IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Scale-Sensitive IOU Loss: an improved Regression Loss Function in Remote Sensing Object Detection

SHUANGJIANG DU[1], BAOFU ZHANG[2], AND PIN ZHANG.[3]
[1]College of Communication Engineering, Army Engineering University of PLA, Nanjing, 210072, China (e-mail: shuangjiangdu@163.com)
[2]College of Communication Engineering, Army Engineering University of PLA, Nanjing, 210072, China (e-mail: zhangbaofu@163.com)
[3]College of Field Engineering, Army Engineering University of PLA, Nanjing 210007, China

Corresponding author: Pin Zhang (e-mail: djcorp@yeah.net).

This work was supported in part by the National Nature Science Foundation of China under Grant 61371121

ABSTRACT Regression loss function in object detection model plays a important factor during training procedure. The IoU based loss functions, such as CIOU loss, achieve remarkable performance, but still have some inherent shortages that may cause slow convergence speed. The paper proposes a Scale-Sensitive IOU(SIOU) loss for the object detection in multi-scale targets, especially the remote sensing images to solve the problem where the gradients of current loss functions tend to be smooth and cannot distinguish some special bounding boxes during training procedure in multi-scale object detection, which may cause unreasonable loss value calculation and impact the convergence speed. A new geometric factor affecting the loss value calculation, namely area difference, is introduced to extend the existing three factors in CIOU loss; By introducing an area regulatory factor γ to the loss function, it could adjust the loss values of the bounding boxes and distinguish different boxes quantitatively. Furthermore, we also apply our SIOU loss to the oriented bounding box detection and get better optimization. Through extensive experiments, the detection accuracies of YOLOv4, Faster R-CNN and SSD with SIOU loss improve much more than the previous loss functions on two horizontal bounding box datasets, i.e, NWPU VHR-10 and DIOR, and on the oriented bounding box dataset, DOTA, which are all remote sensing datasets. Therefore, the proposed loss function has the state-of-the-art performance on multi-scale object detection.

INDEX TERMS scale sensitivity, regression loss function, area difference, object detection

I. INTRODUCTION

REGRESSION loss function is a significant factor that affects the object detection performance, the ℓ_n norm is first used to calculate regression loss of which Smooth L1-norm [14] is an improvement.

The IoU based loss loss functions are also the widely used regression loss functions in many object detection models, of which the first proposed is IOU loss [15] and it performs better than the former in many datasets. Nevertheless, IOU loss has some inherent disadvantages especially when the bounding box do not overlap with the ground truth box, that is, IoU values is 0, and the Generalized IOU (GIOU) loss [16] improves the IOU loss. Distance IOU(DIOU) loss and Complete IOU(CIOU) loss [17] are proposed, arguing that the former two loss functions still have some lacks in theory. In CIOU loss, it summarizes three geometric factors that affect the regression loss value calculation, namely overlap area, center point distance and aspect ratio. DIOU loss and CIOU loss further accelerate the optimization speed of the bounding box and the precision of the model. Furthermore, Efficient IOU loss [18] combines the theory of Focus Loss [19] and add hard example mining mechanism into CIOU loss, which improves the performance of the later one.

CIOU loss takes into account three geometric factors when calculating the regression loss. But in multi-scale detection, the areas of ground truth boxes in the same image change greatly, thus, there are more non-negligible cases as shown in Figure. 1. As for the same ground truth box, there are many cases where the two different bounding boxes of different areas meeting the same conditions as follows:

- The IoU values between the two bounding boxes and the same ground truth box equal with each other, that is, $IoU_1 \approx IoU_2$.
- The ratios between diagonal length of the union box L

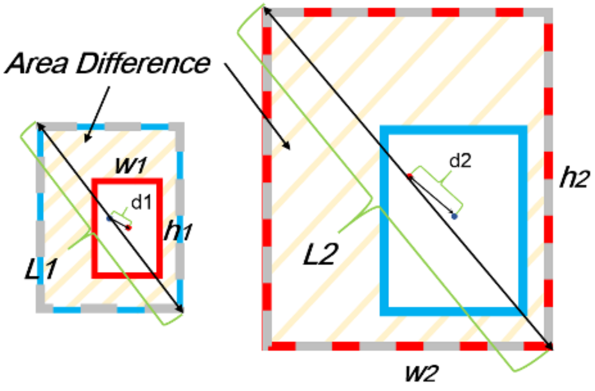


FIGURE 1. The special relations between the bounding box and the ground truth box. The two loss values equal with each other. The Blue box is the ground truth box, the Red box is the bounding box, the Grey box is the union box of them and the Light Orange shadow refers to the area difference.

and the center point distance d equal with each other, that is, $d_1/L_1 \approx d_2/L_2$.

- The aspect ratios of the two bounding boxes equal with each other, that is, $w_1/h_1 \approx w_2/h_2$.

Generally believing that the difference between the two values is approximately equal when it is less than 1e-3. If the bounding boxes meet the cases as mentioned above, then, it is impossible for the current CIOU loss to differentiate them. This problem is particularly prominent when the ground truth box areas vary greatly in one image. At the same time it is sensible to mark these bounding boxes with different loss values. At present, the several mainstream regression loss functions just take into account three geometric factors ,i.e, overlap area, center point distance and aspect ratio to calculate the regression loss. But, through the above analysis, it can be found that not all bounding boxes could be exactly differentiated if just using these three factors. Last but not the least, if the area of the bounding box is much bigger than the ground truth target, the gradients at these points of the loss function become smooth, which may slow down the optimization(in Section IV-B).

To end these problems, our paper proposes the Scale-Sensitive IOU(SIOU) loss, taking into account another geometric factor, namely, area difference, when calculating the regression loss function, as shown in Figure 1. We add an area adjustment factor γ to the CIOU loss to keep the loss values of the bounding boxes of different area different and also raise the gradient around the maximum and minimum loss points, thus, the loss function could differentiate all these bounding boxes theoretically and speed up the optimization procedure.

To thoroughly verify the superior of the proposed method, the paper chooses the most advanced detector of one-stage and two-stage, YOLOv4, SSD and Faster R-CNN to launch comparison experiments, modify the loss functions of them and puts the SIOU loss on them. Selects two mainstream aerial remote sensing datasets, DIOR [20] and NWPU VHR-10 [21], of which the target area scales vary greatly, as the training and testing sets. Meanwhile, we also use SIOU loss to do the oriented bounding box object detection, we replace the ArIoU loss in DRBox [38] with our SIOU loss during training, and the detection accuracy also improves a lot.

The main contributions of our paper are as follows:

1) Propose the Scale-Sensitive(SIOU) loss to improve the CIOU loss, which could differentiate all the bounding boxes in theory and speed up the optimization procedure.
2) Introduce another geometric factor namely area difference when calculating the regression loss values and make the calculation more reasonable.
3) Improve the detection accuracy of multi-scale object detection in both traditional bounding box and the oriented bounding box, which illustrate a broad applicability.

II. RELATED WORKS

A. OBJECT DETECTION

Object detection plays an importance role in many subject field. It could be classified into two-stage and one-stage detections. Two stage detection models, like R-CNN series [1]–[4] and FPN [5] achieve great performance in many datasets. One-stage detection models, like SSD [11], YOLO series [6]–[9], are the most classic models. RefineDet [12] and Retina Net [19] are also widely used. Guo et al [22] used a center-point rectangle loss function(CR loss) in Faster R-CNN to detect the droppers in high-speed railway. It takes the center points of bounding box and ground truth box as the vertex of the rectangle. The rectangle penalty term could quickly move the bounding box close to the ground truth box. But, it is similar to DIOU loss and it is a bit more complex to calculate center-point rectangle than the center point distance. Chen et al [23] combined the GIOU loss and soft-NMS in Faster R-CNN to detect the ships of SAR images. To deal with the imbalance issues in training procedure, Focus loss [19] firstly took hard negative mining mechanism into one-stage detection model; Libra R-CNN [24] proposed a balanced L1 loss to solve the imbalance issues in three aspects; Dynamic R-CNN [33] uses a changeable β values of Smooth L1 loss to dynamically focus on hard samples; DR loss [25] introduced distribution ranking mechanism to choose the hard candidates; Others like RefineDet++ [34], Guided Anchoring [26] and FCOS model [30] are also some effective methods. In order to save the human labor for dataset annotation, Li et al. [35] proposed a weakly supervised deep learning (WSDL) method for remote sensing object detection without costly bounding box annotation. It used class-specific activation maps(CAM) segmentation and a multi-scale scene-sliding-voting strategy to detect the multi-scale targets; To mitigate the impact of error labels in remote sensing scene classification, RSSC-ETDL [36] proposed an error-tolerant method and used the adaptive multi-feature collaborative representation classifier to correct the error labels.

B. REGRESSION LOSS FUNCTION

ℓ_n norm like Smooth ℓ_1 loss [14], mean average error (MAE) are widely used in many deep learning models. They are easy to calculate the loss values. But they also have some inherent lacks, for example, they cannot combine the parameters of the bounding box together, thus it may not get a better optimization result in theory. The IoU series loss functions, like IOU loss [15], GIOU loss [16], DIOU loss [17] and CIOU loss are also some very popular regression loss functions. Many SOTA models([23], AS-YOLO [27], [28]) use these loss functions for detection tasks; IOU loss takes the Intersection over Union between the bounding box and ground truth box as the loss function; GIOU loss adds another area item on the basis of IOU loss. The following DIOU and CIOU loss add extra center point distance and aspect ratio items to make the loss calculation more proper and speed the optimization procedure; Others like Efficient IOU [18] and LIOU [29] point out the convergence speed issue of CIOU loss, and use different method to improve the CIOU loss. Wang et al. [37] revised the $\alpha\nu$ item of CIOU loss in YOLOv4 model, changing the *arctan(h_{gt}/w_{gt})-arctan(h/w)* into *arctan(h/h_{gt})+arctan(w/w_{gt})* to avoid the degradation of CIOU loss when the aspect ratios are the same.

III. PROPOSED METHOD

This section systematically expounds the differences of several existing loss functions, quantitatively compare their characteristics and introduce our SIOU loss.

A. SCALE-SENSITIVE IOU LOSS

The first regression loss function is ℓ_n-norm loss, in which Smooth L1-norm is often used for regression loss calculation, and its formula is as follows:

$$\mathrm{L(x)} = \begin{cases} 0.5|\mathrm{x}|^2 & ,\mathrm{if}\ |\mathrm{x}| < 1 \\ |\mathrm{x}|-0.5 & ,\mathrm{otherwise} \end{cases} \tag{1}$$

where $|x|$ means difference value between the bounding box parameters (*x*, *y*, *w*, *h*) and ground truth box parameters (x^{gt}, y^{gt}, w^{gt}, h^{gt}). ℓ_n-norm loss is an effective loss function in optimization, and different *n* values has different characteristics, used in different deep learning tasks.

The following loss functions are based on IoU. These functions have a common equation as shown below:

$$L(B, B^{gt}) = 1 - \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} + \Re(B, B^{gt}) \tag{2}$$

where

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{3}$$

B means the bounding box parameters while B^{gt} means those of every target box corresponding. For different IoU based loss function, the formula of $\Re(B, B^{gt})$ is variable.

As for GIOU loss:

$$\Re(B, B^{gt}) = \frac{|C - B \cup B^{gt}|}{|C|} \tag{4}$$

where C means the smallest box covering B and B^{gt} at the same time.

As for DIOU loss:

$$\Re(B, B^{gt}) = \frac{\rho^2(B, B^{gt})}{c^2} = \frac{|Center(B) - Center(B^{gt})|^2}{W^2 + H^2} \tag{5}$$

$Center(\cdot)$ means the center point of the box, W, H means the width and height of the box C, and c^2 is the diagonal length of it.

As for CIOU loss:

$$\Re(B, B^{gt}) = \frac{\rho^2(B, B^{gt})}{c^2} + \alpha * \nu \tag{6}$$

where:

$$\nu = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{7}$$

$$\alpha = \frac{\nu}{(1 - IoU + \nu)} \tag{8}$$

According to the mainstream view, calculating the regression loss mainly takes into account three geometric factors, that is, the overlap area, center point distance and aspect ratios between the bounding box and the target box. Among these loss functions, the IOU loss considers the overlap area of the two, while GIOU loss solves this problem from the complementary area; DIOU loss takes into account the center point distance and the CIOU loss adds the aspect ratio into consideration. In this way, the model with CIOU loss has a faster converging speed and a higher detection precision than that with many other loss functions based on bounding box regression in theory.

But when the areas of ground truth boxes in one image vary greatly, there will be some special cases between bounding boxes and ground truth boxes as shown in Figure 2, in which each pair of bounding boxes meet the following conditions:

- $$\frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \approx \frac{|B' \cap B^{gt}|}{|B' \cup B^{gt}|}$$
- $$\frac{\rho^2(B, B^{gt})}{c^2} \approx \frac{\rho^2(B', B^{gt})}{c'^2}$$
- $$\arctan\frac{w}{h} \approx \arctan\frac{w'}{h'}$$

In one hand, as for the two bounding boxes in Figure 2 (a), the *IoU*=0.75, the area difference between the the left bounding box and the target box is $0.25S^{gt}$, while that between the right bounding box and the target box is $0.33S^{gt}$, then, we do not think there are too much difference between the two bounding boxes in scales. But it is obvious that the right one has more information of the target, so it is reasonable to believe it is better than the left one. On the other hand, as for the two bounding boxes in Figure 2 (b), the *IoU*=0.45, the area difference between the the left bounding box and the target box is $0.55S^{gt}$, while that between the right bounding box and the target box is $1.2S^{gt}$, thus, the area of the right

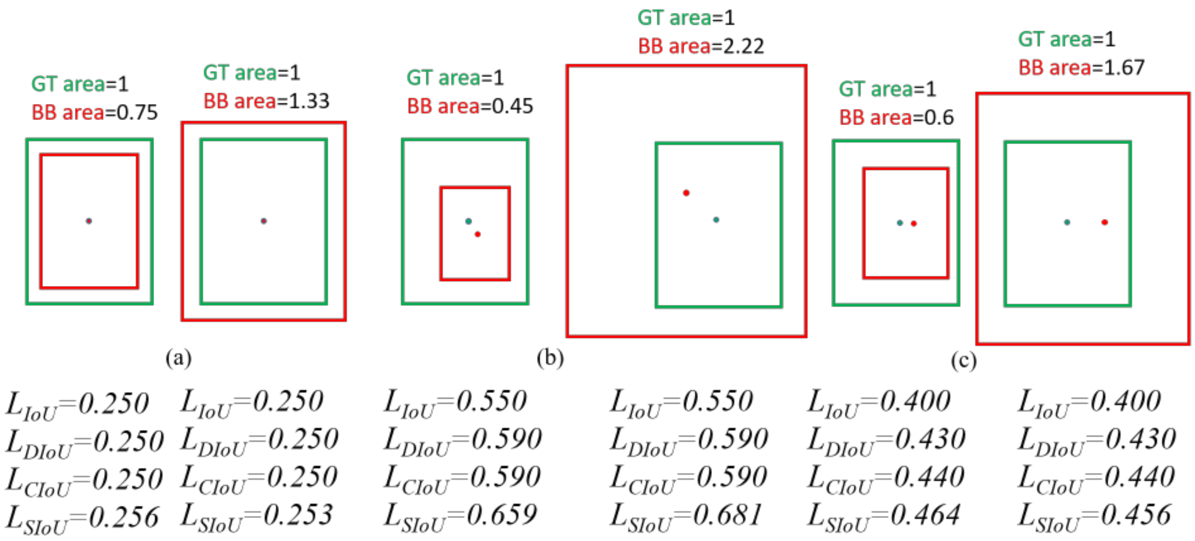

GT area=1
BB area=0.75
GT area=1
BB area=1.33
GT area=1
BB area=0.45
GT area=1
BB area=2.22
GT area=1
BB area=0.6
GT area=1
BB area=1.67
(a)
(b)
(c)
L_{IoU}=0.250 L_{IoU}=0.250 L_{IoU}=0.550 L_{IoU}=0.550 L_{IoU}=0.400 L_{IoU}=0.400
L_{DIoU}=0.250 L_{DIoU}=0.250 L_{DIoU}=0.590 L_{DIoU}=0.590 L_{DIoU}=0.430 L_{DIoU}=0.430
L_{CIoU}=0.250 L_{CIoU}=0.250 L_{CIoU}=0.590 L_{CIoU}=0.590 L_{CIoU}=0.440 L_{CIoU}=0.440
L_{SIoU}=0.256 L_{SIoU}=0.253 L_{SIoU}=0.659 L_{SIoU}=0.681 L_{SIoU}=0.464 L_{SIoU}=0.456

FIGURE 2. The losses of the two bounding boxes are the same when calculated by the previous loss functions in the three cases above, so they can not be differentiated by the previous loss functions. However the SIOU loss can make it. The Green boxes are the ground truth boxes(GT), while the Red boxes are the bounding boxes(BB).

bounding box is much larger than that of the left one, and it is not sure whether the right one contains only one target or much useless even interference information. So it is believe logically and intuitively that the left bounding box is more proper although its area is small than the target box and does not contain all the information of the target. However, the regression loss values of the bounding boxes calculated by the existing loss functions are the same in the above three cases, so it is theoretically impossible to distinguish them. In this way, the area difference between the bounding box and the target will become an important factor affecting the calculation of the regression loss.

To solve the above problems, this paper proposes SIOU loss as follows:

$$\begin{aligned} SIOU\ loss &= (\gamma+1)(1-IoU)+\frac{\rho^2(B,B^{gt})}{c^2}+\alpha*\nu \\ &= 1-IoU+\frac{\rho^2(B,B^{gt})}{c^2}+\alpha*\nu+\gamma*(1-IoU) \\ &= CIOU\ loss+\gamma*(1-IoU) \end{aligned} \tag{9}$$

$$\gamma=\begin{cases} [\tanh(k*AD-2.3)+\tanh(2.3)]/2 & ,IoU>0 \\ 0 & ,IoU=0 \end{cases} \tag{10}$$

$$k=\begin{cases} k_0\ ,AD\geq 0 \\ -2k_0\ ,AD<0 \end{cases} \tag{11}$$

According to formula (10), SIOU loss adds a new item, γ, to CIOU loss and proposes another geometric factor, i.e., area difference(AD), while the CIOU loss just takes three factors into consideration. As shown in formula (12), area difference is different with *IoU* especially when the area differences between bounding box and ground truth box are with different signs.

$$\begin{aligned} AD &= (s-s_{gt})/s_{gt} \\ &= s/s_{gt}-1 \\ &= \begin{cases} IoU-1, & \mathrm{s} < \mathrm{s_{gt}} \\ 1/IoU-1, & \mathrm{s} > \mathrm{s_{gt}} \end{cases} \end{aligned} \tag{12}$$

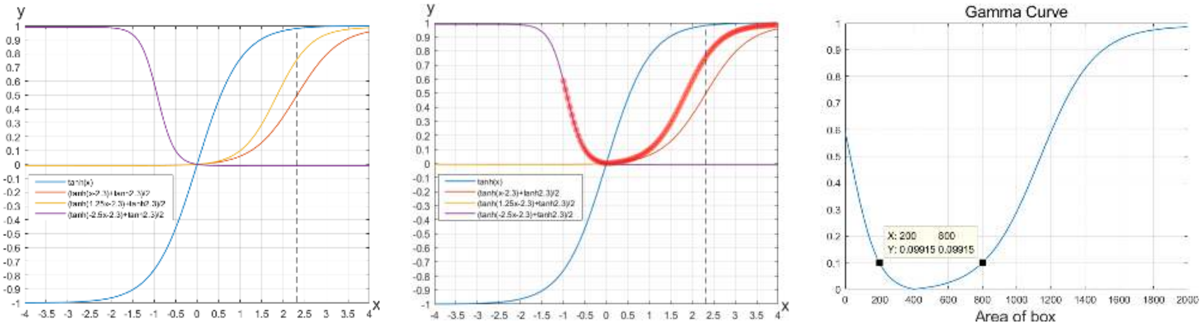
For two bounding boxes in the cases mentioned above, even if the *IoU* values equal with each other, the area differences are not the same.

What it differs from CIOU loss is that the former adds a scale regulating term, γ*(1-*IoU*). Since the purpose of SIOU loss is to adjust for differences in the calculation of loss values caused by changes in area difference, its expression form must be area dependent.

Therefore, in order to make the expression form simple and clear and convenient to calculate, the parameter γ is directly used as a regulation coefficient and multiplied by the (1-*IoU*) term when constructing the SIOU function. In this way, the physical meaning of the original expression is retained, and the role of proper fine-tuning can be really played.

B. METHOD ANALYSIS

The function of γ is to adjust the loss value of bounding boxes in different area scales, so it must be sensitive to the area variation. When the area increases from S_{gt} to $2S_{gt}$, the difference between the areas of the bounding box and the ground truth box is not large, and the γ curve should increase slowly. When the area increases from $2S_{gt}$ to $4S_{gt}$, the difference between the two areas is large, and the γ curve should increase rapidly to adjust the influence of the area. When the area continues to increase, the curve should



(a) (b) (c)

FIGURE 3. γ curve. From (a) to (c) is the procedure of choosing the γ curve. (a) and (b) are the comparison graph of different expressions which are all based on the tanh function. And (c) is the final function curve, which is a piecewise function.

flatten again to avoid the problem of explosion of loss value. Secondly, as a regulating parameter, γ should be valued at [0, 1], so as not to affect the value of the original loss function.

Based on this, the hyperbolic tangent function tanh(x) function is adopted as the basic function, as shown in Figure 3 (a). However, only the middle part meets the requirements, so we need to carry out appropriate transformation of this function to extract the middle part.

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{13}$$

$$\tanh\prime(x) = 1 - \tanh^{2}(x) \tag{14}$$

According to the above formula, tanh(2.3)=0.98 approximately close to 1 in the range of [0, 1], tanh'(2.3)=0.04, at which point, its gradient is relatively gentle, while the gradient at the largest point is tanh'(0)=1 $\gg$ 0.04, so it is also close to zero compared to the largest point. In this way shift the tanh(x) by 2.3 to the right, and by 2.3 to the up, and then in the first quadrant of the axis, it's going to be pretty good. Meanwhile, near the origin, the curve gradient is about zero, the curve tends to be smooth, and the values on both sides will not have a large mutation, which is conducive to the iterative optimization of parameters.

For the negative half of the X-axis, it can be obtained by flipping it symmetrically along the Y-axis directly. But if just flipping it, it can't really tell the difference between the positive area difference and the negative area difference. In order to adjust the loss value problem of large-scale regression box and balance the relationship between the loss value of small-scale and large-scale regression box, We add two different coefficients k to the tanh(x) function in the case of (x <0) and (x >0) to balance the loss values under these two opposite cases. In formula (10) and (12). *IoU*=1-AD, γ=F(k'·AD) when s<s_{gt}; *IoU*=1/(AD+1), γ=F(k·AD) when s>s_{gt}. Now, if we make the *IoU* and γ keep the same with themselves under two the cases, through calculation, we find when s/s_{gt}=k'/k_0, the above two variables are the same with themselves under the two cases, that is to say, if s/s_{gt}=k'/k_0, then, the loss values of the pair bounding boxes are the same. In this way, when we construct the formula of SIOU loss, we make this assumption that when the areas of the pair bounding boxes are 1/2 and 2 times of the ground truth box, we think the regression loss values of the pair bounding boxes are equal, thus k'=2k_0. The γ is to balance the loss values of the pair bounding boxes in the two different cases, When (x >0), we choose k_0 from 0.5 to 2 in arithmetic sequence with increment of 0.25, and k'=-2k_0, when x <0. Then use these serial SIOU loss with different k_0 values to launch the simulation experiments. After constant adjustment and comparison, we choose the k_0 values with the best simulation result. The final decision was made that when x >0, k=1.25, and when x <0, k=2.5. Combine the curves in the positive and negative field of the X-axis together to form the part of the curve marked in red in Figure 3 (b). Figure 3 (c) is the γ curve when the area of the ground truth box is 400. When the area of the bounding box are 200 and 800, that is, 1/2 and 2 times the area of the target box, the γ values are equal. When the area continues to increase, the γ value increases rapidly. When the area increases to 4 times the area of the target box, the growth rate of the γ value slows down and approaches to 1. It can be seen from the image that the state of γ curve change can basically meet the preset requirements of the problem.

C. FUNCTIONS OF SIOU LOSS

The SIOU loss function is compared with the other four *IoU*-based loss functions:

1) γ in SIOU loss, is related to the area difference. SIOU loss can well solve the overlap area, center point distance and the aspect ratio in the regression loss. At the same time, it introduces and solves the problem of area difference, thus makes it more reasonable to calculate the regression loss and differentiate all the bounding boxes in the optimization process, thus making the optimization result more accurate for the multi-scale target boxes in a comprehensive way.
2) When the bounding box and the target box perfectly match, L_{IOU}=L_{GIOU}=L_{DIOU}= L_{CIOU}=L_{SIOU}=0. When bounding box does not overlap with ground truth box, γ = 0, and SIOU loss changes into CIOU

loss, because when the two boxes do not overlap, the area scale difference problem is meaningless and what plays a leading role in the optimization process is $|Center(B) - Center(B^{gt})|$, therefore, the influence of γ on the loss value should be reduced.

3) For the item γ, we know it also ranges in [0, 1], This is the same as the variation range of $\alpha * \nu$ in CIOU loss, but the have different influence stages. For SIOU loss, when the regression loss is large at the beginning of training, it has a main impact. When the loss value decreases, it means that the bounding box and the ground truth box are similar with each other, then, the γ item becomes small, and the $\alpha * \nu$ item starts to play a main role and to adjust the aspect ratio of the bounding box.

4) From the definition of SIOU, it can be concluded that the loss function has a good optimization effect in the object detection under variable scales, and its optimization effect is similar to that of CIOU in theory when the target scale is similar and single in an image.

IV. SIMULATION ANALYSIS

In this section, we used simulate experiments to analyze bounding box regression procedure of five *IoU* based loss functions, i.e., IOU loss, GIOU loss, DIOU loss, CIOU loss and SIOU loss. The algorithm is designed to simulate the optimization process of the bounding box regression, the loss values of the bounding boxes and the target boxes are calculated to visually compare the converging speed of each loss function in the optimization process and quantitatively compare the qualities of the final optimization results. Meanwhile use 3D graphs to compare the values and their gradients of the loss functions at different points.

A. SIMULATION EXPERIMENT

This simulation experiment refers to Zheng et al [17]. The algorithm in detail is shown in Algorithm 1. Some parameters were changed in this experiment considering that this experiment is to simulate the optimization of regression box under multiple scales. As shown in Figure 3, the areas of anchor boxes vary dramatically, the largest area of anchor box is set as 4, while the smallest area is set as 1/4, by which method the optimization ability of SIOU loss can be tested. In Figure 4 (a), there randomly scattered 1,000 points on a circular area with a radius of 3 and a center of (10, 10). The point (10,10) contains three ground truth boxes with an area of 1 and aspect ratios of 1/2, 1, 2. Each scattered point contains 5×6 anchor boxes with areas of 1/4, 1/2, 1, 2, 3, 4 and aspect ratios of 1/3, 1/2, 1, 2, 3 respectively. Therefore, there are a total of 90,000=3×5×6×1,000 regression boxes per iteration.

The loss functions in the iteration process are the five regression loss functions compared above. The final evaluation index error E adopts the ℓ_1 norm, namely $\left|B^t_{n,s,i} - B^{gt}_i\right|$. Figure 4 (b) is the simulation results after fixed iteration of these loss functions. It could be seen from the figure that the IoU Loss was indeed inferior to the other four loss functions in the optimization process. This is caused by the inherent shortcomings of IOU loss, because when bounding box and target box do not overlap, *IoU* =0, resulting in the gradient of the target function remaining zero, which cannot be further optimized. Meanwhile, the optimization results of DIOU loss and CIOU loss were better than that of GIOU loss, which was consistent with the results of literature Zheng et al. More importantly, SIOU loss got the best performance among all these loss functions in speed and result of the optimization.

Algorithm 1 Bounding Box Regression

```
Input: T =200 means the iterations.N=1000 uniformly scattered points within the circular region with center (10, 10) and radius 3. S=6×5 including 6 scales and 5aspect ratios of anchor boxes of each scattered point. $B^{gt}$ means the target boxes fixed at point (10, 10) with area 1 and 3aspect ratios. $B_(n, s, i)^t$ means the predicted bounding box of point N=n, S=s to target box i at iteration T=t. Loss function L($B_(n, s, i)^t$, $B^g_i t$) calculate the loss between the target boxes and the predicted boxes to optimization
Output: Total regression error $E$
Initialize $E$ = 0
Start
for $t$=1 to $T$ do
    for $i$=1 to 3 do
        for $n$=1 to $N$ do
            for $s$=1 to $S$ do
                $\eta = \begin{cases} 0.1 & t \le 0.8T \\ 0.01 & 0.8T < t \le 0.9T \\ 0.001 & t > 0.9T \end{cases}$
                $\nabla B^{t-1}_{n,s,i} = \partial L(B^{t-1}_{n,s,i}, B^{gt}_i)/\partial B^{t-1}_{n,s,i}$
                $B^t_{n,s,i} = B^{t-1}_{n,s,i} + \eta(2 - IoU^{t-1}_{n,s,i})\nabla B^{t-1}_{n,s,i}$
                $E(t) = E(t) + \left|B^t_{n,s,i} - B^{gt}_i\right|$
            end for
        end for
    end for
end for
return $E$
End
```

B. VISUALIZATION OF THE LOSS FUNCTIONS

We drew the visualization simulation graphs to intuitively compare the difference among these loss functions as shown in Figure 5. In subfigure (a), set ground truth box with height of 60, width of 80 and center point(40, 30). Then we changed bounding boxes with different widths and heights from 1 to 160 and from 1 to 120 uniformly, thus, get 160 × 120 =19,200 bounding boxes with uniform scale variation. We use the above five loss functions to calculate the loss values between the bounding box and the ground truth box in Figure 5 (b) $\sim$ (f).

Particularly, Figure 6 shows the gradient variance of the CIOU loss and SIOU loss when the width and height of the bounding box equal.

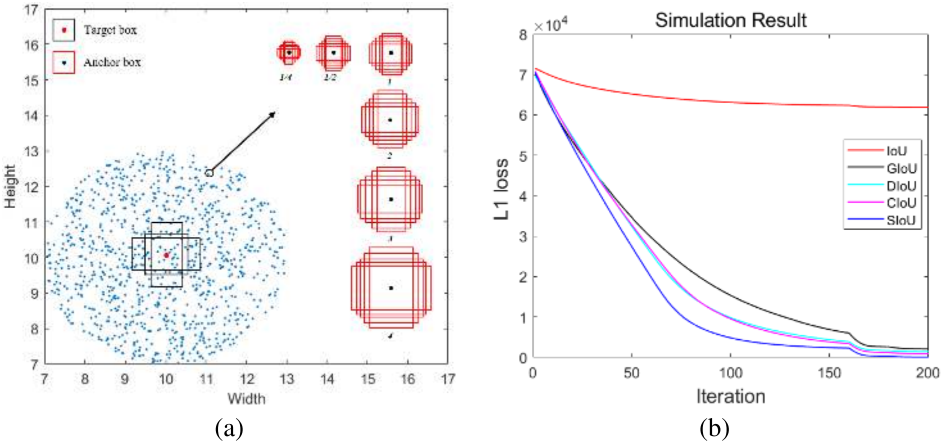


FIGURE 4. (a) is sketch map of the scattered points and boxes. The **Black** boxes are the ground truth boxes, Red boxes are the anchor boxes. (b) is the simulation results.

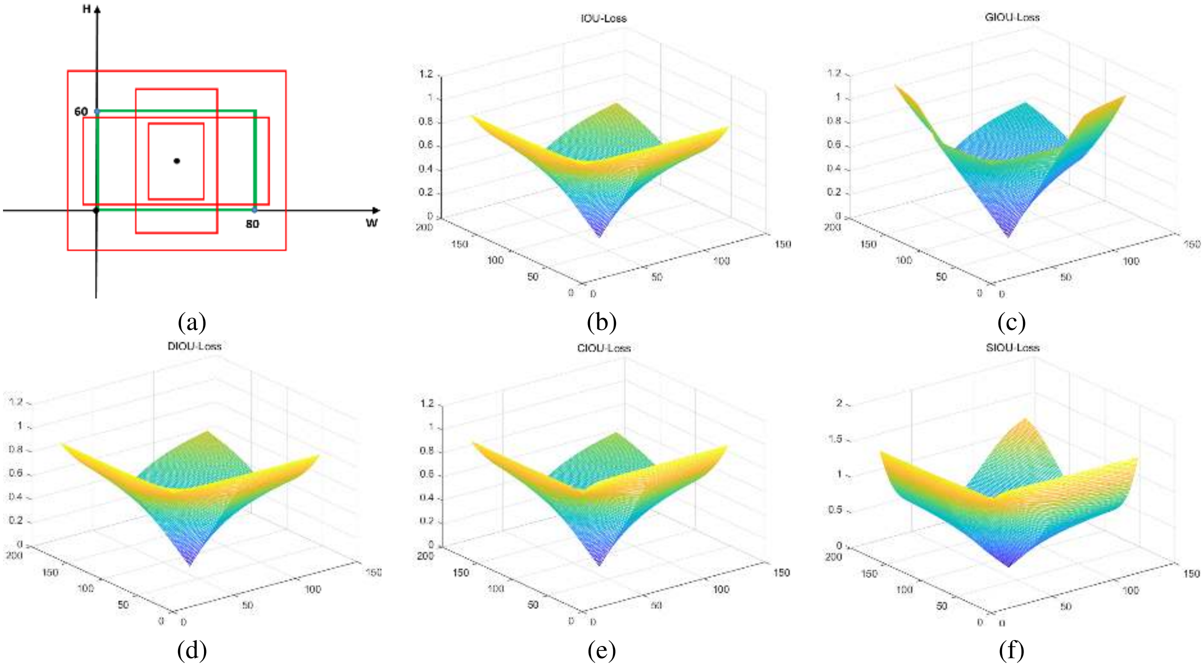


FIGURE 5. The visualization of the loss values drawn by the five loss functions: (a) is the schematic diagram of target boxes and bounding boxes, (b)∼(f) are the visualization graphs of the losses calculated by the five loss functions above.

From the visualization graphs we could intuitively draw the conclusions as follows:

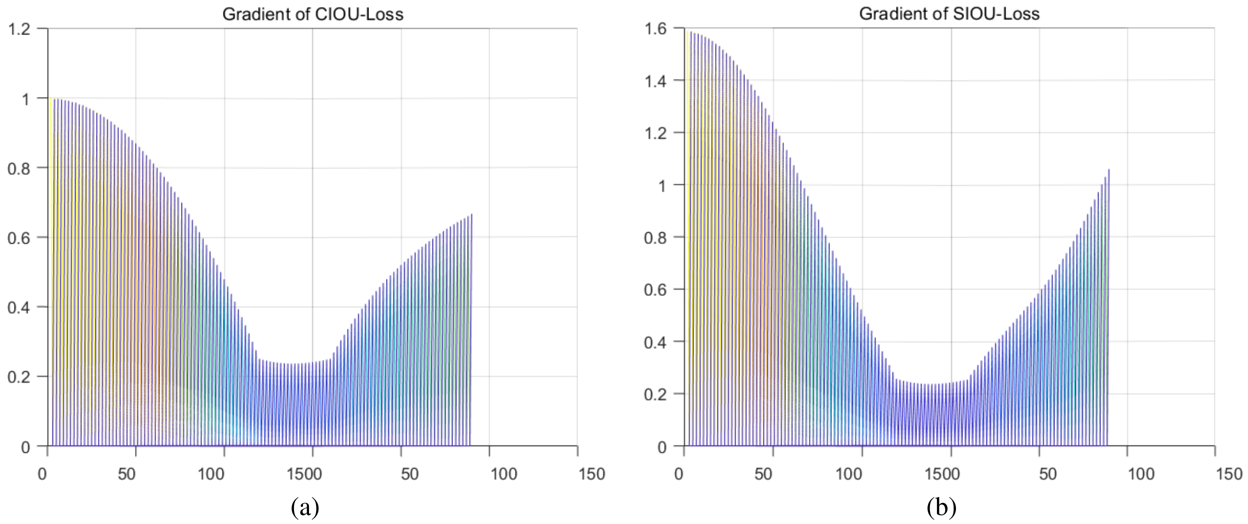
1) For these five graphs in Figure 5, the area of bounding box varies in $[0, 4S_{gt}]$, which has a large area variation range, When the width of the bounding box is 80 and the height is 60, that is, they match perfectly, and the loss value is zero.
2) As for the previous four loss functions,the loss value rapidly increases when the bounding box area is much smaller than the ground truth box. Nevertheless, when it is at the maximum point, $4S_{gt}$, the loss value remains at the largest values and do not change significantly. At the same time, the gradients around the largest and smallest loss values point tend to be flat, which may slow down optimization procedure.
3) Compared with CIOU loss in Figure 6, when the bounding box area changes greatly, the value of SIOU loss also changes rapidly. Influenced by the area adjustment factor γ, the gradient of the SIOU loss is steeper, too, which could promote the optimization process.

Through simulation comparison and visualization analysis, the superiority of the proposed SIOU loss is verified.

V. EXPERIMENT RESULTS

A. EXPERIMENT DATASET

In this section, several data sets are used for experimental verification of SIOU loss. We select IOU loss, DIOU loss, CIOU loss, and another SOTA method—ICIOU loss [37]



(a)

(b)

FIGURE 6. The gradient variance of the CIOU loss(a) and SIOU loss(b) when the width and height of the bounding box equal with each other.

as comparison. The selected datasets are NWPU VHR-10, DIOR and UCAS-AOD, all of which are mainstream aerial remote sensing datasets.

DIOR is a remote sensing dataset used for object detection. The dataset contains 23,463 images and 192,472 objects including 20 object classes. NWPU VHR-10 is a 10-level geographic remote sensing dataset for the detection of space objects, with 650 images containing the targets in 10 categories and 150 background images. UCAS-AOD includes 1,000 aircraft images and 510 vehicle images, of which the objects in the dataset are step-by-step uniform and consistent in scale.

Sample images of the three datasets are shown in Figure 7. The three datasets all have the characteristics of low resolution and relative high density of targets, among which, NWPU VHR-10 and DIOR datasets have great differences in the area scale of the targets in one image.

This experiments do not choose MS COCO and PASCAL VOC datasets because first of all, this study was aimed at detecting aerial remote sensing targets. Secondly, through analysis, it was found that target boxes in one image in the two datasets above do not meet the requirements of target density and large scale difference, of which the targets tend to be conventional objects, such as faces, pedestrians, furniture and animals. The image resolution is high and the features are obvious. In order to prove the rationality of the datasets selected in this experiments, we use Coefficient of Variance(CV) [35] analysis to quantitatively compare the image differences of the four datasets of DIOR, NWPU VHR-10, UCAS-AOD and Pascal VOC-07. Coefficient of Variance is a statistic that reflects the fluctuation of several sets of data. The formula is as follows:

$$V_s = \frac{\sigma}{X} \tag{15}$$

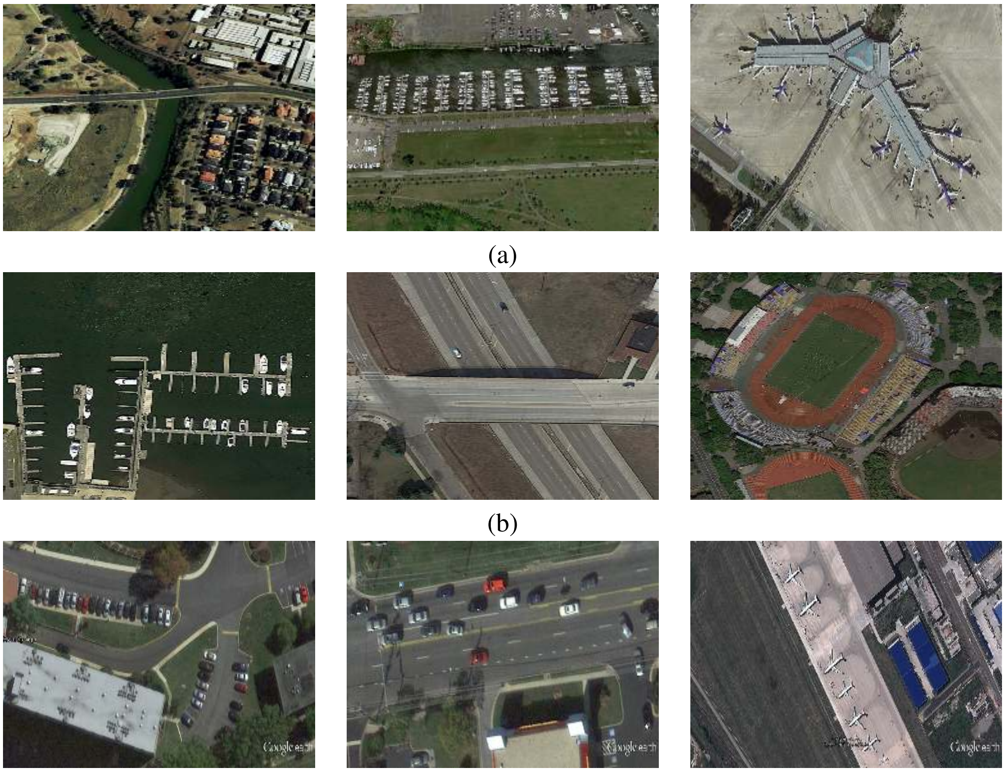
V_s means the sample standard deviation, X means the sample mean, generally the bigger V_s, the higher the fluctuation of the samples are. When $V_s > 1$, generally believe samples fluctuate greatly, when $V_s > 1$, believe less fluctuation. The Coefficient of Variance of the target area scale of each image in each dataset are calculated separately, and set the average of the Coefficient of Variance of all images in one dataset as the overall variance degree of it. As shown in Table 1:

It can be seen that the fluctuation of DIOR and NWPU VHR-10 is greater than that of PASCAL VOC-07, and the target scale in UCAS-AOD is the stablest and smallest. Figure 8 is the statistical histogram of the target scale dispersion coefficient of each image in the datasets of DIOR, PASCAL VOC-07 and UCAS-AOD. It can be seen from the figure that the DIOR dataset contains more images.

B. YOLOV4 ON NWPU VHR-10 AND DIOR

The YOLOV4 model has a high detection accuracy in the MS COCO dataset and is the most representative model in the YOLO series. The feature extraction network uses CSPDarknnet-53; The Necknet adopts SPP module to integrate candidate box feature vectors of different sizes into the same dimension; The PAN module fuses feature images of three different scales by up sampling and down sampling. The Head part use the classification network of Yolov3, and the prediction results of the three scales were output simultaneously. In addition to the innovative model structure, Yolov4 also uses some excellent Bag of Freebie (BOF) and Bag of Special (BOS) training strategies and techniques, such as using CIOU loss as its regression loss function; The feature extraction network uses Mosaic data augmentation to augment the training data; Cosine annealing scheduler [9] is used in the learning rate during the training, making the learning rate update more reasonable. The above features make YOLOV4 have a high detection accuracy not only for large scale objects, but also much higher than other models for small scale objects.

During experiment, YOLOV4 model was first used for


(a)

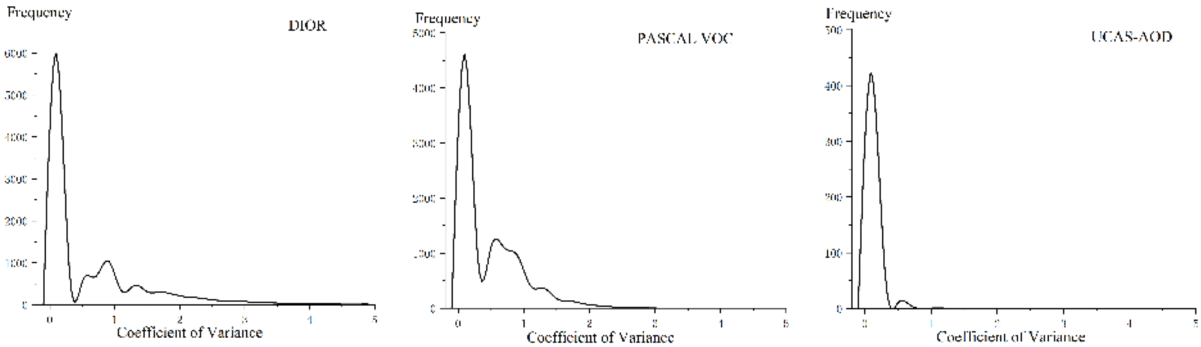
(b)

(c)

FIGURE 7. The examples of three different datasets. (a) is NWPU VHR-10 dataset, (b) is DIOR dataset and (c) is UCAS-AOD dataset.

TABLE 1. Mean Coefficient of Variance of Four Datasets.

Dataset	DIOR	NWPU VHR-10	PASCAL VOC-07	UCAS-AOD
Image num	11,726	500	9,964	510
Target num	70,359	3,243	29,896	4,591
V_s	0.6357	0.6443	0.4418	0.1604



(a) (b) (c)

FIGURE 8. The examples of three different datasets. (a) is NWPU VHR-10 dataset, (b) is DIOR dataset and (c) is UCAS-AOD dataset.

training and testing on NWPU VR-10. NWPU VR-10 has a total of 650 images, 500 pieces were randomly selected as the training set and 150 pieces as the testing set. Set batch size =8, weight decay=1e-5 during training. The backbone network adopts the MS COCO pre-trained weights. Use Cosine annealing scheduler. Firstly, freeze the feature extraction network and train for 30 epoches, the initial learning rate is 1e-3, then unfreeze the feature extraction network for another 30 epoches of training, the initial learning rate is 1e-4. This can speed up the optimization of model parameters. The test results were evaluated by mean average precision (mAP), threshold=0.5.

Secondly, the model is used for training and testing on the DIOR dataset. The DIOR dataset contains 23,643 images. 60% of them are randomly selected as the training and validation set, and the remaining 40% as the testing set. Since there are 20 categories and more than 10K images in DIOR training dataset, the training epoch of the two stages before and after unfreezing is set to be 60 respectively. The initial learning rate and other parameters were consistent with the previous experiment. The regression loss function of the original algorithm is CIOU Loss, and the loss function algorithm needs to be manually modified during the experiment. Therefore, different algorithms are used to conduct five experiments for each dataset. We selected and plotted the dynamic curves of training losses and validation losses of four different loss functions in the training process on the DIOR Dataset, As is shown in Figure 9, it can be seen from the curve that the training loss of SIOU Loss decreases slightly faster in the first 20 training epochs, which may indicate that SIOU Loss plays a regulating role in the initial stage of training, because in the initial stage of training, there is a great difference in the regression box, and the scale adjustment item can help the Loss value to decrease rapidly..

The detection accuracy results of the models are shown in Table 2. Compared with IOU loss baseline, DIOU loss,CIOU loss and ICIOU loss are indeed improved, which indicates their theoretical superiority and high detection accuracy no matter in remote sensing datasets or in conventional large-scale target detection datasets such as MSCOCO tested in its original article. Meanwhile, SIOU loss has the highest detection accuracy among the other four loss functions. The detection accuracy of SIOU loss in NWPU VHR-10 reaches 88.46%, which is 1.9% higher than that of baseline. The detection accuracy on DIOR reaches 81.46%, which was 1.66% higher than that of baseline, indicating that the loss function proposed in this paper can indeed help to improve the accuracy of object detection.

At the same time, several images from IOU, CIOU and SIOU loss trained models in the DIOR dataset were selected for comparison, as shown in Figure 10. The objects in DIOR dataset image are numerous and dense, with large scale changes. It can be seen intuitively from the figure that the selection of the predicted boxes are more moderate and reasonable in the detection of SIOU under the variable scales, which includes all the information of the object as much as possible while reducing the inclusion of background information.

TABLE 2. Detection Results of YOLOv4 with Different Loss Functions on NWPU VHR-10 and DIOR.

Loss function / Dataset	NWPU VHR-10	DIOR
IOU loss	86.81	80.31
DIOU loss	87.06	80.67
Relative improve %	0.29%	0.45%
CIOU loss	87.34	81.27
Relative improve %	0.61%	1.20%
ICIOU loss	87.77	81.32
Relative improve %	1.11%	1.26%
SIOU loss	88.46	81.64
Relative improve %	1.90%	1.66%

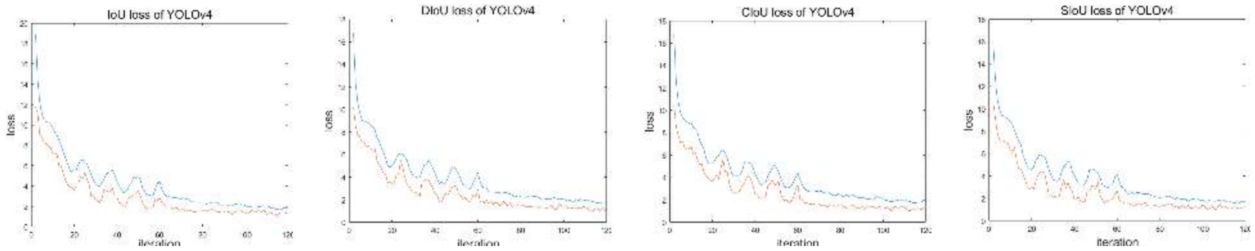
C. FASTER R-CNN ON NWPU VHR-10 AND DIOR

Faster R-CNN detection model is a classic two-stage detection model, which is gradually improved on the basis of R-CNN [1], SPP-net [2] and Fast R-CNN [3], and has good detection accuracy in many datasets. Faster R-CNN model is divided into four parts: Backbone, region proposal network(RPN), region of interest(ROI) and Classifier. Backbone can choose VGG network [13], ResNet series network [10] and so on, while the backbone network selected in this experiment is ResNet-50. RPN is similar to the Selective Search algorithm [31] to generate regional candidate boxes. ROI pooling is similar to the SPP module in YOLOV4, which is responsible for revising candidate boxes of different sizes into fixed lengths. The regression loss function of Faster R-CNN in the original paper adopts Smooth L1 function, which is an end-to-end two-stage target detection model, so the detection speed is faster and the detection accuracy is higher.

This experiment was also carried out first on NWPU VHR-10 and then on DIOR. The configuration of the dataset is consistent with the experiment in the previous section. During the training process, the pre-training weight of ResNet-50 in MS COCO is loaded on the Backbone.

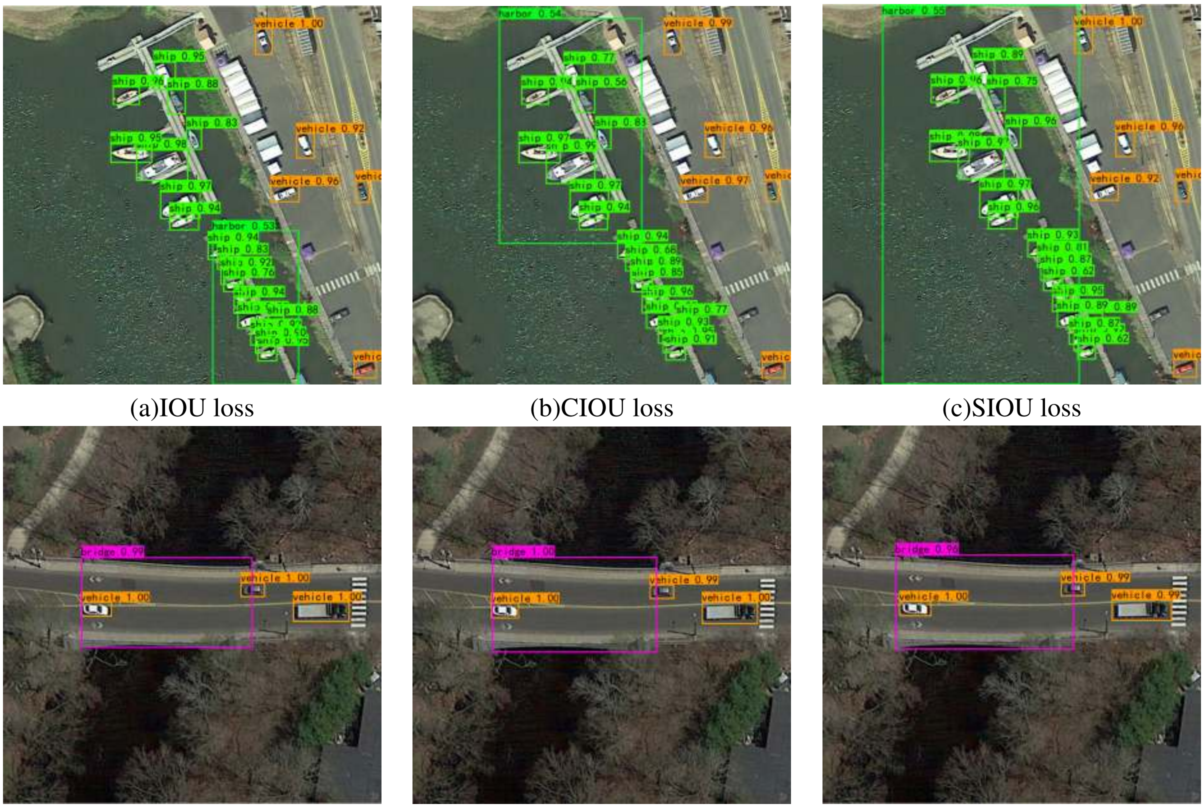
Freeze the backbone training for the first 30 epoches, then unfreeze it training for another 30 epoches, the initial learning rate was 1e-4 and 1e-5 before and after unfreezing respectively. The learning rate descended to 0.95 after each epoch, weight decay=1e-5, batch size =8 during NWPU VHR-10 training. When training on DIOR, the learning rate of the two stages before and after the unfreezing is 60, the other parameters remained unchanged.

The detection accuracy of model after training is shown in Table 3. For Faster R-CNN, the detection accuracy of SIOU loss is also higher than that of the first three loss functions while the ICIOU loss has a similar performance with SIOU loss. It is noteworthy that the accuracy of the model trained by SIOU loss function was significantly improved compared with baseline, increasing by 5.53% and 1.6% compared with CIOU loss function on NWPU VHR-10. It is also 10.2% higher than baseline and 2.5% higher than CIOU loss on DIOR. It should be noted that, as a two-stage detector, in the



(a)IOU loss (b)DIOU loss (c)CIOU loss (d)SIOU loss

FIGURE 9. Training and validation loss dynamic curves from different regression loss functions of YOLOv4 training on DIOR.



(a)IOU loss (b)CIOU loss (c)SIOU loss

(d)IOU loss (e)CIOU loss (f)SIOU loss

FIGURE 10. Comparison of detection results of YOLOv4 on DIOR.

TABLE 3. Detection Result of Faster R-CNN with Different Loss Functions on NWPU VHR-10 and DIOR.

Loss function / Dataset	NWPU VHR-10	DIOR
IOU loss	60.19	57.07
DIOU loss	60.34	57.93
Relative improve %	0.25%	1.51%
CIOU loss	62.55	61.32
Relative improve %	3.29%	7.45%
ICIOU loss	63.24	62.15
Relative improve %	5.07%	8.90%
SIOU loss	63.52	62.89
Relative improve %	5.53%	10.20%

training of Faster R-CNN, parameters are optimized in RPN as well as Classifier. Its total loss value is as follows:

$$\begin{aligned} Tatalloss = rpn_{loss}^{clc} + rpn_{loss}^{bbx} \\ + roi_{loss}^{clc} + roi_{loss}^{bbx} \end{aligned} \tag{16}$$

The first two items on the right of the equation are classification and regression loss values between the predicted output values of RPN network and the ground truth, while the last two items are loss values of the final predicted output of Classifier. When modify the regression loss function, the loss functions of RPN and Classifier are both modified, thus, both of their prediction accuracy has been improved. In the RPN stage, candidate boxes with higher accuracy can be obtained, which promotes the final prediction in the Classifier stage.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3119562, IEEE Access

Therefore, the improvement of the two stages together will lead to a significant increase in the final prediction accuracy.

Figure 11 shows some test result samples of the Faster R-CNN model trained by three different loss functions on DIOR dataset. From the images, some differences can be seen intuitively in the selection of regression boxes and the confidence of objects predicted by the three different loss function models. In general, the detection result of SIOU model is better, especially the selection of regression box is more reasonable and balanced.

D. SSD ON NWPU VHR-10 AND DIOR

SSD is another popular and classic one-stage detection model with multi-scale feature maps for detection, that is, it detects the targets in several lower and upper feature maps directly at the same time, and then use non-maximum suppression to integrate all the results for a final better one. Because of its unique multi-scale prediction, it has a relative higher precision than the early version detection models such as YOLO, Fast R-CNN as well as much faster detection speed than most of the two-stage detectors in many datasets. The regression loss function in its original paper is Smooth L1 loss. The feature extraction network usually use VGG-16, ResNet-50 and so on. SSD is the first one-stage detection model using anchor box mechanism to make regression optimization, meanwhile, to solve the problem that too many anchor boxes may contain plenty of useless background boxes and just a few target boxes thus causing the imbalance of the two kinds of boxes and wrong optimization direction, it use Hard Negative Mining mechanism [32] to select the positive boxes and negative boxes with a ratio of 1:3, which could improve training efficiency.

The configures of training on the two datasets are the same with that in YOLOv4 model. The detection accuracy results are shown in Table 4.

Compared with baseline, SIOU loss improves by 2.04% and 1.65% on NUPU VHR-10 and DIOR respectively, which are some slight improvements compared with Faster R-CNN. It has something to do with the multi-scale detection structure of SSD in theory. The selection of anchor boxes from different convolution layers makes the initial regression boxes are much more similar to the targets in area than that of other models.

TABLE 4. Detection Results of SSD with Different Loss Functions on NWPU VHR-10 and DIOR.

Loss function / Dataset	NWPU VHR-10	DIOR
IOU loss	70.11	64.28
DIOU loss	70.52	64.67
Relative improve %	0.58%	0.61%
CIOU loss	71.16	64.90
Relative improve %	1.50%	0.96%
ICIOU loss	71.64	65.32
Relative improve %	2.18%	1.62%
SIOU loss	71.54	65.34
Relative improve %	2.04%	1.65%

E. YOLOV4 ON UCAS-AOD

UCAS-AOD dataset only has two types of targets: plane and vehicle. Through dispersion coefficient analysis, the target scales of UCAS-AOD dataset are relatively consistent without great changes. In the theoretical model analysis in the previous section, it is pointed out that SIOU loss has advantages in multi-scale target detection and optimization, while for targets with little scale changes, the optimization effect does not improve very obviously. In order to verify the correctness of this theoretical analysis from the opposite side, UCAS-AOD dataset is selected, and used on YOLOV4 model to train and verify it. 500 plane images and 500 vehicle images were selected from the dataset, with a total of 1,000 images. Then, 70% are randomly selected as the train and validation set and the remaining 30% as the test set. During the training process, the parameters were set in accordance with those during the training on the NWPU VHR-10. The detection accuracy results of each model after training are shown in Table 5.

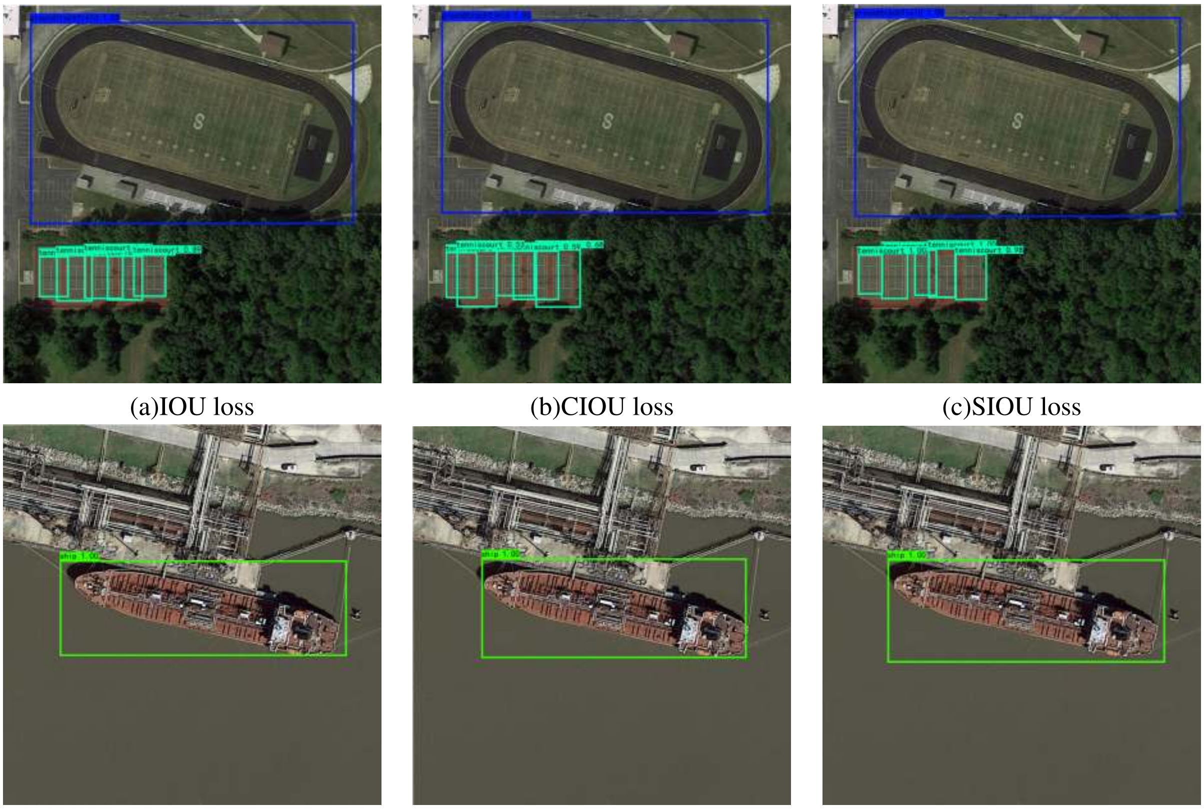
TABLE 5. Detection Results of YOLOv4 with Different Loss Functions on UCAS-AOD.

Loss function / Dataset	UCAS-AOD
IOU loss	93.24
DIOU loss	93.66
Relative improve %	0.45%
CIOU loss	94.21
Relative improve %	1.04%
ICIOU loss	95.08
Relative improve %	1.97%
SIOU loss	93.73
Relative improve %	0.53%

As can be seen from the table, the detection accuracy of the four loss function training models are all relatively high, reaching over 90%, which is related to the dataset itself. As there are only two categories and nearly 700 training images, the dataset is relatively sufficient and the training difficulty is not large. Among the five Loss functions, the detection accuracy of ICIOU loss is the highest, which is 1.97% higher than the baseline. Although the accuracy of SIOU loss is better than that of IOU Loss, it does not have the highest accuracy. After adding γ adjustment item, the accuracy of SIOU loss is slightly lower than that of CIOU and ICIOU loss. This result is within the expectation of theoretical analysis and therefore not an anomaly.

Figure 12 shows some detection results of the three loss functions. The prediction results of the three loss function models are all relatively accurate, but there are slight differences in the selection of the regression boxes.

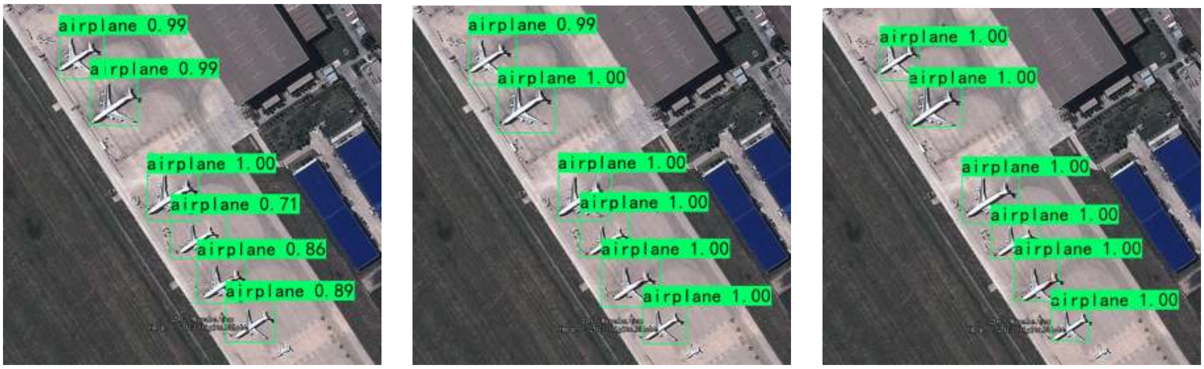
To intuitively compare the detection accuracy of the above groups of experiments, the mAP values of the four models in each group of experiments were drawn into a line chart as shown in Figure 13. The proposed SIOU loss function used on the three classic models have the highest detection accuracy on the two remote sensing dataset, meanwhile specific dataset is also used to verify the characteristics and functions



(a)IOU loss (b)CIOU loss (c)SIOU loss

(d)IOU loss (e)CIOU loss (f)SIOU loss

FIGURE 11. Comparison of detection results of Faster R-CNN on DIOR.



(a)IOU loss (b)CIOU loss (c)SIOU loss

FIGURE 12. Comparison of detection results of YOLOv4 on UCAS-AOD.

of the SIOU loss more comprehensively from the reverse side.

F. ORIENTED BOUNDING BOX DETECTION

To expand the usage of our SIOU loss, we discuss the probability of applying our loss function to oriented bounding box regression and also launched comparison experiment on oriented object detection dataset, DOTA. Compared with the traditional horizontal bounding box, the oriented bounding box has one more location parameter θ, that is, (x,y,w,h,θ). The first four parameters are the same with the traditional bounding box, while θ defines the rotation angle towards the X-axis. When doing optimization, there will also use IoU to calculate the location relationship between the bounding box and the ground truth box. Literature[38] proposed the angle-related IoU(ArIoU) to calculate the IoU values of the oriented boxes, as follows:

$$\begin{aligned} \mathrm{Ar}IoU(A,B) &= \tfrac{area(\hat{A}\cap B)}{area(\hat{A}\cup B)}\left|\cos(\theta_A-\theta_B)\right| \\ &= IoU * \left|\cos(\theta_A-\theta_B)\right| \end{aligned} \tag{17}$$

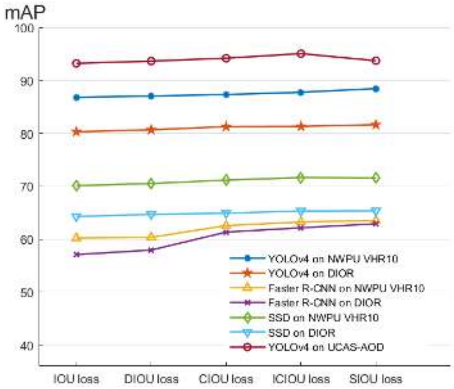


FIGURE 13. Line chart of the mAP. To illustrate the comparison of the mAP of different loss functions, draw them into one line chart.

In literature [38], the author proposed the DRBox model for oriented object detection. When doing training, the model use ArIoU as shown in formula(17), to match the bounding box with the ground truth box. In our comparison experiment, we replace the IoU item in ArIoU with our SIOU, and we select some images from the DOTA dataset for training and detection. The detection result is shown in Table 6:

TABLE 6. Detection Results of DRBox Model with Original ArIoU and SIOU on Oriented object Detection

Method	Dataset	AP(%)	mAP(%)
DRBox	Ship	93.42	92.26
	Vehicle	85.03	
	Airplane	98.34	
SIOU	Ship	94.16	94.53
	Vehicle	90.04	
	Airplane	99.38	

The detection result on oriented bounding box dataset also shows a significant improvement of our proposed method, thus, the SIOU loss could not only be used in traditional object detection, but also be used in oriented object detection.

VI. CONCLUSIONS

The proposed Scale-Sensitive IOU(SIOU) loss in our paper improved the detection accuracy of the existing loss functions. It adjusts the regression loss value calculation and accelerates the convergence speed in multi-scale datasets. Meanwhile, another geometric factor, area difference, expands the current three factors, i.e., overlap area, center point distance and aspect ratio, and could differentiate all the bounding boxes. Compared with the baseline of IOU loss on the two datasets, the detection accuracy of the YOLOV4 improves by 1.66% and 1.9%, Faster R-CNN is used to improve by 10.2% and 5.53%, meanwhile, SSD improves by 2.04% and 1.65% respectively. Furthermore, the SIOU also has promotion on oriented bounding box detection, which illustrates a wide improvement on different models and tasks.

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580-587, doi: 10.1109/CVPR.2014.81.

[2] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015, doi: 10.1109/TPAMI.2015.2389824.

[3] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.

[4] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

[5] T. Lin, P. DollÃąr, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936-944, doi: 10.1109/CVPR.2017.106.

[6] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[7] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6517-6525, doi: 10.1109/CVPR.2017.690.

[8] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[9] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[10] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[11] Liu W. et al. (2016) SSD: Single Shot MultiBox Detector. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision-ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2

[12] S. Zhang, L. Wen, X. Bian, Z. Lei and S. Z. Li, "Single-Shot Refinement Neural Network for Object Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4203-4212, doi: 10.1109/CVPR.2018.00442.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015

[14] Wang H, Nie F, Huang H. Robust distance metric learning via simultaneous l1-norm minimization and maximization[C]//International conference on machine learning. PMLR, 2014: 1836-1844.

[15] Yu J, Jiang Y, Wang Z, et al. Unitbox: An advanced object detection network[C]//Proceedings of the 24th ACM international conference on Multimedia. 2016: 516-520.

[16] 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," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 658-666, doi: 10.1109/CVPR.2019.00075.

[17] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12993-13000.

[18] Zhang Y F, Ren W, Zhang Z, et al. Focal and Efficient IOU Loss for Accurate Bounding Box Regression[J]. arXiv preprint arXiv:2101.08158, 2021.

[19] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999-3007, doi: 10.1109/ICCV.2017.324.

[20] Li. K, Wan. G, Cheng. G, Meng. L, Han. J, Object detection in optical remote sensing images: A survey anda new benchmark. ISPRS J. Photogramm. Remote Sens. 2020, 159, 296-307.

[21] S. Jiang et al., "An Optimized Deep Neural Network Detecting Small and Narrow Rectangular Objects in Google Earth Images," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 1068-1081, 2020, doi: 10.1109/JSTARS.2020.2975606.

[22] Q. Guo, L. Liu, W. Xu, Y. Gong, X. Zhang and W. Jing, "An Improved Faster R-CNN for High-Speed Railway Dropper Detection," in IEEE Access, vol. 8, pp. 105622-105633, 2020, doi: 10.1109/ACCESS.2020.3000506.

[23] H. Xie, Y. Li, X. Li and L. He, "A Method for Surface Defect Detection of Printed Circuit Board Based on Improved YOLOv4," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence

and Internet of Things Engineering (ICBAIE), 2021, pp. 851-857, doi: 10.1109/ICBAIE52039.2021.9390006.

[24] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang and D. Lin, "Libra R-CNN: Towards Balanced Learning for Object Detection," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 821-830, doi: 10.1109/CVPR.2019.00091.

[25] Q. Qian, L. Chen, H. Li and R. Jin, "DR Loss: Improving Object Detection by Distributional Ranking," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12161-12169, doi: 10.1109/CVPR42600.2020.01218.

[26] J. Wang, K. Chen, S. Yang, C. C. Loy and D. Lin, "Region Proposal by Guided Anchoring," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2960-2969, doi: 10.1109/CVPR.2019.00308.

[27] J. Sun, H. Ge and Z. Zhang, "AS-YOLO: An Improved YOLOv4 based on Attention Mechanism and SqueezeNet for Person Detection," 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2021, pp. 1451-1456, doi: 10.1109/IAEAC50856.2021.9390855.

[28] C. Gao, Q. Cai and S. Ming, "YOLOv4 Object Detection Algorithm with Efficient Channel Attention Mechanism," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2020, pp. 1764-1770, doi: 10.1109/ICMCCE51767.2020.00387.

[29] H. Zhai, J. Cheng and M. Wang, "Rethink the IoU-based loss functions for bounding box regression," 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 2020, pp. 1522-1528, doi: 10.1109/ITAIC49862.2020.9339070.

[30] Z. Tian, C. Shen, H. Chen and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9626-9635, doi: 10.1109/ICCV.2019.00972.

[31] Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T. et al. Selective Search for Object Recognition. Int J Comput Vis 104, 154-171 (2013). https://doi.org/10.1007/s11263-013-0620-5

[32] A. Shrivastava, A. Gupta and R. Girshick, "Training Region-Based Object Detectors with Online Hard Example Mining," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 761-769, doi: 10.1109/CVPR.2016.89.

[33] Zhang H., Chang H., Ma B., Wang N., Chen X. (2020) Dynamic R-CNN: Towards High Quality Object Detection via Dynamic Training. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision-ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12360. Springer, Cham.

[34] S. Zhang, L. Wen, Z. Lei and S. Z. Li, "RefineDet++: Single-Shot Refinement Neural Network for Object Detection," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 2, pp. 674-687, Feb. 2021, doi: 10.1109/TCSVT.2020.2986402.

[35] Li Y, Zhang Y, Huang X, et al. Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images[J]. ISPRS journal of photogrammetry and remote sensing, 2018, 146: 182-196.

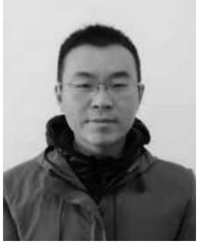
[36] Y. Li, Y. Zhang and Z. Zhu, "Error-Tolerant Deep Learning for Remote Sensing Image Scene Classification," in IEEE Transactions on Cybernetics, vol. 51, no. 4, pp. 1756-1768, April 2021, doi: 10.1109/TCYB.2020.2989241.

[37] X. Wang and J. Song, "ICIoU: Improved Loss Based on Complete Intersection Over Union for Bounding Box Regression," in IEEE Access, vol. 9, pp. 105686-105695, 2021, doi: 10.1109/ACCESS.2021.3100414.

[38] Liu L, Pan Z, Lei B. Learning a rotation invariant detector with rotatable bounding box[J]. arXiv preprint arXiv:1711.09405, 2017.

SHUANGJIANG DU was born in Xiangyang, Hubei, China, in 1996. He received his B.S. in aircraft engineering from National University of Defense Technology, Changsha, Hunan, China in 2018. He is currently pursuing the masterâĂŹs degree in optical information with the College of Communication Engineering, Army Engineering University. His research interests include machine learning, remote sensing and object detection and camouflage.

BAOFU ZHANG received the B.S. degree and M.S. degree in technical physics from Xidian University in 1987 and 1990, respectively. He is currently a professor with the College of Communication Engineering, Army Engineering University of PLA, the expert of Nanjing Workstation of China PLA General Political Deportment, the member of the expert group of doctoral supervisors, the senior member of China Electronic Association and China Institute of Communications. He was the reviewers of Acta Electronica Sinica, Journal on Communications and Acta Optica Sinica. He has published more than 30 articles in core journals, 4 monographs, 1 textbook for the 11th Five-Year Plan for higher education. His research interests include microwave photonics, photoelectric measurement, intelligent signal processing, and photoelectric reconnaissance and countermeasures. He received three first prizes for Teaching Achievements of PLA University of Science and Technology, two Third prizes for military Science and Technology Progress.



PIN ZHANG received the M.S. degree in optical engineering and the Ph.D. degree in science and technology of weapon from PLA University of Science and Technology, in 2013 and 2016, respectively. His research interests include electromagnetic camouflage, microwave photonics and optical remote sensing. He received the Special support of China Postdoctoral Fund, general First-class support of China Postdoctoral Fund, Jiangsu Provincial Natural Science Fund and the Technical Fund of the Foundation Strengthening Plan of the Military Commission of Science and Technology.