Dot Distance for Tiny Object Detection in Aerial Images

Chang Xu*, Jinwang Wang*, Wen Yang†, Lei Yu
School of Electronic Information, Wuhan University
Wuhan, China
{xuchangeis, jwwangchn, yangwen, ly.wd}@whu.ehu.cn

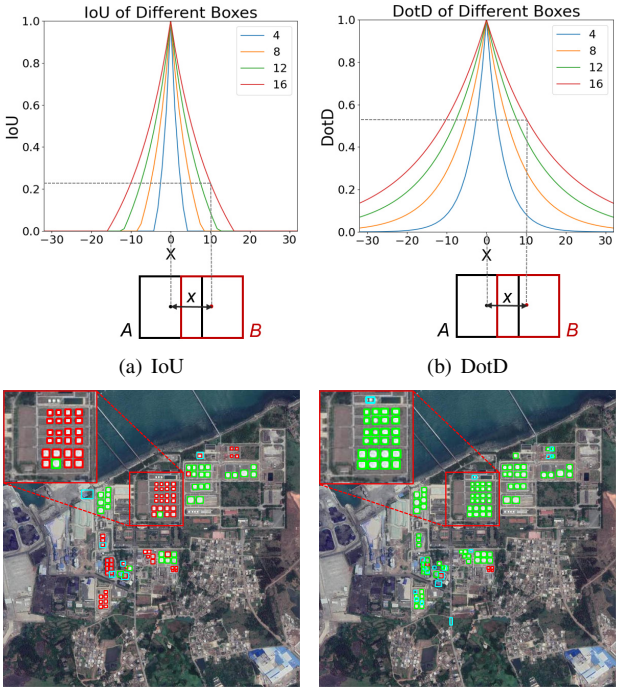
Abstract

Object detection has achieved great progress with the development of anchor-based and anchor-free detectors. However, the detection of tiny objects is still challenging due to the lack of appearance information. In this paper, we observe that Intersection over Union (IoU), the most widely used metric in object detection, is sensitive to slight offsets between predicted bounding boxes and ground truths when detecting tiny objects. Although some new metrics such as GIoU, DIoU and CIoU are proposed, their performance on tiny object detection is still below the expected level by a large margin. In this paper, we propose a simple but effective new metric called Dot Distance (DotD) for tiny object detection where DotD is defined as normalized Euclidean distance between the center points of two bounding boxes. Extensive experiments on tiny object detection dataset show that anchor-based detectors' performance is highly improved over their baselines with the application of DotD.

1. Introduction

Object detection is one of the main branches of computer vision. Recently, with the development of Convolutional Neural Network (CNN), the performance of object detection has achieved great progress. However, tiny object detection, which is a common-seen issue in remote sensing, driving assistance and surveillance [15, 31], is still challenging. Different from *small*, *medium* and *large* objects defined in MS COCO benchmark [18], tiny object is defined as the object whose size is smaller than 16×16 pixels as AI-TOD benchmark [31] in this work. Tiny objects are featured by fewer pixels and are easier to be confused with backgrounds compared with larger objects. Therefore normal object detectors have poor performance on tiny object detection.

To obtain better performance on tiny object detection, much effort is put in designing special networks to learn more appropriate features from tiny objects [16, 19]. For instance, Pyramidbox [27] combines high-order seman-



(a) IoU (b) DotD

(c) IoU-based Detection Results (d) DotD-based Detection Results

Figure 1. Comparison of IoU-based and DotD-based detectors. (a) and (b) are the IoU-X and DotD-X curves, respectively. Without losing generality, assuming there are two horizontal square bounding boxes A and B with same size in each sub-figure, the center of bounding box A is fixed at the origin of coordinates, and bounding box B moves along horizontal direction. The value of the curve denotes the IoU or DotD of A and B, and the vertical projection of any point on the curve on the X axis is the center point of box B when the curve is at this value. Different colors mean different bounding box sizes. (c) and (d) are the detection results by using IoU or DotD. Detection results marked with green, blue and red boxes denote true positive (TP), false positive (FP) and false negative (FN) predictions respectively.

*The first two authors contributed equally to this work.
†Corresponding author.

tic information with low-order geometric features. Yu *et al*. [34] align the scale distribution of dataset for network pre-training and the dataset for detector learning. Nevertheless, not only are special detectors necessary, essential theoretical and mathematical analysis of the fundamental reason for deterioration of normal detectors on tiny objects is also indispensable. It is surprising that Intersection over Union (IoU), which is the most widely used metric in all kinds of detectors, has been used as the metric of defining positive and negative samples without change for a long time. However, IoU is not suitable for tiny object detection. Fig. 1(a) shows the IoU curves with different bounding box sizes, each curve is drawn by keeping the bounding box A fixed and moving the bounding box B along horizontal direction. We can see that the smaller the scale of bounding box, the faster the curve drops. This property implies that the minor changes in the positional relationship between bounding boxes may lead to a great change of IoU value for tiny objects, and the use of IoU will deteriorate the performance on tiny object detection.

Specifically, IoU plays an important role in anchor-based object detectors. Anchor-based detectors first preset a number of pre-defined anchors on the image, then classify the categories and regress the coordinates of these anchors, finally output refined anchors as detection results [35]. For instance, in the Region Proposal Network (RPN) [23] which first proposes the anchor mechanism, the anchors will be classified as positive and negative samples according to their IoU with ground-truth bounding boxes. However, the sensitivity of IoU on tiny objects leads to the result that many positive anchors are classified as negative anchors in the assignment process. Besides, in the Non-Maximum Suppression (NMS) module, IoU is applied to decide whether a predicted bounding box should be classified as the false prediction. Nevertheless, the sensitivity of IoU on tiny objects will make the NMS module treat some true bounding box predictions as redundant bounding boxes. Therefore, IoU is not a good criterion for suppressing redundant bounding boxes of tiny size.

To handle the weaknesses of IoU, we propose a new metric called Dot Distance (DotD). The DotD is defined as the normalized Euclidean distance between two center points of bounding boxes. Fig. 1(b) shows the DotD curves of different bounding box sizes. We can see that when B moves far away from A, the DotD value decreases slower than that of IoU, especially when the object size is very tiny. Therefore, the use of DotD in RPN can provide more high quality positive samples for training anchor-based detectors. Fig. 1(c) and Fig. 1(d) show the detection results by using IoU and DotD, we can see that the DotD-based detector can detect more tiny objects than IoU-based detector. The comparison results between DotD and other metrics on different anchor-based detectors show our proposed DotD is more suitable for tiny object detection.

The main contributions of this work can be summarized as follows:

- We analyze the properties of different metrics on tiny objects and propose a new metric called DotD to overcome the weaknesses of traditional metrics on tiny objects.
- We apply the DotD to positive and negative assignment module in RPN and NMS module. Experimental results show our proposed DotD achieves notable improvement over baselines and obtains the state-of-the-art on AI-TOD dataset.

2. Related Work

Current CNN-based object detection methods can be classified into anchor-based and anchor-free detectors. Anchor-based detectors densely preset anchors of different sizes on the image and then classify and regress these anchors to output final predictions. Different from anchor-based detectors, anchor-free detectors directly generates bounding boxes from key points or center points to detect objects. Current state-of-the-art results are still held by anchor-based methods on standard detection benchmarks [35], but anchor-free detectors have less computational cost. For instance, anchor-based detectors include Faster R-CNN [23], Cascade R-CNN [2], SSD [19], RetinaNet [17], *etc*., and anchor-free detectors include CenterNet [6], CornerNet [11], FCOS [28], YOLO series [21, 9, 22, 1], *etc*. In the following part, we will discuss current tiny object detectors and metrics in detail.

2.1. Tiny Object Detection

We will review tiny object detection methods from three aspects, including multi-scale feature learning, context-based detection and designing better training strategy [29].

Multi-scale Feature Learning: Image pyramid is a classic way of scale transformation, the original image is up-sampled or down-sampled to obtain a series of images with different sizes to construct different scale spaces, improving tiny object detection performance. Besides, Feature Pyramid Network (FPN) [16], combines feature information between different feature layers can improve the tiny object detection performance without introducing much additional overhead. DMNet [12] generates a density map and learns scale information based on density intensities to crop regions with objects, and then cropped images are resized to larger size for training.

Context-based Detection: Context information plays a crucial role in tiny object detection. For instance, Hu *et al*. [8] propose the relationship network by using appearance and geometric features to establish an association model between objects, which improves the detection performance of tiny objects to a certain extent. Pyramid-

box [27] uses a semi-supervised method to supervise high-order semantic feature learning, and combines high-order semantic information with low-order geometric features to improve the accuracy of tiny face detection. To enhance the accuracy of detecting tiny objects, Chen *et al*. [3] use the context patch that is in parallel to the proposal patch produced from RPN and augments R-CNN.

Designing Better Training Strategy: A simple but effective way is to lower the threshold of IoU when defining positive and negative samples in RPN. It can make matching easier, but meanwhile introduces some low quality anchors. Furthermore, Zhang *et al*. [35] propose an Adaptive Training Sample Selection (ATSS) strategy to automatically select positive and negative samples according to the statistical characteristics of objects. The work in [10] proposes an anchor assignment strategy which adaptively separates anchors into positive and negative samples in a probabilistic manner. Besides, Singh *et al*. [26] propose the scale normalization method SNIP, which selectively trains objects within a certain scale range. SNIP solves the problem of network performance deterioration results from the dramatic changes in the object size to some extent. In addition, Yu *et al*. [34] propose Scale Match and align the scale distribution of dataset for network pre-training and the dataset for detector learning.

In this paper, we mainly concentrate on methods of designing better training strategy for tiny objects and our proposed method can be classified into this category. We use distances between center points of tiny boxes to constrain the selection of positive and negative anchors and improve the quality of NMS, ameliorating the quality of anchor-based tiny object detectors comprehensively.

2.2. Metrics in Object Detection

As the most widely used metric in object detection, IoU has limitations for evaluating the positional relationship between two bounding boxes. Precisely, if there is no overlap between two bounding boxes, IoU will always be equal to zero and it can no more reflect the distance of two bounding boxes. Therefore, some new metrics are established. On the basis of IoU, GIoU [24] is introduced by adding the weight of the minimum closure of two bounding boxes. Besides, Zheng *et al*. introduce the center point distance between two bounding boxes and the length of the diagonals of the smallest closure based on IoU to structure DIoU [36], and take width and height of boxes into account based on DIoU to form CIoU [36]. However, on the one hand, these improvements are essentially fine-tuning of IoU. On the other hand, limited by mathematical property of non-normalized value range, GIoU, DIoU, CIoU are initially designed as loss function [36, 24]. These improvements do not essentially solve the problem that tiny objects are sensitive to IoU.

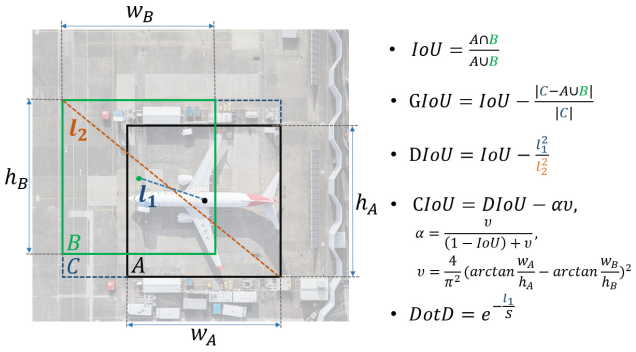


Figure 2. A Brief Comparison of Metrics.

3. Analysis of Metrics on Tiny Object Detection

Fig. 2 shows four commonly used evaluation metrics (*i.e*., IoU, GIoU, DIoU, and CIoU) for evaluating the positional relationship between two bounding boxes (*e.g*., a ground-truth bounding box and a prediction bounding box). The details of these metrics are described as follows.

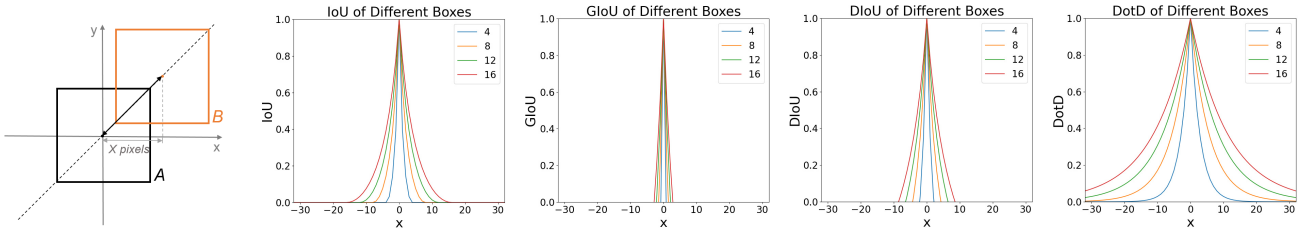
IoU Metric: IoU is the most popular metric in object detection which can be represented as:

$$IoU = \frac{A \cap B}{A \cup B}, \tag{1}$$

where A, B represents two bounding boxes, *e.g*., A and B are the ground-truth and predicted bounding boxes, respectively in object detection. The value of IoU reflects the overlap ratio of A and B, and the value range is $[0, 1]$. Due to the property that IoU is a normalized distance measure, it is a great metric to serve in the positive and negative anchor assigning module of RPN and the NMS module. However, IoU has some weaknesses. On the one hand, it can not reflect the position relationship if two bounding boxes are in vicinity of each other or very far from each other [24]. On the other hand, as shown in Fig. 1(a), minor relative movement between two tiny bounding boxes may lead to a great change of IoU. This implies that IoU is too sensitive to be used as the metric in tiny object detection. Due to the sensitivity, plenty of positive samples are wrongly divided into negative samples in RPN, the training becomes hard to converge, and the accuracy on tiny object detection is much lower than large object detection. Meanwhile, the sensitivity will make NMS treat some true predicted bounding boxes as redundant bounding boxes in tiny object detection, the detection accuracy is further reduced.

GIoU Metric: Generalized Intersection over Union (GIoU) [24] is based on IoU. It uses the minimum closure of the two bounding boxes A, B to solve the problems of IoU mentioned above. It is represented as follows:

$$GIoU = IoU - \frac{|C - A \cup B|}{|C|}, \tag{2}$$



(a) IoU (b) GIoU (c) DIoU (d) DotD

Figure 3. IoU, GIoU, DIoU and DotD of different sizes of bounding boxes. Without losing generality, assuming there are two horizontal square bounding boxes A, B with same size under each graph, the center of bounding box A is fixed at the origin of coordinates. Box B moves diagonally as shown in the left figure. The abscissa value is the number of pixels offset between the center points of A and B. The ordinate value denotes the IoU, GIoU, DIoU and DotD of box A and box B. And curves of different colors as marked in the graph denote different side length of boxes.

where C is the smallest box covering bounding box A and B. GIoU is originally designed as loss function, and its range is $[-1, 1]$, which indicates that it does not have a normalized form as IoU. Hence, it is difficult to serve as metric of determining positive and negative samples and fit into RPN with no change. With the introduction of the weight of other factors, GIoU is even more sensitivity to tiny offsets between tiny objects as shown in Fig. 3(b).

DIoU Metric: Distance Intersection over Union (DIoU) [36] takes the distance between the target and the anchor, overlap rate and scale into consideration:

$$DIoU = IoU - \frac{l_1^2}{l_2^2}, \tag{3}$$

where l_1 is the Euclidean distance between the center of A and B, and l_2 is the diagonal length of the smallest enclosing box covering the two boxes. When it is applied in loss function, the regression of the target box becomes faster and more stable. Different from GIoU, DIoU can also be applied in NMS.

CIoU Metric: Complete Intersection over Union (CIoU) [36] takes aspect ratio into consideration on the basis of the three metrics above:

$$CIoU = DIoU - \alpha\upsilon, \tag{4}$$

where α is a positive trade-off parameter:

$$\alpha = \frac{\upsilon}{(1 - IoU) + \upsilon}, \tag{5}$$

where υ measures the consistency of aspect ratio:

$$\upsilon = \frac{4}{\pi^2}(arctan\frac{w_A}{h_A} - arctan\frac{w_B}{h_B})^2, \tag{6}$$

where w_A, h_A denote the width and height of bounding box A, w_B, h_B denote the width and height of bounding box B. When the aspect ratio of the two boxes is the same, CIoU will degenerate to DIoU. DIoU has a fatal flaw that the denominator of α is equal to zero in the case where the two boxes are completely coincident. DIoU and GIoU are proposed by the same author, and the original purpose is to improve the performance of loss function. Therefore, they have the same problem as GIoU that they are hard to serve as threshold with little modification. It can be easily found from Fig. 3(c) that their curves drop even faster than IoU and it can also be seen in Tab. 1 that these new metrics brings no improvement to the network when serving as threshold of defining positive and negative anchors. Therefore, adding more weights on the basis of IoU can not essentially solve the problem.

In summary, IoU is a good metric of threshold, however, due to its sensitivity, it will deteriorate the performance of tiny object detectors. GIoU, DIoU, CIoU are modified metrics based on IoU, they have solved some problems in IoU-based loss function, and DIoU can be used in NMS. On the one hand, the range of GIoU, DIoU, CIoU is [-1,1], (-1,1], [-1,1] respectively, they do not have a normalized form and are hard to serve as threshold with little change. On the other hand, they have not solved the problem of IoU on tiny objects fundamentally. It is urgent to especially design a new metric for tiny objects.

4. Dot Distance for Tiny Object Detection

4.1. Definition of DotD

The absolute size and relative size of object A are calculated as:

$$AS(A) = \sqrt{w_A \times h_A} \tag{7}$$

$$RS(A) = \sqrt{\frac{w_A \times h_A}{W \times H}} \tag{8}$$

where AS is the abbreviation of absolute size, RS is the abbreviation of relative size. w_A, h_A denote the width and height of bounding box A, W, H denote the width and height of image [34].

We propose Dot Distance (DotD) based on the feature that the absolute size and relative size of tiny bounding

boxes are much smaller than medium or large bounding boxes. For instance, the average absolute size and relative size of tiny object detection dataset AI-TOD is 12.8 pixels, 0.016 respectively, and some examples of tiny objects is shown in Fig. 4. It can be seen that tiny objects can be viewed as points and the significance of width and height is much lower than that of the position of center points. Therefore, we define the DotD as:

$$DotD = e^{-\frac{D}{S}} \tag{9}$$

where D denotes the Euclidean distance between the center of two horizontal bounding boxes, and S denotes the average size of all objects in a certain dataset. In order to obtain a value range between 0 and 1, we use exponential form to normalize it. The expressions of D and S are shown below:

$$D = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2} \tag{10}$$

$$S = \sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} w_{ij} \times h_{ij}}{\sum_{i=1}^{M} N_i}} \tag{11}$$

where (x_A, y_A) and (x_B, y_B) respectively denote the center coordinate of bounding boxes A and B, M denotes the number of images in a certain dataset, N_i denotes the number of labeled bounding boxes in the i-th image, w_{ij}, h_{ij} denote the width and height of j-th bounding box in i-th image.

The proposed DotD inherits some properties from IoU, it also has its own characteristics:

- They both have a normalized form. The range of IoU and DotD is [0,1] and (0,1], respectively. When the centers of two bounding boxes coincide, $DotD = 0$. When distance between centers of two bounding boxes is far, $DotD \to 1$.
- IoU evaluates the positional relationship between whole bounding boxes, but DotD only concentrates on the positional relationship between center points which is more suitable for tiny objects of absolute size less than 16 pixels.
- IoU is sensitive to slight offset between two bounding boxes, but DotD is insensitive to slight offset between two bounding boxes and the curve descends slowly from the peak as shown in Fig. 1(b).

4.2. DotD-based Detectors

DotD can be easily integrated into many parts of anchor-based detectors. In traditional RPN, anchors of different sizes are firstly generated, then positive and negative anchors are defined according to their IoU with ground-truth bounding boxes, and then positive samples are selected for further regression. DotD can serve as the threshold of deciding positive and negative anchors for further regression, which conquers the problem that some potential positive anchors are defined as negative anchors under the IoU metric, and experiments have indicated that its sampling quality on tiny objects outperforms some other methods.

Figure 4. Samples of annotated images in AI-TOD dataset. Best viewed in color and zoomed in.

In the post-processing, DotD is a better metric for tiny bounding box NMS, and central point distance between two tiny boxes is much more important than their width and height when suppressing redundant boxes. For the predicted box N with the highest score, the DotD-NMS can be formally defined as

$$s_i = \begin{cases} s_i, & DotD(N, B_i) < \varepsilon \\ 0, & DotD(N, B_i) \geq \varepsilon \end{cases} \tag{12}$$

where box B_i is removed only by distance between central points of two boxes, and s_i denotes the classification score and ε is the NMS threshold. Although DotD-NMS takes less factors into consideration than IoU and DIoU, it is well-applied to tiny objects and easy to be integrated into object detection pipeline, which is proved in Tab. 4.

5. Experiments

5.1. Experiment Setting

Dataset. Our proposed method is evaluated on AI-TOD [31] which is a challenging dataset for tiny object detection in aerial images. It comes with 700,621 object instances across 28,036 aerial images with 800 × 800 pixels. There are eight object categories: airplane (AI), bridge (BR), storage-tank (ST), ship (SH), swimming-pool (SP), vehicle (VE), person (PE), wind-mill (WM). The mean absolute size of AI-TOD is only 12.8 pixels, which is much smaller than other aerial image detection dataset like DOTA (55.3 pixels) [32] and DIOR (65.7 pixels) [13]. The AI-TOD defines object size (pixels) in the range of [2,8], [8,16], [16,32] and [32, 64] as *very tiny*, *tiny*, *small* and *medium*, respectively. In the analysis experiments, 11,214 images in the `train set` is used for training, and 2,804 images in the `validation set` is used for validation. To obtain the final performance compared with state-of-the-art methods, we also train and evaluate the final model on `trainval set` and `test set`, respectively. Note that,

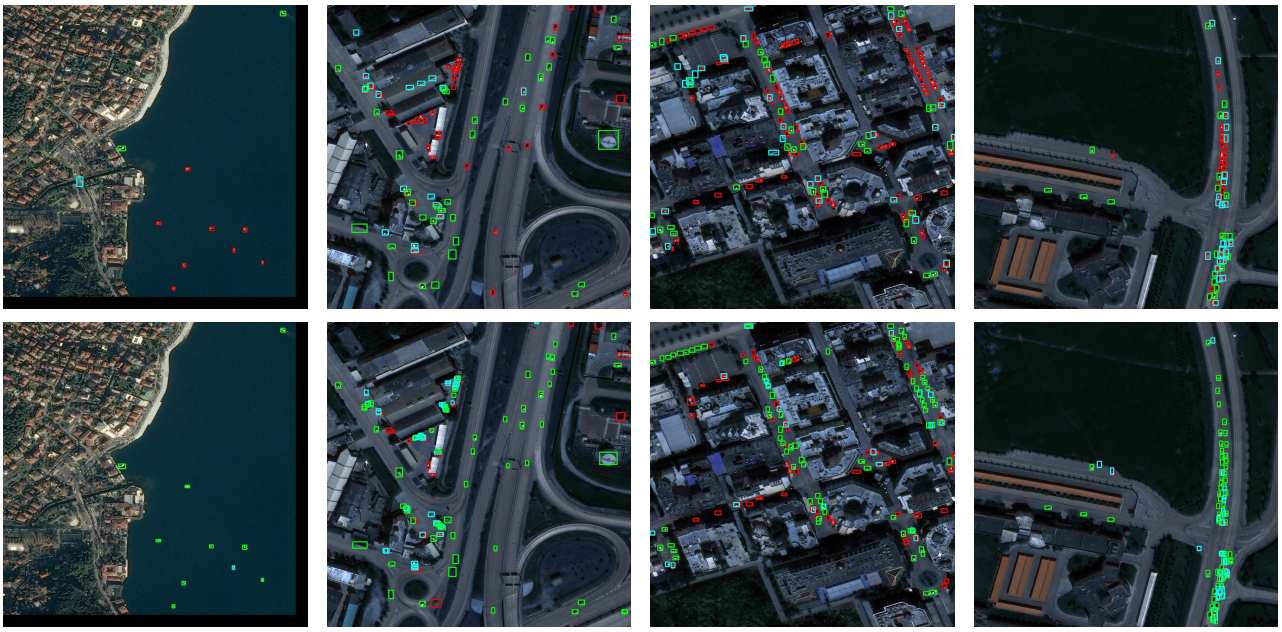


Figure 5. Example results from AI-TOD validation using IoU-based detector and DotD-based detector. The first row is the result of IoU-based detector, the second row is the result of DotD-based detector. Detection results marked with green, blue and red box denote true positive (TP), false positive (FP) and false negative (FN) predictions respectively.

there are 67.8% and 79.0% of objects larger than 16 pixels in DOTA and DIOR datasets, respectively, they are not suitable for training and evaluating tiny object detectors [31].

Implementation Details. We use the ImageNet [25] pretrained ResNet-50 [7] with FPN [16] as the backbone, unless specified otherwise. All experiments are based on MMDetection [4] code library and trained on a computer with one NVIDIA RTX 3090 GPU. All models are trained using the Stochastic Gradient Descent (SGD) optimizer for 12 epochs with 0.9 momentum, 0.0001 weight decay and 4 batch size. We set the initial learning rate as 0.005 and decay it at epoch 8 and 11. 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 to generate top 100 confident bounding boxes. The above training and inference parameters are used in all experiments, unless specified otherwise.

In addition, we use the commonly used AP (Average Precision) metric to evaluate the performance of the proposed method. Firstly, the precision and recall are given as follows:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

where TP, FP and FN denote the number of true positive, false positive, and false negative, respectively. We take $Recall$ as the abscissa and calculate the corresponding $Precision$ value as the ordinate, and plot the Precision-Recall curve. AP is defined as the area under Precision-Recall curve. Defined in MS COCO benchmark [18], AP_{50} means the IoU threshold of defining TP is 0.5, AP_{75} means the IoU threshold of defining TP is 0.75, AP means the average value from AP_{50} to AP_{95}, and the interval of IoU is 0.05. Note that AP_{50}, AP_{75} and AP take objects of all scales into consideration. Moreover, $\mathrm{AP}_{\mathrm{vt}}$, AP_{t}, AP_{s} and AP_{m} are for *very tiny, tiny, small* and *medium* scale evaluation in AI-TOD [31], $\mathrm{AP}_{\mathrm{sc}}$ is for *small* scale evaluation in MS COCO [18], respectively.

Metric	AP	AP_{50}	$\mathrm{AP}_{\mathrm{vt}}$	AP_{t}	AP_{s}	$\mathrm{AP}_{\mathrm{sc}}$
DIoU	3.8	7.8	0.0	2.6	13.7	3.9
CIoU	4.0	8.2	0.0	2.3	12.6	3.7
GIoU	6.0	14.7	0.0	2.3	12.6	3.7
IoU	7.8	18.0	0.0	5.2	**19.7**	7.0
DotD	**12.2**	**32.5**	**5.3**	**15.1**	14.9	**12.7**

Table 1. Comparison of DotD with other metrics on AI-TOD `validation set`. These metrics are used on both the positive and negative anchor assigning module and the NMS module in Faster R-CNN.

5.2. Comparison of Different Metrics

The commonly used metrics like IoU, GIoU [24], DIoU [36], CIoU [36] and our proposed DotD are introduced in Sec. 3 and Sec. 4, respectively. In this work, we reimplement the aforementioned metrics on the same basic

Method	PT	NT	MP	IoU	ATSS	DotD	AP	AP_{50}	AP_{vt}	AP_t	AP_s	AP_{sc}
Faster R-CNN [23]	0.7	0.3	0.3	✓			7.8	18.0	0.0	5.2	19.7	7.0
	0.5	0.1	0.1	✓			11.7	29.8	0.8	12.8	15.1	11.5
	-	-	-	✓	✓		11.7	28.4	1.4	10.2	**20.0**	11.0
	0.7	0.3	0.3			✓	**12.2**	**32.5**	**5.3**	**15.1**	14.9	**12.7**
Cascade R-CNN [2]	0.7	0.3	0.3	✓			9.0	20.9	0.0	6.7	**20.6**	8.1
	0.7	0.3	0.3			✓	**12.3**	**31.1**	**6.1**	**14.8**	13.8	**12.7**
Cascade RPN [30]	0.7	0.3	0.3	✓			9.4	21.7	2.1	9.2	**15.6**	9.2
	0.7	0.3	0.3			✓	**12.3**	**30.4**	**5.9**	**14.2**	14.9	**12.4**

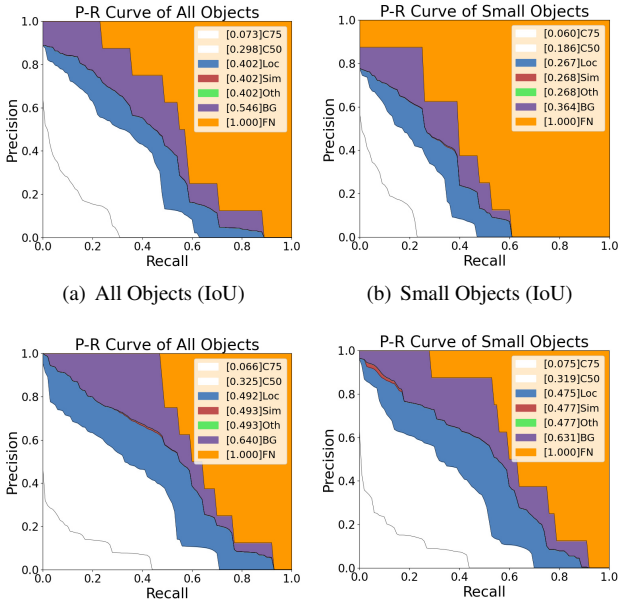
Table 2. Comparison with baseline detectors on AI-TOD `val set`. PT, NT, MP denote the positive threshold, negative threshold and minimum positive threshold of assigning anchors in RPN. Note that the evaluation criterion is IoU. The backbone is ResNet-50 with FPN.

detector Faster R-CNN [23] for a fair comparison. Note that the GIoU, DIoU and CIoU are used as loss functions in the original work, but these metrics are used on both the positive and negative anchor assigning module and the NMS module in this work. Specifically, the positive and negative anchor assigning module is designed for training RPN which will assign a binary class label to each anchor. We assign a positive label to an anchor when it has highest metric value with a ground-truth bounding box or has an metric value higher than 0.7 with any ground-truth bounding box. Meanwhile, we assign a negative label to a negative anchor if its metric value is lower than 0.3 for all ground-truth bounding boxes. Anchors that are neither positive nor negative do not contribute to the training. The NMS module is designed to filter out redundant bounding boxes by the metric value. Note that we keep the threshold parameters same as Faster R-CNN.

As shown in Tab. 1, we find that IoU-based metrics (*i.e.* DIoU, CIoU, GIoU and IoU) are worse than DotD on AP, we argue that the accuracy gap results from the difficulty in assigning high quality positive anchors to ground truth in the training stage. On the one hand, for low signal-to-noise ratio tiny object, its boundary maybe confused with background which leads to the center distance is more important than width and height in the positive and negative sample assigning module. On the other hand, IoU-based metrics are sensitive to tiny objects as discussed in Sec. 3. Therefore, using DotD metric which only considers the center distance between bounding boxes can achieve better results.

Constant	AP	AP_{50}	AP_{vt}	AP_t	AP_s	AP_m
8.0	12.1	33.0	4.4	15.2	13.5	14.8
10.0	12.4	**34.2**	5.2	15.4	14.1	16.8
12.8	**12.7**	33.6	**5.5**	**15.5**	**16.0**	**17.8**
14.0	12.2	32.5	5.1	14.6	14.8	17.7
16.0	9.9	27.8	4.2	12.2	11.5	15.0
32.0	6.6	19.3	3.0	8.0	6.7	12.5

Table 3. Performance of different constant on AI-TOD. The average absolute size of AI-TOD is 12.8.



(a) All Objects (IoU) (b) Small Objects (IoU)

(c) All Objects (DotD) (d) Small Objects (DotD)

Figure 6. Precision-Recall Curve. Note that **All Objects** denotes objects of all sizes, **Small Objects** denotes objects smaller than 32×32 pixels. **IoU** or **DotD** in brackets denote the metric used in assignment and NMS module.

5.3. Improvements over Baseline Detectors

To demonstrate the effectiveness of the DotD-based detector in baseline networks, experiments are performed with three detectors: Faster R-CNN [23], Cascade R-CNN [2], and Cascade RPN [30]. To further verify its advantage, we compare it with ATSS [35] and simply lowering the threshold of IoU of RPN in Faster R-CNN. We tried ten different combinations of thresholds, and the best combination of positive threshold, negative threshold and minimum positive threshold of assigning anchors by IoU is 0.5, 0.1, 0.1. Experiment results in Tab. 2 have shown that although simply lowering the IoU can improve AP to some extent, the best performance of IoU-based detector after fine-tuning is not better than the performance of DotD-based without fine-

Detector	Assigning	NMS	Evaluation	AP	AP_{50}	AP_{vt}	AP_t	AP_s	AP_{sc}
Faster R-CNN [23]	IoU	IoU	IoU	7.8	18.0	0.0	5.2	**19.7**	7.0
	DotD	IoU	IoU	**12.7**	33.6	**5.5**	15.5	16.0	13.2
	DotD	DotD	IoU	12.2	**37.8**	4.8	**16.3**	18.3	**14.2**

Table 4. Ablation Study on AI-TOD `validation set`.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_{vt}	AP_t	AP_s	AP_m
RetinaNet [17]	ResNet-50-FPN	4.7	13.6	2.1	2.0	5.4	6.3	7.6
SSD-512 [19]	VGG-16	7.0	21.7	2.8	1.0	4.7	11.5	13.5
TridentNet [14]	ResNet-50	7.5	20.9	3.6	1.0	5.8	12.6	14.0
RepPoints [33]	ResNet-50-FPN	9.2	23.6	5.3	2.5	9.2	12.9	14.4
FCOS [28]	ResNet-50-FPN	9.8	24.1	5.9	1.4	8.0	15.1	17.4
Faster R-CNN [23]	ResNet-50-FPN	11.4	27.0	8.0	0.0	8.3	<u>23.1</u>	24.5
Grid R-CNN [20]	ResNet-50-FPN	12.2	27.7	9.0	0.2	10.3	22.6	23.3
CenterNet [37]	DLA-34	13.4	39.2	5.0	3.8	12.1	17.7	18.9
Cascade R-CNN [2]	ResNet-50-FPN	13.8	30.8	<u>10.5</u>	0.0	10.6	**25.5**	**26.6**
M-CenterNet [31]	DLA-34	14.5	**40.7**	6.4	6.1	15.0	19.4	20.4
Cascade RPN w/ DotD	ResNet-50-FPN	13.7	34.0	8.7	<u>6.9</u>	14.8	15.8	<u>24.6</u>
Faster R-CNN w/ DotD	ResNet-50-FPN	<u>14.9</u>	38.5	9.3	7.2	<u>16.1</u>	17.9	23.7
Cascade R-CNN w/ DotD	ResNet-50-FPN	**16.1**	<u>39.2</u>	**10.6**	**8.3**	**17.6**	18.1	22.1

Table 5. Performance of state-of-the-art detectors on the proposed AI-TOD `test set`. Bold and underline fonts indicate the best and the second best performances for each metric, respectively.

tuning. DotD on tiny object dataset AI-TOD outperforms the best performance of others by 0.5% AP, 2.7% AP_{50}, 2.3% AP_t in Faster R-CNN. It can also be concluded that the tinier the size of the object is, the more effective DotD is. A comparison between IoU-based detector and DotD-based detector is shown in Fig. 5.

Moreover, we find that besides AP, DotD-based detectors can also improve the average recall rate (AR) to a great extent, which indicates that the ratio of the number of correctly identified objects to the number of all objects in the validation set is higher. A comparison of Precision-Recall (P-R) Curve [5] of DotD with the best performance of IoU is shown in Fig. 6.

5.4. Ablation Study

In this section, we further verify the effectiveness of DotD by ablation study. Note that after fine-tuning, the NMS is applied with threshold of 0.2 per class to generate top 100 confident detections per image.

Hyper-parameter: In this paper, we use the average absolute size as the selection of hyper-parameter in DotD by default. Besides, we have also experimentally set the hyper-parameter to different constant. The results are shown in Tab. 3, we find the average absolute size is the best one in AI-TOD dataset, and the AP of larger objects is related to the hyper-parameter. In our future work, we will investigate the adaptive hyper-parameter in the case where the object scales change drastically and further boost the performance.

Different Parts of the Detector: In Faster R-CNN, we gradually replace IoU with DotD in positive and negative anchor assigning module and the NMS module. We can see from Tab. 4 that the comprehensive performance of detectors gradually improves with more parts switched to DotD.

5.5. Comparison with State-of-the-art Detectors

We compare our method on the AI-TOD with recently state-of-the-art object detectors. In this experiment, we use `trainval set` of AI-TOD for training and `test set` for validation as in [31]. Tab. 5 shows the comparison results, our proposed DotD improves Faster R-CNN [23], Cascade R-CNN [2] by 3.5%, 2.3% on AP respectively. Compared to existing state-of-the-art methods, our DotD-based detector outperforms M-CenterNet [31] by 1.6% on AP, 2.2% on AP_{vt} and 2.6% on AP_t.

6. Conclusion

In this paper, we propose a novel metric called Dot Distance (DotD) for tiny object detection. The proposed DotD gets rid of the problem that IoU is sensitive to slight offsets between bounding boxes when detecting tiny objects. Extensive experiments have been conducted to verify its effectiveness of defining positive and negative anchors and NMS on anchor-based detectors. When replacing IoU with DotD, object detectors achieve considerable improvements over baseline methods and obtain the state-of-the-art performance on tiny object dataset AI-TOD.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6154–6162, 2018.

[3] Chenyi Chen, Ming Yu Liu, Oncel Tuzel, and Jianxiong Xiao. R-cnn for small object detection. In *Asian Conference on Computer Vision (ACCV)*, 2017.

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[5] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *International Conference on Machine Learning (ICML)*, pages 233–240, 2006.

[6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6569–6578, 2019.

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

[8] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3588–3597, 2018.

[9] Ali Farhadi Joseph Redmon. Yolo9000: Better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[10] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. *arXiv preprint arXiv:2007.08103*, 2020.

[11] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[12] Changlin Li, Taojiannan Yang, Sijie Zhu, Chen Chen, and Shanyue Guan. Density map guided object detection in aerial images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, June 2020.

[13] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS J P&RS)*, 159:296–307, 2020.

[14] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6054–6063, 2019.

[15] Yundong Li, Han Dong, Hongguang Li, Xueyan Zhang, Baochang Zhang, and Zhifeng Xiao. Multi-block ssd based on small object detection for uav railway scene surveillance. *Chinese Journal of Aeronautics*, 33(6):1747–1755, 2020.

[16] 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 (CVPR)*, pages 2117–2125, 2017.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2980–2988, 2017.

[18] 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 (ECCV)*, pages 740—-755, 2014.

[19] 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 (ECCV)*, pages 21–37. Springer, 2016.

[20] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid r-cnn. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7363–7372, 2019.

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[22] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 39(6):1137–1149, 2017.

[24] 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 (CVPR)*, pages 658–666, 2019.

[25] 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 (IJCV)*, 115(3):211–252, 2015.

[26] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3578–3587, 2018.

[27] Xu Tang, Daniel K Du, Zeqiang He, and Jingtuo Liu. Pyramidbox: A context-assisted single shot face detector. In *European Conference on Computer Vision (ECCV)*, pages 797–813, 2018.

[28] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9627–9636, 2019.

[29] 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.

[30] Thang Vu, Hyunjun Jang, Trung X Pham, and Chang D Yoo. Cascade rpn: Delving into high-quality region proposal network with adaptive convolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[31] Jinwang Wang, Wen Yang, Haowen Guo, Ruixiang Zhang, and Gui-song Xia. Tiny object detection in aerial images. In *International Conference on Pattern Recognition (ICPR)*, 2021.

[32] 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 (CVPR)*, 2018.

[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 (ICCV)*, pages 9657–9666, 2019.

[34] Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1257–1265, 2020.

[35] 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 (CVPR)*, 2020.

[36] 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 *National Conference on Artificial Intelligence (AAAI)*, volume 34, pages 12993–13000, 2020.

[37] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *CoRR*, abs/1904.07850, 2019.