Similarity Distance-Based Label Assignment for Tiny Object Detection

Shuohao Shi[1,†], Qiang Fang[1,†,*], Xin Xu[1], Tong Zhao[1]

Abstract— Tiny object detection is becoming one of the most challenging tasks in computer vision because of the limited object size and lack of information. The label assignment strategy is a key factor affecting the accuracy of object detection. Although there are some effective label assignment strategies for tiny objects, most of them focus on reducing the sensitivity to the bounding boxes to increase the number of positive samples and have some fixed hyperparameters need to set. However, more positive samples may not necessarily lead to better detection results, in fact, excessive positive samples may lead to more false positives. In this paper, we introduce a simple but effective strategy named the Similarity Distance (SimD) to evaluate the similarity between bounding boxes. This proposed strategy not only considers both location and shape similarity but also learns hyperparameters adaptively, ensuring that it can adapt to different datasets and various object sizes in a dataset. Our approach can be simply applied in common anchor-based detectors in place of the IoU for label assignment and Non Maximum Suppression (NMS). Extensive experiments on four mainstream tiny object detection datasets demonstrate superior performance of our method, especially, 1.8 AP points and 4.1 AP points of very tiny higher than the state-of-the-art competitors on AI-TOD. Code is available at: https://github.com/cszzshi/SimD.

I. INTRODUCTION

With the popularization of drone technology and autonomous driving, applications of object detection are becoming increasingly widespread in daily life. General object detectors have achieved significant progress in both accuracy and detection speed. For example, the latest version of the YOLO series, YOLOv8, achieves a mean average precision (mAP) of 53.9 percent on the COCO detection dataset and takes only 3.53 ms to detect objects in an image when implemented on the NVIDIA A100 GPU using TensorRT. Nevertheless, despite this significant progress in general object detectors, when they are directly applied for tiny object detection tasks, their accuracy sharply decreases.

In a recent survey of small object detection, Cheng et al. [1] proposed dividing small objects into three categories (extremely, relatively and generally small) in accordance with their mean area. Two major challenges faced in tiny object detection are information loss and the lack of positive samples. There are many possible approaches to improve the accuracy of tiny object detection, such as feature fusion, data augmentation, and superresolution.

Because sufficiently numerous and high-quality positive samples are crucial for object detection, the label assignment

[1]National University of Defense Technology, Changsha, Hunan, China e867cda2b@126.com, qiangfang@nudt.edu.cn, xinxu@nudt.edu.cn, zhaotong@nudt.edu.cn

[†]These authors contributed equally and should be co-first authors
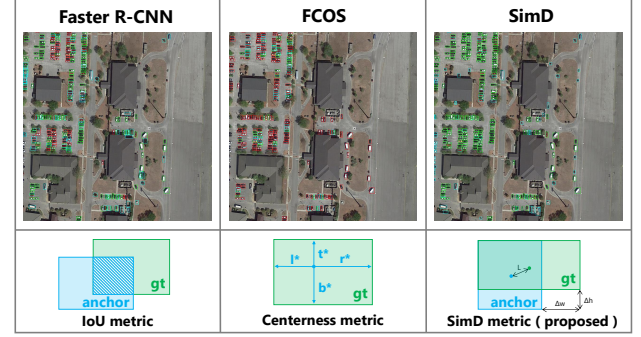
[*]corresponding author

Fig. 1.   Comparison between traditional label assignment metrics and our SimD metric. The first row shows typical detection results achieved with these methods, and the second row presents diagrammatic sketches of these metrics. The Δw and Δh in SimD respectively represent the difference of width and height between anchor and ground truth. The green, blue and red boxes respectively denote true positive (TP), false positive (FP) and false negative (FN) predictions.

strategy is a core factor affecting the final results. The smaller the bounding boxes are, the higher the sensitivity of the IoU metric [2], this is the main reason why it is not possible to label as many positive samples as tiny objects as can be labeled as general objects. A simple comparison between traditional anchor-based and anchor-free metrics and our SimD metric is shown in Fig. 1.

The current research on tiny object label assignment strategies mainly focuses on reducing the sensitivity to the bounding box size. From this perspective, Xu et al. [2] proposed using the Dot Distance (DotD) as the assignment metric in place of the IoU. Later, NWD [3] and RFLA [4] were proposed as attempts to model the ground truth and anchor as Gaussian distributions and then use the distance between these two Gaussian distributions to evaluate the two bounding boxes. In fact, these methods have enabled considerable progress in label assignment, but there are also some problems that they may not consider.

First, most of these methods focus on reducing the sensitivity to the bounding box size, thereby increasing the number of positive samples. However, as we know, excessive positive samples may have a detrimental impact on an object detector, leading to many false positives.

Second, the essence of these evaluation metrics is to measure the similarity between bounding boxes. For anchor-based methods, the similarity between the ground truth and the anchor is considered. This similarity includes two aspects: shape and location. However, some methods only consider the locations of the bounding boxes, the others consider both shape and location, but they also have a
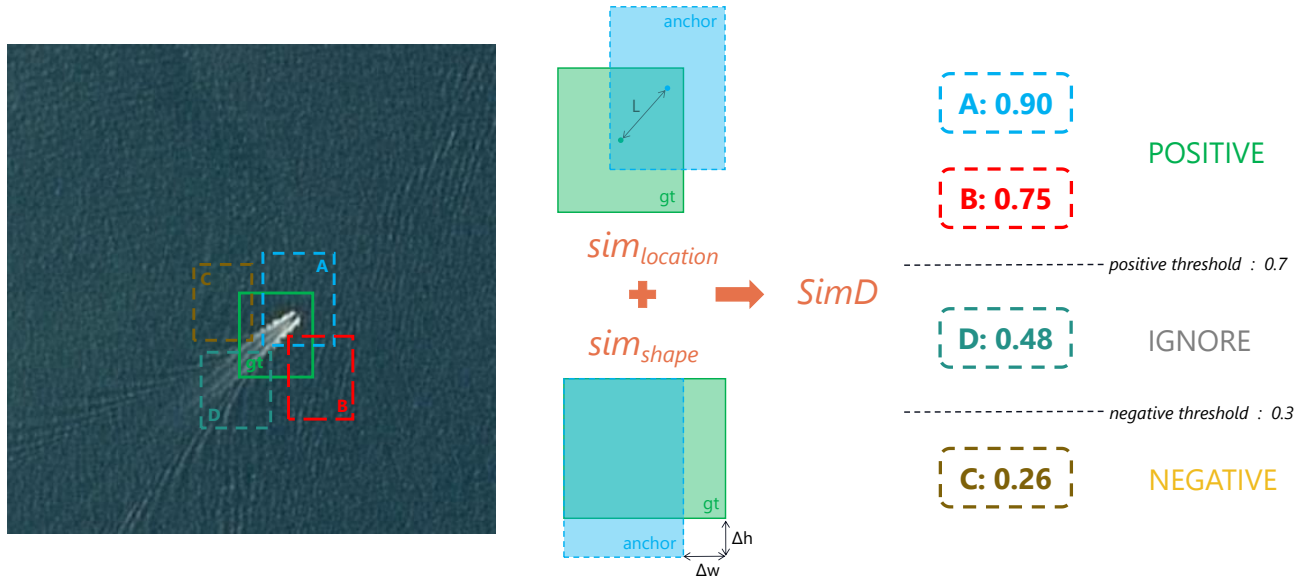
Fig. 2. The processing flow of the SimD-based label assignment strategy. We first obtain the coordinates of the ground truth and anchors and then calculate the Similarity Distance (SimD) between the ground truth and each anchor. Subsequently, we follow the traditional label assignment strategy to obtain positive and negative samples in accordance with corresponding thresholds. For a ground truth that does not have any associated positive sample based on this strategy, we assign the anchor with the maximum SimD value as a positive sample, as long as this SimD value is larger than a minimum positive threshold.

hyperparameter that needs to be chosen.

Finally, although the object sizes in a tiny object detection dataset tend to be fairly similar, there are still differences in the scales of different objects. For example, the sizes of the objects in the AI-TOD dataset range from 2 to 64 pixels. The discrepancy is more pronounced in the VisDrone2019 dataset, which contains both small and general-sized objects. In fact, the smaller the size of object, the more difficult it is to obtain positive samples. Unfortunately, most of the existing methods may pay less attention to this problem.

In this paper, to solve these problems, we introduce a new evaluation metric to take the place of the traditional IoU, the processing flow of our method is shown in Fig. 2. The main contributions of this paper include the following:

• We propose a simple but effective strategy named Similarity Distance (SimD) to evaluate the relationship between two bounding boxes. It not only considers both location and shape similarity but also can effectively adapt to different datasets and different object sizes in a dataset without the need to set any hyperparameter.

• Extensive experiments prove the validity of our method. We use several general object detectors and simply replace the IoU-based assignment module with the proposed method based on our SimD metric, in this way, we achieve the state-of-the-art performance on four mainstream tiny object detection datasets.

II. RELATED WORK

In recent years, applications of object detection have become increasingly widespread in various industries. This technology offers considerable convenience. For example, rescue operations can be quickly carried out by identifying

ground objects in remote sensing images. With the development of deep learning technology, especially the introduction of ResNet [5], the accuracy and speed of detection have significantly increased.

General object detectors can be divided into two categories: one-stage and two-stage detectors. A two-stage detector first generates a list of proposal regions and then determines the position and category of the object. Such algorithms include R-CNN [6], Fast R-CNN [7] and Faster R-CNN [8]. The structure of one-stage detectors is simpler. They can directly output the coordinates and category of an object from the input image. Some classic one-stage detectors include YOLO [9] and SSD [10].

A. Tiny Object Detection

Despite the significant progress in object detection achieved with deep learning technology, the detection accuracy will sharply decrease when the objects to be detected are tiny. Small objects are usually defined as objects with sizes less than a certain threshold value. For example, in Microsoft COCO [11], if the area of an object is less than or equal to 1024, it is considered a small object. In many cases, however, the objects of interest are in fact much smaller than the above definition. For example, in the AI-TOD dataset, the average edge length of an object is only 12.8 pixels, far smaller than in other datasets.

As stated in a previous paper [1], due to the extremely small size of the objects of interest, there are three main challenges in tiny object detection. First, most object detectors use downsampling for feature extraction, which will result in a large amount of information loss for tiny objects. Second, due to the limited amount of valid information they

TABLE I

COMPARISON BETWEEN EXISTING LABEL ASSIGNMENT METHODS FOR TINY OBJECTS AND OUR METHOD

Method	Formula	Insensitive	Comprehensive	Adaptive
DotD	$DotD = e^{-\frac{D}{S}}$, $D = \sqrt{(x_g - x_a)^2 + (y_g - y_a)^2}$, $S = \sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} w_{ij} \times h_{ij}}{\sum_{i=1}^{M} N_i}}$	✓	✗	✗
NWD	$NWD = e^{-\frac{W}{C}}$, $W = \sqrt{(x_g - x_a)^2 + (y_g - y_a)^2 + \left((w_g - w_a)^2 + (h_g - h_a)^2\right) \times 0.25}$	✓	✓	✗
RFLA	$RFD = \frac{1}{1+RFDC}$, $RFDC = \frac{0.5\beta \times w_a^2}{w_g^2} + \frac{0.5\beta \times h_a^2}{h_g^2} + \frac{2(x_a - x_g)^2}{w_g^2} + \frac{2(y_a - y_g)^2}{h_g^2} + \ln \frac{w_g}{\beta \times w_a} + \ln \frac{h_g}{\beta \times h_a} - 1$	✓	✓	✗
SimD	$SimD = e^{-\left(sim_{location} + sim_{shape}\right)}$	✓	✓	✓

contain, small objects are easily disturbed by noise. Finally, the smaller an object is, the more sensitive it is to changes in the bounding box [2]. Consequently, if we use traditional label assignment metrics, such as the IoU, GIoU [12], DIoU [13], and CIoU [13], for object detection, the number of positive samples obtained for tiny objects will be very small.

Many methods and have been proposed to improve the accuracy and efficiency of tiny object detection. For example, from the perspective of data augmentation, Kisantal et al. [14] proposed increasing the number of training samples by copying tiny objects, randomly transforming the copies and then pasting the results into new positions in an image.

B. Label Assignment Strategies

The label assignment strategy plays a significant role in object detection. Depending on whether each label is either strictly negative or strictly positive, such strategies can be divided into hard label assignment strategies and soft label assignment strategies. In a soft label assignment strategy, different weights are set for different samples based on the calculation results, examples include GFL [15], VFL [16], TOOD [17] and DW [18]. Hard label assignment strategies can be further divided into static and dynamic strategies depending on whether the thresholds for designating positive and negative samples are fixed. Static label assignment strategies include those based on the IoU and DotD [2] metrics as well as RFLA [4]. Examples of dynamic label assignment strategies include ATSS [19], PAA [20], OTA [21], and DSLA [22]. From another perspective, label assignment strategies can be divided into prediction-based and prediction-free strategies. A prediction-based method assigns a sample a positive/negative label based on the relationship between the ground-truth and predicted bounding boxes, while a prediction-free method assigns labels based only on the anchors or other existing information.

C. Label Assignment Strategies for Tiny Objects

Although there have been many studies on label assignment strategies for object detection, most such strategies are designed for conventional datasets, with few specifically designed for tiny objects. When these traditional label assignment strategies are directly used for tiny object detection, they suffer a significant decrease in accuracy. To date, the label assignment strategies and metrics designed specifically for tiny objects mainly include S^3FD [23], DotD [2], NWD-RKA [24] and RFLA [4].

In S^3FD, the threshold value is first reduced (from 0.5 to 0.35) to obtain more positive samples for ground truth and then is further lowered to 0.1 to obtain positive samples for those ground truths that were not addressed with the first threshold reduction. However, S^3FD also uses the traditional IoU metric to compute the similarity between the ground truth and anchor. To overcome the weakness of the IoU metric, the novel DotD formula was introduced to reduce the sensitivity to the bounding box size. Based on this metric, more positive samples can be obtained for the ground truth. In NWD-RKA, a normalized Wasserstein distance is introduced as a replacement for the IoU, and a ranking-based strategy is used to assign the top-k samples as positive. RFLA explores the relationship between the ground truth and anchor from the perspective of the receptive field, on this basis, the ground truth and anchor are modeled as Gaussian distributions. Then, the distance between these two Gaussian distributions is calculated based on the Kullback–Leibler divergence (KLD), which is used in place of the IoU metric.

III. METHOD

A. Similarity Distance Between Bounding Boxes

One of the most important steps in label assignment is to calculate a value that reflects the similarity between different bounding boxes. Specifically, for an anchor-based label assignment strategy, the similarity between the anchors and ground truths must be quantified before assigning labels.

Common label assignment metrics, such as the IoU, GIoU [12], DIoU [13] and CIoU [13], are usually based on the overlap between the anchor and the ground truth. These metrics have the serious problem that if the overlap is zero, which is often the case for tiny objects, these metrics may become invalid. Some more suitable methods use distance-based evaluation metrics or even use Gaussian distributions to model the ground truth and anchor, such as the DotD [2], NWD [3] and RFLA [4]. We present a simple comparison between the existing metrics and our SimD metric from three perspectives in Table I. For example, DotD only considers location similarity and may not adapt different object sizes in a dataset, so it is not comprehensive or adaptive. NWD and RFLA are not adaptive because they respectively have

a hyperparameter C and β need to set. Following the existing approaches, we consider proposing an adaptive method without any hyperparameters.

In this paper, we introduce a novel metric named Similarity Distance (SimD) to better reflect the similarity between different bounding boxes. The Similarity Distance is defined as follows:

$$SimD = e^{-\left(sim_{location}+sim_{shape}\right)} \tag{1}$$

$$sim_{location} = \sqrt{\left(\frac{x_g - x_a}{\frac{1}{m} \times (w_g + w_a)}\right)^2 + \left(\frac{y_g - y_a}{\frac{1}{n} \times (h_g + h_a)}\right)^2} \tag{2}$$

$$sim_{shape} = \sqrt{\left(\frac{w_g - w_a}{\frac{1}{m} \times (w_g + w_a)}\right)^2 + \left(\frac{h_g - h_a}{\frac{1}{n} \times (h_g + h_a)}\right)^2} \tag{3}$$

where m and n are shown below:

$$m = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{Q_i} \frac{|x_{ij} - x_{ik}|}{w_{ij} + w_{ik}}}{\sum_{i=1}^{M} N_i \times Q_i} \tag{4}$$

$$n = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{Q_i} \frac{|y_{ij} - y_{ik}|}{h_{ij} + h_{ik}}}{\sum_{i=1}^{M} N_i \times Q_i} \tag{5}$$

The SimD contains two parts, location similarity, $sim_{location}$ and shape similarity, sim_{shape}. As shown in (2), (x_g, y_g) and (x_a, y_a) respectively represent the center coordinate of ground truth and anchor, w_g, w_a, h_g, h_a represent the width and height of ground truth and anchor, the main part of the formula is the distance between the center point of ground truth and anchor, similar to the DotD formula [2]. The difference is that we use the widths and heights of the two bounding boxes multiplied by corresponding parameters to eliminate the differences between bounding boxes of different sizes. This process is similar to the idea of normalization. This is also why our metric can easily adapt to different object sizes in a dataset. The shape similarity in (3) is similar to (2).

The definitions of the two normalization parameters are shown in (4) and (5). Because they are quite similar, we take the parameter m in (4) as an example for further discussion. m is the average ratio of the distance in the x-direction to the sum of the two widths for all ground truths and anchors in each image in the whole train set. M represents the number of images in the train set, N_i and Q_i represent the number of ground truths and anchors, respectively, in the i-th image. x_{ij} and x_{ik} respectively represent the x-coordinate of the center point of j-th ground truth and k-th anchor in i-th image. w_{ij} and w_{ik} respectively represent the width of j-th ground truth and k-th anchor in i-th image. Because the two normalization parameters are computed based on the train set, our metric can also automatically adapt to different datasets.

To facilitate label assignment, we use an exponential function to scale the value of SimD to the range between zero and one. If the two bounding boxes are the same, the value below the root sign will be zero, so SimD will be equal

to one. If the two bounding boxes greatly differ, this value will be very large, so SimD will approach zero.

B. Similarity Distance-based Detectors

The novel SimD metric defined in (1) can well reflect the relationship between two bounding boxes and is easy to compute. Therefore, it can be used in place of the IoU in scenarios where the similarity between two bounding boxes needs to be calculated.

SimD-based label assignment. In traditional object detectors, such as Faster R-CNN [8], Cascade R-CNN [25] and DetectoRS [26], the label assignment strategy for the RPN and R-CNN models is MaxIoUAssigner. MaxIoUAssigner considers three threshold values: a positive threshold, a negative threshold and a minimum positive threshold. Anchors for which the IoU with respect to the ground truth is higher than the positive threshold are positive samples, those for which the IoU is lower than the negative threshold are negative samples, and those for which the IoU is between the positive threshold and the negative threshold are ignored. For tiny object detection, Xu et al. introduced the RKA [24] and HLA [4] label assignment strategies, which do not use fixed thresholds to divide positive and negative samples. In RKA, the top-k anchors associated with a ground truth are simply selected as positive samples, this strategy can increase the number of positive samples because the assignment of positive labels is not limited by a positive threshold. However, introducing too many low-quality positive samples may cause the detection accuracy to degrade.

In this paper, we follow the traditional MaxIoUAssigner strategy and simply use SimD in place of the IoU. The positive threshold, negative threshold, and minimum positive threshold are set to 0.7, 0.3, and 0.3, respectively. Our label assignment strategy is named MaxSimDAssigner.

SimD-based NMS. Non Maximum Suppression (NMS) is one of the most important components of postprocessing. Its purpose is to eliminate predicted bounding boxes that are repeatedly detected by preserving only the best detection result. In the traditional NMS procedure, first, the IoUs between the bounding box with the highest score and all other bounding boxes are computed. Then, bounding boxes with an IoU higher than a certain threshold will be eliminated. Considering the advantages of SimD, we can simply use it as the metric for NMS in place of the traditional IoU metric.

IV. EXPERIMENTS

To verify the reliability of our proposed method, we designed a series of experiments involving the application of traditional object detectors to several open-source tiny object detection datasets.

A. Datasets

There are two main types of tiny object detection datasets: one contains only small objects, such as AI-TOD [27], AI-TODv2 [24], and SODA-D [1], the other contains both small and medium-sized objects, such as VisDrone2019 [28] and TinyPerson [29].

Method	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_{vt}	AP_t	AP_s	AP_m
TridentNet [33]	ResNet-50	7.5	20.9	3.6	1.0	5.8	12.6	14.0
Faster R-CNN [8]	ResNet-50	11.1	26.3	7.6	0.0	7.2	23.3	33.6
Cascade R-CNN [25]	ResNet-50	13.8	30.8	10.5	0.0	10.6	25.5	36.6
DetectoRS [26]	ResNet-50	14.8	32.8	11.4	0.0	10.8	18.3	38.0
DotD [2]	ResNet-50	16.1	39.2	10.6	8.3	17.6	18.1	22.1
DetectoRS w/ NWD [3]	ResNet-50	20.8	49.3	14.3	6.4	19.7	29.6	**38.3**
DetectoRS w/ RFLA [4]	ResNet-50	24.8	55.2	18.5	9.3	24.8	30.3	38.2
SSD-512 [10]	ResNet-50	7.0	21.7	2.8	1.0	4.7	11.5	13.5
RetinaNet [34]	ResNet-50	8.7	22.3	4.8	2.4	8.9	12.2	16.0
ATSS [19]	ResNet-50	12.8	30.6	8.5	1.9	11.6	19.5	29.2
RepPoints [35]	ResNet-50	9.2	23.6	5.3	2.5	9.2	12.9	14.4
AutoAssign [36]	ResNet-50	12.2	32.0	6.8	3.4	13.7	16.0	19.1
FCOS [37]	ResNet-50	12.6	30.4	8.1	2.3	12.2	17.2	25.0
M-CenterNet [27]	DLA-34	14.5	40.7	6.4	6.1	15.0	19.4	20.4
Faster R-CNN w/ SimD	ResNet-50	23.9	54.5	17.6	11.9	24.8	28.0	33.5
Cascade R-CNN w/ SimD	ResNet-50	25.0	53.2	19.1	13.2	25.8	29.0	35.7
DetectoRS w/ SimD	ResNet-50	**26.6**	**55.9**	**21.2**	**13.4**	**27.5**	**30.9**	37.8

AI-TOD. AI-TOD (Tiny Object Detection in Aerial Images) is an aerial remote sensing small object detection dataset collected to address the shortage of available datasets for aerial image object detection tasks. It contains 28,036 images and 700,621 object instances divided into eight categories with accurate annotations. Due to the extremely small size of its object instances (the mean size is only 12.8 pixels), it can be effectively used to test the capabilities of tiny object detectors.

SODA-D. The SODA (Small Object Detection dAtasets) series includes two datasets: SODA-A and SODA-D. SODA-D was collected from MVD [30], which consists of images captured from streets, highways and other similar scenes. There are 25,834 extremely small objects (with areas ranging from 0 to 144) in SODA-D [1], making it an excellent benchmark for tiny object detection tasks.

VisDrone2019. VisDrone2019 is a dataset from the VisDrone Object Detection in Images Challenge. For this competition, 10,209 static images were captured by an unmanned drone in different locations at different heights and angles. VisDrone2019 is also an excellent dataset for evaluating tiny object detectors because it contains not only extremely small objects but also normally sized objects.

B. Experimental settings

In the following series of experiments, we use a computer with one NVIDIA RTX A6000 GPU and implement various models based on the object detection framework MMDetection [31] and PyTorch [32]. We use general object detectors such as Faster R-CNN, Cascade R-CNN and DetectoRS as the base models and simply replace the MaxIoUAssigner module with our SimD assignment module. Our method can effectively adaptive to any backbone networks and anchor-based detectors. Following the main stream settings, for all of the models, ResNet-50-FPN pretrained on ImageNet is used

as the backbone, and stochastic gradient descent (SGD) is used as the optimizer, with a momentum of 0.9 and a weight decay of 0.0001. The batch size is set to 2, and the initial learning rate is 0.005. The number of RPN proposals is 3000 in both the training and testing stages. For the VisDrone2019 dataset, the number of epochs for training is set to 12, and the learning rate decays at the 8th and 11th epochs. For AI-TOD, AI-TODv2 and SODA-D, the number of training epochs is 24, and the learning rate decays in the 20th and 23rd epochs. For NMS, we use the IoU metric, and the IoU threshold is set to 0.7 for RPN and 0.5 for R-CNN. Other aspects of the configuration, such as the data preprocessing and pipeline, follow the defaults in MMDetection.

To facilitate comparison with previous research results, during the testing stage, we use the AI-TOD benchmark evaluation metrics, which comprise the Average Precision (AP), $AP_{0.5}$, $AP_{0.75}$, AP_{vt}, AP_t, AP_s and AP_m, for AI-TOD, AI-TODv2 and VisDrone2019. For the SODA-D dataset, we use the COCO evaluation metrics.

C. Results

We design four groups of experiments on the AI-TOD, AI-TODv2, VisDrone2019 and SODA-D datasets. In each group, we replace the IoU metric with our SimD metric in the RPN module and then apply this module in combination with traditional object detection models, including Faster R-CNN, Cascade R-CNN and DetectoRS.

The results on AI-TOD are shown in Table II, where we compare our method to several typical object detection methods. The detectors in the first seven rows are two-stage anchor-based detectors, those in the next three and four rows are one-stage anchor-based and anchor-free detectors, respectively, and the last three rows show the results of our method. Compared with Faster R-CNN, Cascade R-CNN and DetectoRS, we achieve AP improvements of 12.8, 11.2 and

TABLE III

MAIN RESULTS ON AI-TODV2. ALL OF THE SETTINGS ARE AS SAME AS THOSE FOR AI-TOD

Method	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_{vt}	AP_t	AP_s	AP_m
TridentNet [33]	ResNet-50	10.1	24.5	6.7	0.1	6.3	19.8	31.9
Faster R-CNN [8]	ResNet-50-FPN	12.8	29.9	9.4	0.0	9.2	24.6	37.0
Cascade R-CNN [25]	ResNet-50-FPN	15.1	34.2	11.2	0.1	11.5	26.7	38.5
DetectoRS [26]	ResNet-50-FPN	16.1	35.5	12.5	0.1	12.6	28.3	**40.0**
DetectoRS w/ NWD-RKA [24]	ResNet-50-FPN	24.7	57.4	17.1	9.7	24.2	29.8	39.3
YOLOv3 [38]	DarkNet-53	4.1	14.6	0.9	1.1	4.8	7.7	8.0
RetinaNet [34]	ResNet-50-FPN	8.9	24.2	4.6	2.7	8.4	13.1	20.4
SSD-512 [10]	VGG-16	10.7	32.5	4.0	2.0	8.7	16.8	28.0
RepPoints [35]	ResNet-50-FPN	9.3	23.6	5.4	2.8	10.0	12.3	18.9
FoveaBox [39]	ResNet-50-FPN	11.3	28.1	7.4	1.4	8.6	17.8	32.2
FCOS [37]	ResNet-50-FPN	12.0	30.2	7.3	2.2	11.1	16.6	26.9
Grid R-CNN [40]	ResNet-50-FPN	14.3	31.1	11.0	0.1	11.0	25.7	36.7
Faster R-CNN w/ SimD	ResNet-50-FPN	24.5	56.3	17.2	10.3	24.9	28.9	35.4
Cascade R-CNN w/ SimD	ResNet-50-FPN	25.2	55.2	19.1	9.7	25.3	30.3	37.4
DetectoRS w/ SimD	ResNet-50-FPN	**26.5**	**57.7**	**20.5**	**11.6**	**26.7**	**30.9**	38.4

TABLE IV

MAIN RESULTS ON VISDRONE2019. ALL MODELS ARE TRAINED ON THE TRAIN SET AND TESTED ON THE VAL SET

Method	AP	$AP_{0.5}$	AP_{vt}	AP_t
Faster R-CNN	22.3	38.0	0.1	6.2
Cascade R-CNN	22.5	38.5	0.5	6.8
DetectoRS	25.7	41.7	0.5	7.6
DetectoRS w/ NWD-RKA	27.4	46.2	4.4	12.6
DetectoRS w/ RFLA	27.4	45.3	4.5	12.9
Faster R-CNN w/ SimD	26.5	47.9	7.5	16.3
Cascade R-CNN w/ SimD	27.7	48.6	6.8	16.1
DetectoRS w/ SimD	**28.7**	**50.3**	**7.6**	**17.3**

TABLE V

MAIN RESULTS ON SODA-D. ALL MODELS ARE TRAINED ON THE TRAIN SET AND TESTED ON THE TEST SET

Method	AP	$AP_{0.5}$	$AP_{0.75}$
Faster R-CNN [8]	28.9	59.7	24.2
Cascade R-CNN [25]	31.2	59.9	27.8
RetinaNet [34]	28.2	57.6	23.7
FCOS [37]	23.9	49.5	19.9
RepPoints [35]	28.0	55.6	24.7
ATSS [19]	26.8	55.6	22.1
YOLOX [41]	26.7	53.4	23.0
RFLA [4]	29.7	**60.2**	25.2
Faster R-CNN w/ SimD	31.1	59.8	28.1
Cascade R-CNN w/ SimD	32.4	59.2	30.5
DetectoRS w/ SimD	**32.8**	59.4	**31.3**

TABLE VI

ABLATION STUDY ON AI-TOD BASED ON FASTER R-CNN

norm_width	norm_height	AP	$AP_{0.5}$	$AP_{0.75}$
		20.4	49.6	13.2
✓		22.1	52.0	14.8
	✓	22.0	52.3	14.5
✓	✓	**23.9**	**54.5**	**17.6**

11.8 points, respectively, by using SimD in place of the IoU in RPN. We also compare our method with some specialized detectors for tiny objects, namely, DotD, NWD and RFLA, relative to which our method improves the AP by 10.5, 5.8 and 1.8 points, respectively.

The performance of our method on tiny objects is worthy of special attention. Because of the extremely small size of these objects (very tiny refers to a size range from 2 to 8 pixels), the AP_{vt} of general object detectors is 0, whereas with the use of SimD, the AP_{vt} values of Faster R-CNN, Cascade R-CNN and DetectoRS are increased from 0 to 11.9, 13.2 and 13.4 points, respectively.

In addition to AI-TOD, our method also achieves the best performance on AI-TODv2, VisDrone2019 and SODA-D, as shown in Table III, Table IV and Table V, respectively. On AI-TODv2 and SODA-D, the AP of our method is 1.8 and 1.6 points higher, respectively, than that of its best competitor. On VisDrone2019, which contains both tiny and general-sized objects, our method also performs well, in particular, it achieves a 1.3 points improvement over RFLA. In Table V, the $AP_{0.5}$ is almost at the same level with RFLA but $AP_{0.75}$ is much higher, this may indicate that our method is more capable of tiny object detection. Some typical visual comparisons between the IoU metric and SimD are shown

in Fig. 3. We can find there is an obvious improvement of detection performance after using our method.

D. Ablation Study

In our proposed method, an important operation is normalization based on the width and height of the ground truth and anchor. To verify the effectiveness of the normalization operation, we conduct a set of ablation study. As shown in Table VI, we respectively compare not normalizing, normalizing only width, only height, and both width and height. The experimental results show that normalization operation achieves a 3.5 points improvement, mainly benefit by the ability to adapt to objects of different sizes in a dataset, and the normalization parameters m, n can be adaptively adjusted according to different datasets.
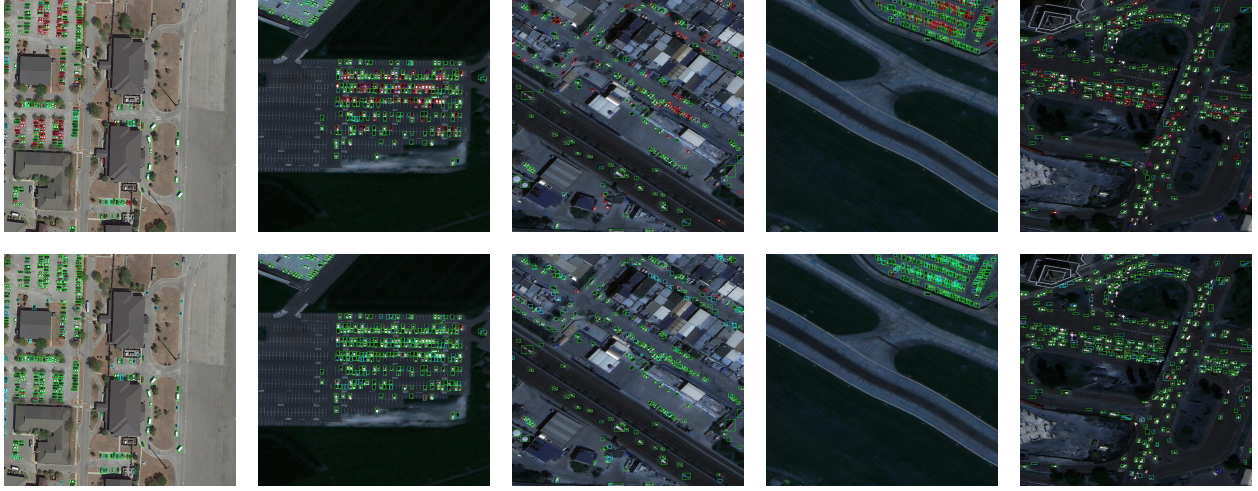
Fig. 3. Comparison of detection results on AI-TOD dataset between label assignment with the traditional IoU metric and SimD. The first row shows the results of Faster R-CNN with the IoU metric, and the second row is also based on Faster R-CNN but with the SimD metric. The green, blue and red boxes respectively denote true positive (TP), false positive (FP) and false negative (FN) predictions. The improvement achieved with our method is obvious.

E. Analysis

From the experimental results shown in Table II to Table V, we find that our method achieves the best AP on all four datasets. In addition, on the AI-TOD, AI-TODv2, and VisDrone2019 datasets, our method achieves the best results for very tiny, tiny and small objects. Three main achievements of our method can be summarized.

First, our method effectively solves the problem of low accuracy for tiny objects. The most fundamental reason is that our method fully accounts for the similarity between two bounding boxes, including both location and shape similarity, therefore, only the highest-quality anchors will be chosen as positive samples when using the SimD metric. Compared with VisDrone2019, the performance improvements on AI-TOD and AI-TODv2 are more obvious because the objects are much smaller in these two datasets, this phenomenon may also reflects the effectiveness of our method on tiny object detection.

Second, our method can adapt well to different object sizes in a dataset. In Table IV, both the AP and AP_{vt} values of our method are the best and much higher than those of other methods. The main reason is that normalization is applied in the SimD metric when calculating the similarity between bounding boxes, so it can eliminate the differences arising from bounding boxes of different sizes. Some typical detection results are shown in Fig. 4.

Finally, our method achieves the state-of-the-art results on four different datasets. Although the characteristics of the objects in different datasets vary, we use the relationships between the ground truths and anchors in the train set when computing the normalization parameters, enabling our metric to automatically adapt to different datasets. In addition, there are no hyperparameters that need to be set in our formula.

V. CONCLUSION

In this paper, we point out that most of the existing methods may not be able to automatically adapt to objects

Fig. 4. Some typical detection results on VisDrone2019 val set, which contains both tiny and general-sized objects.

of different sizes and include some hyperparameters need to be chosen. To this end, we propose a novel evaluation metric named Similarity Distance (SimD), which not only considers both location and shape similarity but also can automatically adapt to different datasets and different object sizes in a dataset. In addition, there are no hyperparameters in our formula. Finally, we conduct extensive experiments on four classic tiny object detection datasets, in which our method achieves the state-of-the-art results. Although our proposed SimD metric is adaptive, it also based on the existing label assignment strategy with fixed thresholds. In the future, we aim to further improve the effectiveness of label assignment for tiny object detection.

REFERENCES

[1] G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, X. Xie, and J. Han, "Towards Large-Scale Small Object Detection: Survey and Benchmarks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13467-13488, 2022.

[2] C. Xu, J. Wang, W. Yang, and L. Yu, "Dot Distance for Tiny Object Detection in Aerial Images," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1192-1201.

[3] J. Wang, C. Xu, W. Yang, and L. Yu, "A Normalized Gaussian Wasserstein Distance for Tiny Object Detection," in *Computer Vision and Pattern Recognition*, 2021.

[4] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "RFLA: Gaussian Receptive Field based Label Assignment for Tiny Object Detection," in *European Conference on Computer Vision*, 2022, pp. 526-543.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580-587.

[7] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440-1448.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779-788.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. Berg, "SSD: Single Shot MultiBox Detector," in *Computer Vision–ECCV 2016: 14th European Conference*, 2016, pp. 21–37.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, "Microsoft COCO: Common Objects in Context," in *Computer Vision–ECCV 2014: 13th European Conference*, 2014, pp. 740–755.

[12] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 658-666.

[13] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12993–13000, 2020.

[14] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," in *9th International Conference on Advances in Computing and Information Technology (ACITY 2019)*, 2019.

[15] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection," in *Advances in Neural Information Processing Systems*, 2020, pp. 21002-21012.

[16] H. Zhang, Y. Wang, F. Dayoub, and N. Sünderhauf, "VarifocalNet: An IoU-aware Dense Object Detector," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8510-8519.

[17] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned One-stage Object Detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3490-3499.

[18] S. Li, C. He, R. Li, and L. Zhang, "A Dual Weighting Label Assignment Scheme for Object Detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9377-9386.

[19] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9756-9765.

[20] K. Kim and H. S. Lee, "Probabilistic Anchor Assignment with IoU Prediction for Object Detection," in *Computer Vision–ECCV 2020: 16th European Conference*, 2020, pp. 355–371.

[21] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, "OTA: Optimal Transport Assignment for Object Detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 303-312.

[22] H. Su, Y. He, R. Jiang, J. Zhang, W. Zou, and B. Fan, "DSLA: Dynamic smooth label assignment for efficient anchor-free object detection," *Pattern Recognition*, vol. 131, pp. 108868, 2022.

[23] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and SZ. Li, "S^{3}FD: Single Shot Scale-invariant Face Detector," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 192-201.

[24] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 79–93, 2022.

[25] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154-6162.

[26] S. Qiao, L.-C. Chen, and A. Yuille, "DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10208-10219.

[27] J. Wang, W. Yang, H. Guo, R. Zhang, and G.-S. Xia, "Tiny Object Detection in Aerial Images," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 3791-3798.

[28] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and Tracking Meet Drones Challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380-7399, 2022.

[29] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han, "Scale Match for Tiny Person Detection," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1246-1254.

[30] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5000-5009.

[31] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, and C. C. Loy, "MMDetection: Open MMLab Detection Toolbox and Benchmark," in *Computer Vision and Pattern Recognition*, 2019.

[32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, vol. 32, 2019.

[33] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, "Scale-Aware Trident Networks for Object Detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6053-6062.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980-2988.

[35] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point Set Representation for Object Detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9656-9665.

[36] B. Zhu, J. Wang, Z. Jiang, F. Zong, S. Liu, Z. Li, and J. Sun, "AutoAssign: Differentiable Label Assignment for Dense Object Detection," in *Computer Vision and Pattern Recognition*, 2020.

[37] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9626-9635.

[38] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," in *Computer Vision and Pattern Recognition*, 2018.

[39] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyound Anchor-Based Object Detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 7389-7398, 2020.

[40] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid R-CNN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7363-7372.

[41] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," in *Computer Vision and Pattern Recognition*, 2021.