}_g), \quad (9)$$

where \mathcal{N}_p is the Gaussian distribution model of prediction box P , \mathcal{N}_g is the Gaussian distribution model of *gt* box G . According to the introduction in Sec. 3.2, NWD-based loss can provide gradient even in both cases $|P \cap G| = 0$ and $|P \cap G| = P$ or G .

4 Experiments

We evaluate the proposed method on AI-TOD [29] and VisDrone2019 [4] datasets. The ablation study is conducted on AI-TOD, which is a challenging dataset designed for tiny object detection. It comes with eight categories, 700,621 object instances across 28,036 aerial images with 800×800 pixels. The mean absolute size of AI-TOD is only 12.8 pixels, which is much smaller than other object detection dataset like PASCAL VOC (156.6 pixels) [5], MS COCO (99.5 pixels) [16], and

Table 1: Comparison of different metrics in label assignment, NMS and loss function.

Metric	Assigning			NMS			Loss		
	AP	AP _{0.5}	AP _t	AP	AP _{0.5}	AP _t	AP	AP _{0.5}	AP _t
DIoU	5.4	11.3	3.9	11.2	26.8	7.8	10.7	25.1	6.7
CIoU	5.9	12.5	4.4	10.9	25.7	7.2	10.6	24.9	6.8
GIoU	11.0	26.5	7.7	11.5	26.5	7.6	10.9	25.1	6.9
IoU	11.1	26.5	7.8	11.1	26.5	7.8	10.8	25.3	7.1
NWD	16.1	43.8	17.4	11.9	27.5	8.0	12.1	27.5	8.9

DOTA (55.3 pixels) [31]. In addition, VisDrone2019 [4] is an UAV dataset for object detection. It consists of 10,209 images with 10 categories. VisDrone2019 has many complex scenes and large numbers of tiny objects since images are captured in different places at different height.

We adopt the same evaluation metric as AI-TOD [29] dataset, including AP, AP_{0.5}, AP_{0.75}, AP_{vt}, AP_t, AP_s and AP_m. Specifically, AP is averaged mAP across different IoU thresholds IoU={0.5, 0.55, ···, 0.95}, AP_{0.5} and AP_{0.75} are APs at IoU threshold of 0.5 and 0.75, respectively. In addition, AP_{vt}, AP_t, AP_s and AP_m are for *very tiny* (2-8 pixels), *tiny* (8-16 pixels), *small* (16-32 pixels) and *medium* (32-64 pixels) scale evaluation in AI-TOD [29].

We conduct all the experiments on a computer with 4 NVIDIA Titan X GPUs, and the codes are used for our experiments are based on MMdetection [3] code library. The ImageNet [23] pretrained ResNet-50 [8] with FPN [14] is used as the backbone, unless specified otherwise. All models are trained using the SGD optimizer for 12 epochs with 0.9 momentum, 0.0001 weight decay and 8 batch size. We set the initial learning rate as 0.01 and decay it at epoch 8 and 11 by a factor of 0.1. Besides, the batch size of RPN and Fast R-CNN are set to 256 and 512, respectively, and the sampling ratio of positive and negative samples is set to 1/3. The number of proposals generated by RPN is set to 3000. In the inference stage, we use the preset score 0.05 to filter out background bounding boxes, and NMS is applied with the IoU threshold 0.5. The above training and inference parameters are used in all experiments, unless specified otherwise.

4.1 Comparison with Other Metrics based IoU

There are some IoU-based metrics can be used to measure the similarity between bounding boxes as mentioned in Sec. 2. In this work, we re-implement the aforementioned four metrics (*i.e.* IoU, GIoU, CIoU and DIoU) and our proposed NWD on the same basic network (*i.e.* Faster R-CNN) to compare their performance on tiny objects. Specifically, they are applied in label assignment, NMS and loss function, respectively. Experimental results on AI-TOD dataset are shown in Tab. 1.

Comparison in label assignment. Note that the metric in assigning modules of RPN and R-CNN are both modified. It can be seen that NWD achieves the highest AP of 16.1% and improves 9.6% on AP_t when comparing with IoU metric, revealing that the NWD-based label assignment can provide more high quality training samples for tiny objects. In addition, to analyze the essential of the improvement, We make a group of statistical experiment. Specifically, we respectively calculate the average number of positive anchors matched by each *gt* box when using IoU, GIoU, DIoU, CIoU and NWD under the same default threshold, the number is 0.72, 0.71, 0.19, 0.19 and 1.05 respectively. It can be found that only NWD can ensure a considerable number of positive training samples. Moreover, although simply lowering the threshold of IoU-based metrics can provide more positive anchors for training, the performance of IoU-based tiny object detector after threshold fine-tuning is not better than the performance of NWD-based detector, which will be further discussed in supplementary materials. It attributes to the fact that NWD can solve the sensitivity of IoU to tiny object location deviation.

Comparison in NMS. We only modify the NMS module of RPN in this experiment since only the NMS in RPN can directly affect the training processing of detector. It can be seen that using different metrics to filter out redundant predictions during training can also affect the detection performance. Concretely, NWD achieves the best AP of 11.9%, which is 0.8% higher than the commonly used IoU. This implies that the NWD is a better metric for filtering out redundant bounding boxes when detecting tiny objects.

Table 2: Ablation experiments when NWD is applied to single module.

Method	Assigning		NMS		Loss		AP
	RPN	R-CNN	RPN	R-CNN	RPN	R-CNN	
Baseline							11.1
NWD	✓	✓	✓	✓	✓	✓	17.3
							14.3
							11.9
							10.8
							12.1
							12.4

Table 3: Ablation experiments when NWD is applied to multiple modules.

Method	Assigning		NMS		Loss		AP 12 epochs	AP 24 epochs
	RPN	R-CNN	RPN	R-CNN	RPN	R-CNN		
Baseline							11.1	12.6
NWD	✓		✓		✓	✓	17.8	19.7
		✓		✓		✓	13.8	16.8
	✓	✓	✓	✓	✓	✓	15.2	18.8

Comparison in loss function. Note that we modify the loss function both in RPN and R-CNN, which can both affect the convergence of the detector. It can also be seen that NWD-based loss function achieves the highest AP of 12.1%.

4.2 Ablation Study

In this section, Faster R-CNN [21] are used as the baseline, and it consists of two stages: RPN and R-CNN. Our proposed method can both be applied in the label assignment, NMS, loss function module of RPN and R-CNN, therefore there are totally six modules that can be switched from IoU metric to NWD metric. To verify the effectiveness of our proposed method in different modules, we make the following two groups of ablation study: comparison of applying NWD into one of the six modules and comparison of applying NWD into all modules in RPN or R-CNN.

Applying NWD into single module. Experimental results are shown in Tab. 2. Compared to baseline method, NWD-based assigning module in RPN and R-CNN respectively achieves the highest and second-highest AP improvement of 6.2% and 3.2%, which indicates that the problem of tiny object training label assignment resulting from IoU is the most noticeable, and our proposed NWD-based assignment strategy greatly improves assignment quality. It can also be observed that our proposed method improves the performance in 5 out of 6 modules, which significantly verifies the effectiveness of our NWD-based method. And the performance drop in NMS of R-CNN may owe to the fact that the default NMS threshold is sub-optimal, and it needs fine-tuning to boost the performance.

Applying NWD into multiple modules. Tab. 3 lists the experimental results. When training for 12 epochs, the detection performance all achieves significant improvement when using NWD in RPN, R-CNN or all modules. And the best performance of 17.8% is achieved when we apply NWD into all three modules of RPN. However, we find that when using NWD in all six modules, the AP has a drop of 2.6% compared with merely using NWD in RPN. In order to analyze the reason for performance drop, we add a group of experiments and train the network for 24 epochs. It can be seen that the performance gap decreases from 2.6% to 0.9%, which reveals that the network needs more time to converge when using NWD in R-CNN. Therefore, we only use NWD in RPN to achieve a considerable performance improvement with less time in the following experiments.

4.3 Main Results

To reveal the effectiveness of NWD on TOD, we conduct experiments on tiny object detection datasets AI-TOD [29] and VisDrone2019 [4].

Main results on AI-TOD. To verify that NWD can be applied into any anchor-based detector and boost TOD performance, we select five baseline detectors, including one-stage anchor-based detectors

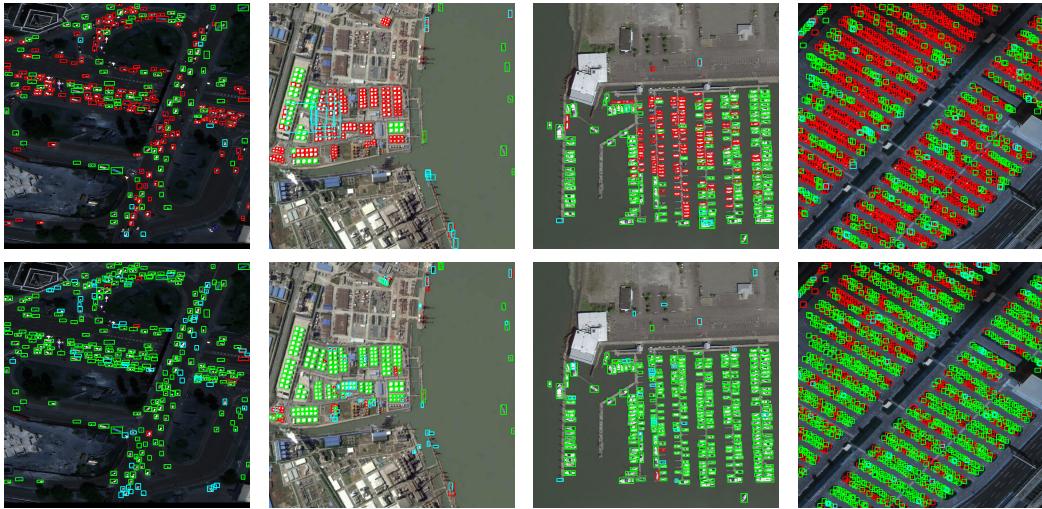


Figure 3: Visualization of detection results using IoU-based detector (first row) and NWD-based detector (second row) of AI-TOD dataset. The green, blue and red boxes denote true positive (TP), false positive (FP) and false negative (FN) predictions, respectively.

Table 4: Quantitative comparison of the baselines and NWD (with *) on AI-TOD test set.

Method	Backbone	AP	AP _{0.5}	AP _{0.75}	AP _{vt}	AP _t	AP _s	AP _m
SSD-512 [18]	ResNet-50	7.0	21.7	2.8	1.0	4.7	11.5	13.5
TridentNet [13]	ResNet-50	7.5	20.9	3.6	1.0	5.8	12.6	14.0
FoveaBox [11]	ResNet-50	8.1	19.8	5.1	0.9	5.8	13.4	15.9
PepPonits [33]	ResNet-50	9.2	23.6	5.3	2.5	9.2	12.9	14.4
FCOS [27]	ResNet-50	9.8	24.1	5.9	1.4	8.0	15.1	17.4
CenterNet [39]	DLA-34	13.4	39.2	5.0	3.8	12.1	17.7	18.9
M-CenterNet [29]	DLA-34	14.5	40.7	6.4	6.1	15.0	19.4	20.4
RetinaNet [15]	ResNet-50	4.7	13.6	2.1	2.0	5.4	6.3	7.6
RetinaNet*	ResNet-50	9.2	24.9	5.0	3.2	10.0	13.1	16.9
ATSS [36]	ResNet-50	12.8	30.6	8.5	1.9	11.6	19.5	29.2
ATSS*	ResNet-50	13.5	33.2	8.6	2.1	11.1	20.9	31.9
Faster R-CNN [21]	ResNet-50	11.1	26.3	7.6	0.0	7.2	23.3	33.6
Faster R-CNN*	ResNet-50	17.8	43.8	11.0	2.5	17.0	26.1	34.3
Cascade R-CNN [2]	ResNet-50	13.8	30.8	10.5	0.0	10.6	25.5	26.6
Cascade R-CNN*	ResNet-50	18.7	44.2	12.9	3.6	17.4	26.5	35.6
DetectoRS [20]	ResNet-50	14.8	32.8	11.4	0.0	10.8	28.3	28.0
DetectoRS*	ResNet-50	20.8	49.3	14.3	6.4	19.7	29.6	38.3

(*i.e.*, RetinaNet [15], ATSS [36]) and multi-stage anchor-based detectors (*i.e.*, Faster R-CNN [21], Cascade R-CNN [2], DetectoRS [20]). Experimental results are shown in Tab. 4. It can be seen that AP_{vt} of current state-of-the-art detectors is extremely low and close to zero, that means they cannot produce satisfactory results on tiny objects. In addition, our proposed NWD-based detectors improve AP metric of RetinaNet, ATSS, Faster R-CNN, Cascade R-CNN and DetectoRS by 4.5%, 0.7%, 6.7%, 4.9% and 6.0%, respectively. The performance improvement is even more obvious when objects are extremely tiny. It is worth noticing that NWD-based DetectoRS achieves state-of-the-art performance (20.8% AP) on AI-TOD. Some visualization results using IoU-based detector (first row) and NWD-based detector (second row) on AI-TOD dataset are shown in Fig. 3. We can observe that NWD can significantly reduce false negative (FN) compared with IoU.

Main results on Visdrone. Besides AI-TOD, we use VisDrone2019 [4] which contains many tiny objects with different scenarios to verify the generalization of NWD-based detectors. The results are shown in Tab. 5. It can be seen that NWD-based anchor-based detectors all achieve considerable improvements over their baselines.

Table 5: Quantitative comparison of the baselines and NWD (with *) on VisDrone2019 val set.

Method	Faster R-CNN	Faster R-CNN*	Cascade R-CNN	Cascade R-CNN*
AP _{0.5}	38.0	38.5	38.5	40.3
AP _{vt}	0.1	3.8	0.5	2.9
AP _t	6.2	10.2	6.8	11.1
AP _s	20.0	21.4	21.4	22.2

5 Conclusion

In this paper, we observe that IoU-based metrics is sensitive to the location deviation of tiny objects, which drastically deteriorates the tiny object detection performance. To handle this problem, we propose a new metric dubbed Normalized Wasserstein Distance (NWD) to measure the similarity between bounding boxes for tiny objects. Based on that, we further present a novel NWD-based tiny object detector by embedding NWD into label assignment, non-maximum suppression, and loss function of anchor-based detectors to replace original IoU metric. Experimental results show that our proposed method can improve the tiny object detection performance by a large margin and achieve state-of-the-art on AI-TOD dataset.

References

- [1] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *European Conference on Computer Vision*, pages 206–221. Springer, 2018.
- [2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.
- [3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, and et al. MMDetection: Open mmlab detection toolbox and benchmark. *CoRR*, abs/arXiv:1906.07155, 2019.
- [4] Dawei Du, Pengfei Zhu, Longyin Wen, and et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *IEEE International Conference on Computer Vision Workshops*, pages 213–226, 2019.
- [5] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.
- [6] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.
- [7] Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [9] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *European Conference on Computer Vision*, pages 355–371, 2020.
- [10] Yonghyun Kim, Bong-Nam Kang, and Daijin Kim. San: Learning relationship between convolutional features for multi-scale object detection. In *European Conference on Computer Vision*, September 2018.
- [11] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyond anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020.
- [12] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1222–1230, 2017.
- [13] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *IEEE International Conference on Computer Vision*, pages 6054–6063, 2019.
- [14] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.
- [15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

- [16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [17] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.
- [18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.
- [19] Junhyug Noh, Wonho Bae, Wonhee Lee, Jinhwan Seo, and Gunhee Kim. Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In *IEEE International Conference on Computer Vision*, pages 9725–9734, 2019.
- [20] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. *arXiv preprint arXiv:2006.02334*, 2020.
- [21] 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.
- [22] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.
- [23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [24] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3578–3587, 2018.
- [25] Bharat Singh, Mahyar Najibi, and Larry S Davis. Sniper: Efficient multi-scale training. In *Advances in Neural Information Processing Systems*, pages 9310–9320, 2018.
- [26] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020.
- [27] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *IEEE International Conference on Computer Vision*, pages 9627–9636, 2019.
- [28] Kang Tong, Yiquan Wu, and Fei Zhou. Recent advances in small object detection based on deep learning: A review. *Image and Vision Computing*, 97:103910, 2020.
- [29] Jinwang Wang, Wen Yang, Haowen Guo, Ruixiang Zhang, and Gui-Song Xia. Tiny object detection in aerial images. In *International Conference on Pattern Recognition*, pages 3791–3798, 2021.
- [30] 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*, 59(5):4307–4323, 2020.
- [31] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018.
- [32] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. In *International Conference on Machine Learning*, 2021.
- [33] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *IEEE International Conference on Computer Vision*, pages 9657–9666, 2019.
- [34] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. pages 516–520, 2016.
- [35] Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *IEEE Workshops on Applications of Computer Vision*, pages 1257–1265, 2020.
- [36] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9759–9768, 2020.
- [37] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. In *AAAI Conference on Artificial Intelligence*, pages 9259–9266, 2019.

- [38] 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 *AAAI Conference on Artificial Intelligence*, volume 34, pages 12993–13000, 2020.
- [39] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *CoRR*, abs/arXiv:1904.07850, 2019.