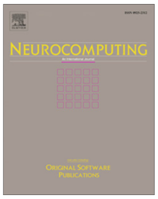
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Manhattan-distance IOU loss for fast and accurate bounding box regression and object detection

Yanyun Shen [a], Feizhao Zhang [a], Di Liu [a], Weihua Pu [b], Qingling Zhang [a,*]

[a] *School of Aeronautics and Astronautics, Sun Yat-sen University, Shenzhen Campus, Shenzhen, China*
[b] *Shenzhen Aerospace Dongfanghong Satellite Ltd., Shenzhen, China*

ARTICLE INFO

Article history:
Received 20 December 2021
Revised 5 May 2022
Accepted 14 May 2022
Available online 18 May 2022
Communicated by Zidong Wang

Keywords:
Manhattan-distance IOU loss
Bounding box regression
Object detection
Convolutional neural network

ABSTRACT

Bounding box regression is a crucial step in most object detection algorithms, and directly affects the positioning accuracy and regression speed of convolutional neural networks (CNN). The existing loss functions commonly used in bounding box regression suffer two main disadvantages: firstly, the l_n-norm loss does not match the evaluation metric Intersection over Union (IOU), leading to poor regression performance. Second, some recently proposed IOU-based loss functions are beneficial to IOU metric, but the negative effects of some terms in these loss functions on bounding box regression lead to slow convergence and inaccurate regression results. To solve these shortcomings, we proposed a Manhattan-Distance IOU (MIOU) loss function here. It takes into account that the Euclidean distance term in the Complete IOU (CIOU) loss and the Efficient IOU (EIOU) loss is unstable in training due to the huge gradient in the early stage of regression, and the Manhattan distance is added to effectively alleviate this defect. In addition, the denominator of the Euclidean distance term in the two loss functions discussed above has an antagonistic effect on loss reduction, and setting it as a normalized coefficient without participating in backpropagation can effectively improve the convergence speed. The effectiveness of the proposed MIOU loss was verified with designed simulation experiments. Moreover, object detection is usually applied to natural scenes and remote sensing scenes, but the application of detection methods are often limited due to varied image characteristics in different scene settings. We incorporated the MIOU loss into YOLO v4 and other mainstream object detection networks to examine its effectiveness in remote sensing and natural object detection scenarios. The experimental results on real remote sensing datasets DOTA and natural datasets MS COCO demonstrate that the MIOU loss has strong robustness in both remote sensing object detection tasks and natural object detection tasks. In summary, as a general regression loss function, the MIOU loss shows excellent performance in the above two types of scenes.

1. Introduction

Object detection algorithms emerged in the field of computer vision and now are widely used in many fields such as remote sensing, intelligent robot, unmanned driving, intelligent monitoring and so on. Since Everingham et al. [1] and Lin et al. [2] proposed some large-scale object detection datasets, a large number of object detection algorithms such as R-CNN [3] and YOLO [4] have sprung up, but with most of them are oriented to natural imagery.

As the main data source in the field of Earth observation, remote sensing images have long been widely used in disaster warning, municipal planning, surveying and mapping, resource exploration, and other applications. Remote sensing images often contain a large amount of land surface information and various types of objects, such as airplanes, ships, vehicles, and so on [5,6], retrieving which require effective object detection methods. For the earth observation community, it is a challenging but worthwhile task to develop efficient, accurate, and practical methods for remote sensing object detection.

Bounding box regression is one of the most critical steps in object detection. The two main tasks of object detection are classification and positioning. Classification [7] is the systematic arrangement in groups or categories according to established criteria and is the main task of deep learning networks. Positioning refers to determining the coordinates of target objects from an image. How to quickly and accurately regress anchor boxes, which represent the initial values of predefined candidate boxes, to ground-truth boxes during training stages has been a great challenge.

[*] Corresponding author.
E-mail address: zhangqling@mail.sysu.edu.cn (Q. Zhang).

https://doi.org/10.1016/j.neucom.2022.05.052
0925-2312/© 2022 Elsevier B.V. All rights reserved.

The most popular evaluation metric for bounding box regression is the Intersection over Union (IOU),

$$IOU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{1}$$

where $B^{gt} = \left(x^{gt}, y^{gt}, w^{gt}, h^{gt}\right)$ is a ground-truth box and $B = (x, y, w, h)$ is the corresponding predicted box. Generally, we call predicted boxes with IOU greater than a certain value (e.g. 0.6) as positive samples and those with IOU less than a certain value (e.g. 0.4) as negative samples. The intermediate predicted boxes (e.g. 0.4–0.6) are often abandoned because they can often confuse a network.

Currently, most object detection networks use the l_n-norm (e.g. n = 1 or 2) loss (in Eq. (3)) to measure the distance between a predicted box and its corresponding ground-truth box [4,8–11]. But Yu et al. [12] pointed out that at different regression stages when the l_n-norm loss between the predicted box and the ground-truth box are the same, their IOU might not be necessarily the same. Consequently, the l_n-norm loss is not a suitable choice to obtain the optimal IOU. The IOU loss was proposed to better match the IOU metric,

$$L_{IOU} = 1 - IOU \tag{2}$$

The IOU loss has achieved better IOU than the l_n-norm loss [12]. However, if there is no overlap between bounding boxes, it would stop working and no longer contribute to the gradient. To alleviate this problem, Rezatofighi et al. [13] proposed the Generalized IOU (GIOU) loss (in Eq. (5)). GIOU effectively relieves gradient vanishing problems caused by non-overlapping through adding an area compensation item (in Eq. (6)), but still has some limitations. As shown in the first row of Fig. 1, when the loss function is set as GIOU, during the regression process a predicted box first expands to encompass the target box from 50th iteration to 100th iteration, and then changes its area to increase the IOU from 100th iteration to 200th iteration. However, the regression process is often slow and might fail, because in this case GIOU cannot overcome the aspect ratio gap between the predicted box and the target box. To correct that weakness of GIOU, Zheng et al. [14] proposed the Distance IOU (DIOU) loss (in Eq. (7)) and the CIOU loss (in Eq. (9)), where the DIOU loss added the Euclidean distance between the center points of a predicted box and its corresponding ground-truth box to accelerate the regression process, and the CIOU loss further optimized the DIOU loss by adding a new item to reduce the aspect ratio gap between the predicted box and the target box. The DIOU loss regression example in Fig. 1 shows that the center point distance is reduced before the predicted box meets the target box, and then the center point distance and overlap area work simultaneously after the boxes intersect each other. The CIOU loss regression example in Fig. 1 shows that the predicted box starts to approach to aspect ratios of the target box during the regression process. But both the DIOU loss and the CIOU loss eventually fail due to their failure to cope with the huge aspect ratio gap. Zhang et al. [15] believed that the CIOU loss failed to reflect real relations between w and w^{gt} or h and h^{gt} and only reflected discrepancy between two aspect ratios, and proposed a new loss function called the EIOU (in Eq. (13)) loss. As shown in Fig. 1, the width and height of the predicted box in the EIOU regression case indeed get close to the target box during the regression process, but it still fails to regress at the end. Although the CIOU loss and the EIOU loss all contain Euclidean distance specially used to reduce the center point distance between the predicted box and the target box, we find that the center point of their prediction box is not clearly close to the corresponding target box before the 100th iteration (Fig. 1). The Euclidean distance might be the main reason the CIOU loss and the EIOU loss failed to regress. Through the above analysis, we can reach two conclusions: firstly, the Euclidean distance term (in Eq. (8)) in the two loss functions is unstable during training due to huge gradient in the early stages of regression; Secondly, the denominator of the Euclidean distance term in the two loss functions has an antagonistic effect on the loss reduction (more details are in the method section).

To overcome those drawbacks of the CIOU loss and the EIOU loss, we propose a faster and more accurate Manhattan-Distance IOU (MIOU) loss, adding a Manhattan distance term (in Eq. (16)) to the existing loss functions and setting all distance normalized denominator terms as coefficients so that they do not participate in backpropagation. Manhattan distance is the distance between two points measured along axes at right angles. It is often called l_1 loss (in Eq. (3)) when used in the loss function and can provide a stable gradient in the early stage of training, preventing gradient explosion, and increasing regression speed. As shown in Fig. 1, at the 50th iteration, the center points of bounding boxes in the MIOU regression case are already close to overlap, and the overlapped area and aspect ratios are significantly better than the final iteration of all previous IOU-based losses, ensuring a reliable regression. Fig. 1 is an example of regression experiment and the experimental setting can be found in Sec. 4.1. Please see Fig. 4 (a) for the complete simulated regression process.

The main contributions of this paper can be summarized as the following:

(1) We found that the main reason the CIOU loss and the EIOU loss sometimes fail to converge in bounding box regression is that the Euclidean distance term used by them converges slowly and even causes gradient explosion due to the large gradient in the early stage of training. At the same time, the normalized denominator of Euclidean distance term also plays an antagonistic role to the convergence of loss in the whole training stage.
(2) We designed the MIOU loss function for faster and more accurate bounding box regression to overcome the shortcomings of the CIOU loss and the EIOU loss.
(3) We incorporated the MIOU loss into YOLO v4 [16] to realize real-time high-precision multi-class remote sensing object detection without increasing additional calculation burden.
(4) The proposed method can be easily embedded into state-of-the-art detection algorithms to improve their performance.

2. Related works

2.1. Deep learning object detection methods in computer vision community

Deep learning object detection is one of the most important researches in the field of computer vision. In order to encourage researchers to design better object detection algorithms, object detection datasets such as PASCAL VOC [1] and Microsoft COCO (MS COCO) [2] were proposed and open sourced. According to whether the candidate boxes are extracted by a specific extraction method first, current deep learning object detection methods can be grouped into regression-based one-stage ones and region-based two-stage ones. Among them, one-stage methods pay more attention to the promotion of inference speed but with slightly lowered accuracy, while two-stage methods can achieve high-precision detection but with reduced speed.

The one-stage object detection algorithms were originated from Overfeat [17] and subsequently developed classic algorithms such as YOLO series, SSD, RetinaNet [4,8,16,18–20]. Overfeat [17] first uses CNN to implement the three main tasks of computer vision:

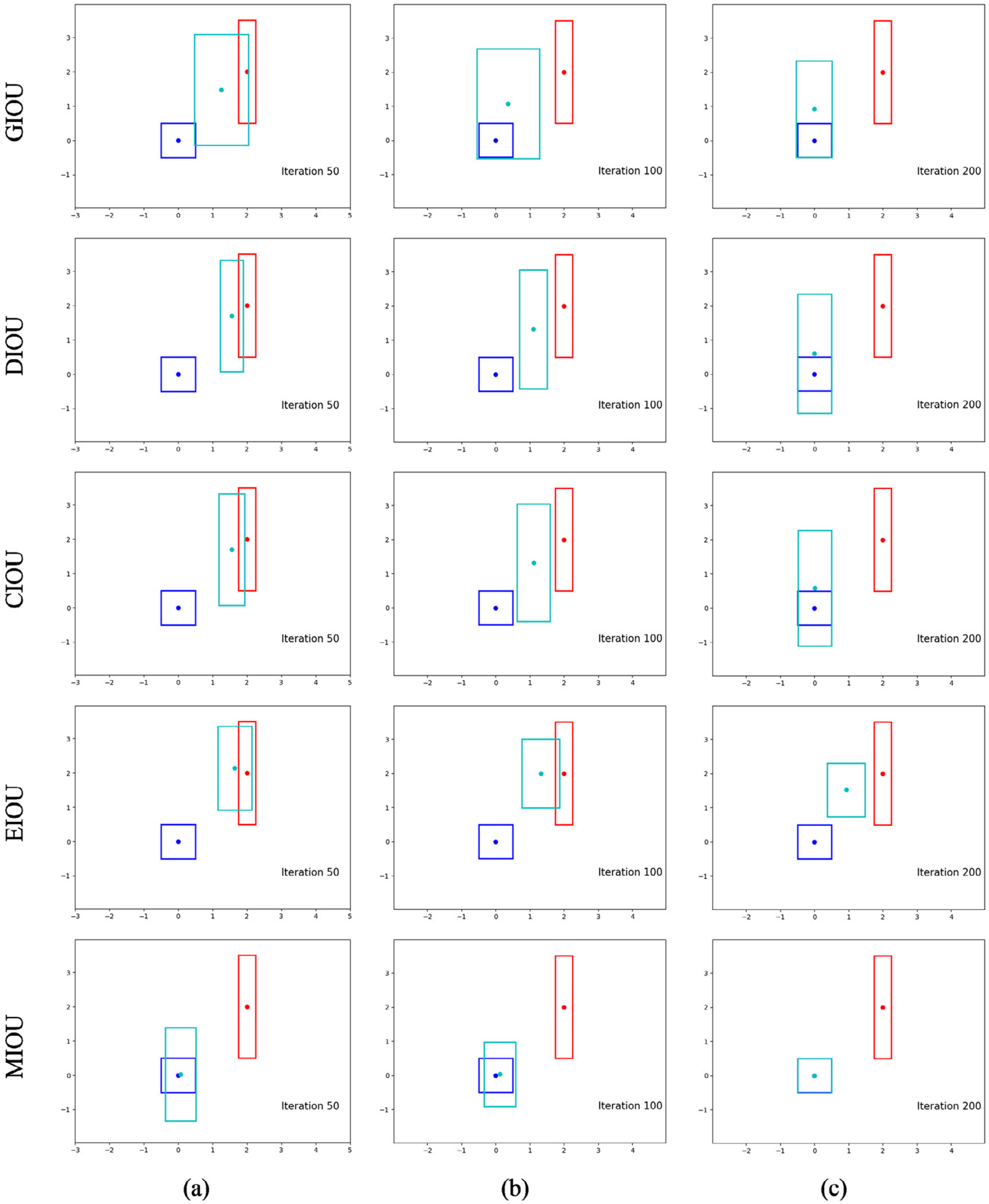


Fig. 1. Regression process of the bounding box under different loss functions simulation. The first row to the fifth row represents GIOU, DIOU, CIOU, EIOU, MIOU (proposed) respectively. In each panel, the blue box represents the target box, the red box represents the anchor box, and the green box represents the regression of the predicted box in each iteration, where iteration $\in$ [1,200]. Anchor box refers to the candidate box to be regressed generated according to certain rules. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

classification, positioning, and detection, opening the era of deep learning object detection. YOLO v1 [4] regards object detection as a regression problem, directly inputs the image to the network, and then regresses the bounding box position and category of the object at the output layer of the network, realizing an end-to-end network structure that became a classic algorithm in the industry later soon. SSD [20] classifies and regresses on the feature maps of different receptive fields, which improves the speed and accuracy compared with YOLO v1. YOLO v2 [18] introduces the anchor idea in Faster R-CNN [11] into YOLO v1, and replaces the full connection layer in the output layer of YOLO v1 with a convolution layer, which finally improves the accuracy and the speed. RetinaNet [8] analyzes that the main reason for the accuracy difference between one-stage and two-stage methods is the imbalance between foreground and background categories. It realizes hard example mining by proposing the focal loss, which greatly improves the accuracy of one-stage object detection methods. YOLO v3 [19] uses a better backbone DarkNet-53 and feature pyramid network (FPN) [21] structure to realize the multi-scale feature fusion. Compared with RetinaNet, YOLO v3 has achieved the same level of precision but its speed is four times faster, which has become the most popular object detection network at the time.

YOLO v4 [16] proposes CSPDarkNet53 as the backbone, uses PAN [22] to replace FPN [21], and applies a large number of tricks to make the accuracy reach the advanced level comparable to that of the two-stage networks.

Two-stage networks were originated from R-CNN [3], and subsequently a series of methods such as SPP-Net, Fast R-CNN, Faster R-CNN, R-FCN, Mask R-CNN [9–11,23,24] emerged. R-CNN [3] first proposes the concept of region proposal, which divides object detection into two steps: firstly, candidate boxes are obtained through selective search, and then feature extraction, classification, and regression are carried out according to the candidate boxes. SPP-Net [23] proposes a spatial pyramid pooling layer so that the network can input images of different sizes to avoid information loss caused by image cropping. Fast R-CNN [9] adds an ROI pooling layer on the basis of R-CNN, which extracts the location information of the region proposal from the feature map and converts it into a fixed size output. In addition, Fast R-CNN uses softmax instead of SVM classifier to increase the speed of the network. Faster R-CNN [11] proposes a new region proposal network (RPN) instead of selective search to generate candidate boxes, realizing end-to-end object detection and greatly improving the network detection speed. R-FCN [24] solves the contradiction between translation-invariance in image classification and translation-variance in object detection, and uses position-sensitive score maps to improve the detection speed as well as accuracy at the same time. Mask R-CNN [10] adds a new mask prediction branch based on the Faster R-CNN and simultaneously realizes two tasks: object detection and semantic segmentation, which is called instance segmentation.

2.2. Deep learning object detection methods in earth observation

Due to the peculiarities of remote sensing images such as large-scale, complex background and dense arrangement of targets, the existing object detection algorithms developed for natural scene imagery cannot be directly adopted for remote sensing object detection. Consequently, many object detection methods for remote sensing imagery have been developed. Bearing in mind the characteristics of large-scale and small targets of remote sensing images, Adam et al. [25] made adaptive improvements to the YOLO v2 [18] and proposed a complete remote sensing object detection process: image cutting, detection, and splicing for the first time, and solved the problem of a large amount of image information loss caused by directly scaling the image to a smaller size. In view of the complex background of remote sensing images, Li et al. [26] proposed a parallel down-up fusion network for salient object detection in optical remote sensing images, which takes full advantage of the in-path low- and high-level features and cross-path multi-resolution features to distinguish diversely scaled salient objects and suppress the cluttered backgrounds. Chen et al. [27] proposed a cascading attention network composed of a patching self-attention module and a supervised spatial attention module for enhancing feature representations for objects of interest and suppressing background noises in FPN from coarse to fine. In remote sensing imagery, densely arranged small targets such as ships and vehicles and large aspect ratio targets such as bridges and harbors are very important peculiarities, significantly different than those in natural images. To solve that challenge, Yang et al. [28] suppressed the noise and highlighted the object feature by joining the supervised pixel attention network and the channel attention network into the sampling fusion network to increase the detection accuracy for densely arranged small targets. Xu et al. [29] proposed an efficient feature-aligned single-shot detector (ASSD) to deal with the misalignment among predefined anchors, objects, and features extracted with a standard convolution kernel in both spatial and scale terms, especially for long-narrow and multi-scale geospatial objects.

Most remote sensing object detection networks mentioned above mainly focus on the improvement of accuracy, and some networks do focus on speed but have poor accuracy. In this paper, benefiting from the universality of the MIOU loss, we embedded it into YOLO v4 to achieve the optimal balance between accuracy and speed.

2.3. Loss functions for bounding box regression

Bounding box regression is a crucial step in object detection, referring to continuously refining the position of a predicted box from an initial anchor box to a target box. The l_n-norm loss has been widely applied in some classical object detection networks [4,8–11], and is defined as the following:

$$l_n(x) = \begin{cases} |x|, n = 1 \\ x^2, n = 2 \end{cases} \tag{3}$$

where x is the difference between a predicted box and its corresponding target box. Manhattan distance and Euclidean distance are used for l_1 loss and l_2 loss, respectively. Fast R-CNN [9] proposes the smooth l_1 loss to correct the instability of the l_2 loss (in Eq. (3)) used in R-CNN due to excessive gradient at the beginning of training. YOLO v1 [4] retains the l_2 loss on x and y of a bounding box and uses the square root on w and h to alleviate the scale sensitivity problem. The dynamic smooth l_1 loss [30] adds a dynamic factor to the smooth l_1 loss, making the network focus more on high-quality anchor boxes. Pang et al. [31] called the samples with a total loss less than 1.0 as inliers and those with a total loss greater than 1.0 as outliers. Based on the smooth l_1 loss, the balanced smooth l_1 loss was proposed to clip the large gradients produced by outliers with a maximum value of 1.0 and add factors to improve the regression loss of inliers.

The premise of using the l_n-norm loss is the assumption that the four parameters (x, y, w, h) of a bounding box are irrelevant. But as a matter of fact, this assumption is not always true, because different l_n-norm loss values may lead to the same IOU calculation results. In order to solve these problems, Yu et al. [12] proposed the IOU loss and achieved better results. Generally, an IOU-based loss can be defined as

$$L = 1 - IOU + R\left(B, B^{gt}\right) \tag{4}$$

where $R\left(B, B^{gt}\right)$ is the penalty term for a predicted box B and a target box B^{gt}.

Subsequently, Rezatofighi et al. [13] proposed the GIOU loss to alleviate gradient vanishing with the IOU loss for non-overlapping conditions, which has since been widely used in object detection networks such as ATSS [32] and YOLO v4. The GIOU loss is defined as the following:

$$L_{GIOU} = 1 - IOU + R_{GIOU} \tag{5}$$

$$R_{GIOU} = \frac{|C - B \cup B^{gt}|}{|C|} \tag{6}$$

where C is the smallest box enclosing B and B^{gt}. Considering that the IOU loss and the GIOU loss are only applied to axis-aligned 2D bounding boxes, Zhou et al. [33] broadened their application scenarios through the designing improved versions based on rotated bounding boxes and 3D bounding boxes [34–36]. Recently, the DIOU loss and the CIOU loss [14] have been proposed to deal with the slow regression phenomenon of the GIOU loss and achieved better regression speed and higher accuracy. Inspired by hard example mining of the focal loss [8], Zhang et al. [15] proposed the focal EIOU loss and applied it to Mask R-CNN, RetinaNet, and other object detection networks to prove its effectiveness. To overcome the slow convergence caused by the Euclidean distance term and the normalized denominator in the CIOU loss and the EIOU loss, this paper proposes a new Manhattan distance IOU loss, which has faster regression speed and higher regression accuracy.

3. The proposed method

In this section, we first analyze the inadequacy of the existing loss functions CIOU and EIOU, and then put forward the MIOU loss function. We then apply the new loss function to YOLO v4 to obtain an efficient and accurate remote sensing object detection network. We will introduce the MIOU loss and YOLO v4 in the next two parts.

3.1. Manhattan-distance intersection over union loss

3.1.1. Limitations of the IOU-based losses

3.1.1.1. Limitations of the distance IOU loss and the complete IOU loss. In order to solve the slow convergence problem of the GIOU loss and improve the regression accuracy, the DIOU loss uses the Euclidean distance to accelerate convergence and is defined as the following:

$$L_{DIOU} = 1 - IOU + R_{DIOU} \tag{7}$$

$$R_{DIOU} = \frac{\rho^2\left(b, b^{gt}\right)}{c^2} \tag{8}$$

where b and b^{gt} denote center points of B and B^{gt}, $\rho(\cdot)$ is the Euclidean distance, and c is the diagonal length of the smallest enclosing box covering the two boxes. Zheng et al. [14] proposed that a good bounding box regression loss should consider three important geometric factors: overlapping area, center point distance, and aspect ratio. The DIOU loss takes into account the overlapping area and the center point distance, directly minimizes the Euclidean distance between the predicted box and the corresponding target box, and increases the convergence speed. The CIOU loss further takes into account the aspect ratio to accelerate convergence, and is defined as the following:

$$L_{CIOU} = 1 - IOU + R_{CIOU} \tag{9}$$

$$R_{CIOU} = \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{10}$$

where α is a positive trade-off parameter, and v measures the consistency of the aspect ratio,

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}^{2}) \tag{11}$$

And the trade-off parameter α is defined as

$$\alpha = \frac{v}{(1 - IOU) + v} \tag{12}$$

Because both the DIOU loss and the CIOU loss are heavily rely on the R_{DIOU}, they have two common shortcomings. First, the R_{DIOU} and the l_2 loss both include the Euclidean distance and their derivatives to x will be very large when x is large, which will cause instability at the beginning of training and reduce the convergence speed. Second, the R_{DIOU} adopts the normalized Euclidean distance whose denominator term c^2 participates in the gradient backpropagation but will gradually decrease in the process of approaching a predicted box and the corresponding target box, thus counteracting the reduction of the R_{DIOU} term and reducing the convergence speed.

3.1.1.2. Limitations of the efficient IOU loss. As stated in Zhang et al. [15], the CIOU loss has the following defects: first, v only reflects the difference between the aspect ratios, rather than the real relationship between w and w^{gt} or h and h^{gt}; second, since $\frac{\partial v}{\partial w} = -\frac{h}{w}\frac{\partial v}{\partial h}$, $\frac{\partial v}{\partial w}$ and $\frac{\partial v}{\partial h}$ have opposite signs, if either w or h increases, the other will decrease, which is unreasonable, especially when$w < w^{gt}$and$h < h^{gt}$or$w > w^{gt}$and$h > h^{gt}$. To solve this problem of the CIOU loss, Zhang et al. [15] put forward the EIOU loss, which is defined as the following:

$$L_{EIOU} = 1 - IOU + R_{EIOU} \tag{13}$$

$$R_{EIOU} = \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \frac{\rho^2(w, w^{gt})}{c_w^2} + \frac{\rho^2\left(h, h^{gt}\right)}{c_h^2} \tag{14}$$

where c_w and c_h are the width and height of the smallest enclosing box covering the two boxes. However, because the EIOU loss uses the R_{DIOU} (in Eq. (14)), it has the same drawbacks as the CIOU loss does and the denominators c_w^2 and c_h^2 in the new term R_{EIOU} are similar to c^2, counteracting the decrease of R_{EIOU}. From Fig. 4 (a), we can see that compared with the methods proposed in this paper, the CIOU loss and the EIOU loss have poor performance in terms of regression speed, due to the weaknesses discussed above.

3.1.2. The CIOU-based Manhattan-distance IOU loss

As mentioned earlier, Manhattan distance and Euclidean distance are often called l_1 loss and l_2 loss respectively (in Eq. (3)) when used in the loss function. It can be seen from the derivative of the l_n-norm loss (in Eq. (3)) to x that the derivative of the l_1 loss to x is constant, which maintains constant gradient and prevents gradient explosion in the early stage of training. In the later stage of training, x becomes very small. If the learning rate remains unchanged, the l_1 loss will fluctuate around a stable value duo to the constant gradient, causing it difficult to converge to a higher accuracy. The derivative of the l_2 loss to x is positively correlated with x, which is large when the value of x is large, resulting in huge gradient and instability in the early stage of training. In the later stage of training, x is very small, and the corresponding gradient is also very small, which is conducive to the convergence of the loss function. To overcome the instability of the R_{DIOU} in the early stage of training, we add the R_{MIOU} to the CIOU loss to get the $MIOU - C*$ loss, accelerating loss reduction in the early stage of training and ensuring accuracy improvement in the later stage of training. It can be defined as the following:

$$L_{MIOU-C*} = 1 - IOU + R_{CIOU} + R_{MIOU} \tag{15}$$

where $*$ denotes the transition version of the L_{MIOU-C}

$$R_{MIOU} = \frac{\delta_x\left(b, b^{gt}\right)}{c_w} + \frac{\delta_y\left(b, b^{gt}\right)}{c_h} \tag{16}$$

where $\delta(\cdot)$ is the Manhattan distance, which is defined as the following:

$$\begin{cases} \delta_x\left(b, b^{gt}\right) = |x - x^{gt}| \\ \delta_y\left(b, b^{gt}\right) = |y - y^{gt}| \end{cases} \tag{17}$$

where (x, y) and (x^{gt}, y^{gt}) denote center points of a predicted box B and the corresponding target box B^{gt}.

We find that the denominator of the normalized distance term including both the Euclidean distance term and the Manhattan distance term has an antagonistic effect on the decline of the loss function because it provides an opposite gradient in the process of backpropagation. The counteraction of the denominator of the normalized distance term on the R_{DIOU} is proved by the following reasoning.

It is assumed that in the regression process the coordinate values of the four edges of a predicted box $B = (x, y, w, h)$ are greater than those of the corresponding edges of the corresponding target box $B^{gt} = \left(x^{gt}, y^{gt}, w^{gt}, h^{gt}\right)$, i.e.:

$$\begin{cases} x - w/2 > x^{gt} - w^{gt}/2 \\ x + w/2 > x^{gt} + w^{gt}/2 \\ y - h/2 > y^{gt} - h^{gt}/2 \\ y + h/2 > y^{gt} + h^{gt}/2 \end{cases} \tag{18}$$

So the width c_w and the height c_h of the smallest enclosing box enclosing the two boxes B and B^{gt} in (16) can be defined as the following:

$$\begin{cases} c_w = x - x^{gt} + \frac{w + w^{gt}}{2} \\ c_h = y - y^{gt} + \frac{h + h^{gt}}{2} \end{cases} \tag{19}$$

Then we can obtain R_{DIOU} from Eq. (9):

$$R_{DIOU} = \frac{\left(x - x^{gt}\right)^2 + \left(y - y^{gt}\right)^2}{c_w^2 + c_h^2} \tag{20}$$

According to whether the denominator of the R_{DIOU} is regarded as a coefficient without participating in backpropagation, we can get a different derivative $\partial R_{DIOU}/\partial x$, which is defined as the following:

$$\frac{\partial R_{DIOU}}{\partial x} = \begin{cases} \frac{2}{c^2}\left(x - x^{gt}\right) - \frac{2c_w \cdot \rho^2\left(b, b^{gt}\right)}{c^4}, & c^2 \textit{ is not a coefficient} \\ \frac{2}{c^2}\left(x - x^{gt}\right) & , c^2 \textit{ is a coefficient} \end{cases} \tag{21}$$

In the early stage of the regression process, when a predicted box does not contain the corresponding target box, c^4 and $c_w \cdot \rho^2\left(b, b^{gt}\right)$ in (21) gradually decrease, but c^4 obviously decreases faster than $c_w \cdot \rho^2\left(b, b^{gt}\right)$, so $2c_w \cdot \rho^2\left(b, b^{gt}\right)/c^4$ will increase. Therefore, when c^2 *is not a* coefficient, $\partial R_{DIOU}/\partial x$ will continue to decrease compared with the case where c^2 *is a* coefficient, which explains that when the denominator of the R_{DIOU} participates in derivation, the loss will decline slowly due to insufficient gradient in the early training process and thus the gradient of the denominator in (8) will play a confrontational role. In the later stage of the regression process, when a predicted box contains the corresponding target box, c_w/c^4 remains unchanged and $\rho^2\left(b, b^{gt}\right)$ decreases, which will cause $\partial R_{DIOU}/\partial x$ continue to increase when c^2 *is not a* coefficient relative to the case where c^2 *is a* coefficient and finally will result in a too large gradient, causing the loss not to converge.

The case in Eq. (18) means B is in the bottom-right direction of B^{gt}. When B is in the top-right direction of B^{gt}:

$$\begin{cases} x - w/2 > x^{gt} - w^{gt}/2 \\ x + w/2 > x^{gt} + w^{gt}/2 \\ y - h/2 < y^{gt} - h^{gt}/2 \\ y + h/2 < y^{gt} + h^{gt}/2 \end{cases} \tag{22}$$

$\partial R_{DIOU}/\partial x$ is the same as in Eq. (21).When B is in the top-left direction of B^{gt}:

$$\begin{cases} x - w/2 < x^{gt} - w^{gt}/2 \\ x + w/2 < x^{gt} + w^{gt}/2 \\ y - h/2 < y^{gt} - h^{gt}/2 \\ y + h/2 < y^{gt} + h^{gt}/2 \end{cases} \tag{23}$$

Then,

$$\frac{\partial R_{DIOU}}{\partial x} = \begin{cases} \frac{2}{c^2}\left(x - x^{gt}\right) + \frac{2c_w \cdot \rho^2\left(b, b^{gt}\right)}{c^4} & c^2 \textit{ is not a coefficient} \\ \frac{2}{c^2}\left(x - x^{gt}\right) & c^2 \textit{ is a coefficient} \end{cases} \tag{24}$$

In the whole stage of the regression process, when c^2*isnota*coefficient, $\partial R_{DIOU}/\partial x$ produces an additional term $2c_w \cdot \rho^2\left(b, b^{gt}\right)/c^4$ compared with the case c^2*isa*coefficient. At the early stage of the regression process, $2c_w \cdot \rho^2\left(b, b^{gt}\right)/c^4$ will continue to increase, which will accelerate the decline of loss but increase the risk of gradient explosion at the same time. At the later stage of the regression process, the extra gradient $2c_w \cdot \rho^2\left(b, b^{gt}\right)/c^4$ will make the loss difficult to converge.When B is in the bottom-left direction of B^{gt}:

$$\begin{cases} x - w/2 < x^{gt} - w^{gt}/2 \\ x + w/2 < x^{gt} + w^{gt}/2 \\ y - h/2 > y^{gt} - h^{gt}/2 \\ y + h/2 > y^{gt} + h^{gt}/2 \end{cases} \tag{25}$$

$\partial R_{DIOU}/\partial x$ is the same as in Eq. (24). In the above four cases, $\partial R_{DIOU}/\partial y$ and $\partial R_{DIOU}/\partial x$ have similar laws.

When the predicted box regresses from any direction (in Eqs. (18), (22)–(23), (25)):

$$\frac{\partial R_{DIOU}}{\partial w} = \begin{cases} -\frac{2c_w \cdot \rho^2\left(b, b^{gt}\right)}{c^4} & c^2 \textit{ is not a coefficient} \\ 0 & c^2 \textit{ is a coefficient} \end{cases} \tag{26}$$

When c^2 is not a coefficient, a permanent confrontation gradient will be generated in the whole training process, which will prevent the gradient descent and reduce the regression speed. There is a similar rule for $\partial R_{DIOU}/\partial h$.

To sum up, the influence of the denominator term participating in the backpropagation on $\partial R_{DIOU}/\partial x$ and $\partial R_{DIOU}/\partial y$ depends on the regression direction, while the influence of the term on $\partial R_{DIOU}/\partial w$ and $\partial R_{DIOU}/\partial h$ is independent of the direction and it will produce a permanent confrontation gradient to hinder the regression. In order to overcome that shortcoming, we define denominators of the normalized distance term as coefficients without participating in the backpropagation and get the final MIOU loss version, i.e.

$$L_{MIOU-C} = 1 - IOU + \alpha v + \beta \rho^2\left(b, b^{gt}\right) + \gamma \delta\left(b, b^{gt}\right) \tag{27}$$

where the normalization parameter β and γ are defined as

$$\beta = \frac{1}{c^2} \tag{28}$$

$$\gamma = \begin{cases} \frac{1}{c_w}, \delta = \delta_x \\ \frac{1}{c_h}, \delta = \delta_y \end{cases} \tag{29}$$

Simulation experiments in Fig. 4(c) have verified the performance improvement brought by treating the denominator term as a coefficient.

3.1.3. The EIOU-based Manhattan-distance IOU loss

Because the EIOU loss also contains the Euclidean distance term (14), it also has the defects mentioned above. We add the R_{MIOU} to the EIOU loss to get $MIOU - E*$ loss, which is defined as

$$L_{MIOU-E*} = 1 - IOU + R_{EIOU} + R_{MIOU} \tag{30}$$

After defining the denominator of the normalized distance term as a coefficient, the final version obtained is as

$$L_{MIOU-E} = 1 - IOU + \varepsilon \rho^2\left(a, a^{gt}\right) + \beta \rho^2\left(b, b^{gt}\right) + \gamma \delta\left(b, b^{gt}\right) \tag{31}$$

where a and a^{gt} denote w or h of B and B^{gt} respectively, and the normalization parameter ε is defined as.

$$\varepsilon = \begin{cases} \frac{1}{c_w^2}, a = w; a^{gt} = w^{gt} \\ \frac{1}{c_h^2}, a = h; a^{gt} = h^{gt} \end{cases} \tag{32}$$

3.2. MIOU loss-based YOLO v4 for remote sensing object detection

YOLO v4 [16] is one of the most advanced one-stage object detection algorithms at present. Compared with YOLO v3 [19], YOLO v4 uses a lot of deep learning convolution neural network tricks, which comprehensively improves the performance of a network in many aspects.

In order to facilitate the introduction of the network structure, this paper uses three levels to represent the basic units of the network. As shown in Fig. 2, CBL (CONV 3 × 3, batch normalization, and leaky relu activation function), CBM (CONV 3 × 3, batch normalization, and Mish activation function) and spatial pyramid pooling (SPP) are the three primary units of the network. The residual unit (RU) is the secondary level unit of the network, which is mainly composed of CBM and a cross stage feature map addition. For the sake of deepening the number of network layers, the RU uses the residual idea of the ResNet [38] to construct the network structure. Cross Stage Partial X (CSPX) is the third level unit of the network, which is composed of CBM, X RUs and the cascade of cross stage feature maps. The idea of CSPX mainly comes from the CSPNet [39], which noticed that the repetition of gradient information in the network is the main reason for the excessive amount of inference calculation. Therefore, the feature map is divided into two parts: one part passes through the original network channels and the other remains unchanged. At the end the two parts are combined through cross stage cascade to reduce the calculation cost while ensuring the accuracy.

The improvement of YOLO v4 is reflected in data augmentation and network structure update. Mosaic technique is used for data augmentation through randomly scaling and distributing four pictures and finally splicing them into a new picture to increase the diversity of datasets. Network structure update mainly includes backbone, neck, loss, etc. Based on Darknet53 [19], YOLO v4 combines the structure of CSPNet and proposes a new backbone CSPDarknet53 that contains five CSPX units. As shown in Fig. 2, each CSPX unit will perform a downsampling operation because the stride of their first convolution kernel is 2. Therefore, the feature map sizes obtained after passing through the last three CSPX structures are 128×128, 64×64 and 32×32 respectively, which are 8, 16 and 32 times downsampled compared with the input image. In order to further alleviate the overfitting of the network, YOLO v4 uses Dropblock [40] in the CSPDarknet53. Its idea is similar to Dropout and reduces the network complexity by discarding weights. However, the difference is that dropout is mostly applied in the full connection layer, while the convolutional layer and pooling layer can learn the same information through adjacent units, resulting in no obvious effect of dropout on the convolution layer. Dropblock is effective for the convolutional layer by randomly clearing a certain area on the convolutional layer. In the neck part of YOLO v4, PAN [22] is used instead of FPN [21]. As shown in Fig. 2, the red dashed box represents the FPN structure. Through the top-down multi-scale fusion strategy, the high-level semantic information is effectively transmitted to the bottom layer. The blue dashed box is a bottom-up FPN structure, which effectively transmits the location information of the bottom layer to the top layer, PAN realizes effective feature fusion through the combination of two FPN structures and improves the network's inferring accuracy.

YOLO v4 uses advanced CIOU loss [14] as the regression loss function. This paper incorporates the proposed MIOU loss into YOLO v4, which effectively improves the performance of the network and achieves optimal balance between accuracy and speed in remote sensing scenarios.

4. Experiments

4.1. Data description and experiment settings

4.1.1. Data description

In this paper, simulated data and real datasets are used to verify the effectiveness of the proposed method. In the simulation experiment, in order to simulate actual bounding box regression as much as possible, simulation data are set from three aspects: distance, scale and aspect ratio. As shown in Fig. 3, we first use cell boxes with seven different aspect ratios (1:4, 1:3, 1:2, 1:1, 2:1, 3:1 and 4:1) as target boxes, and fixing the center point at (3, 3).

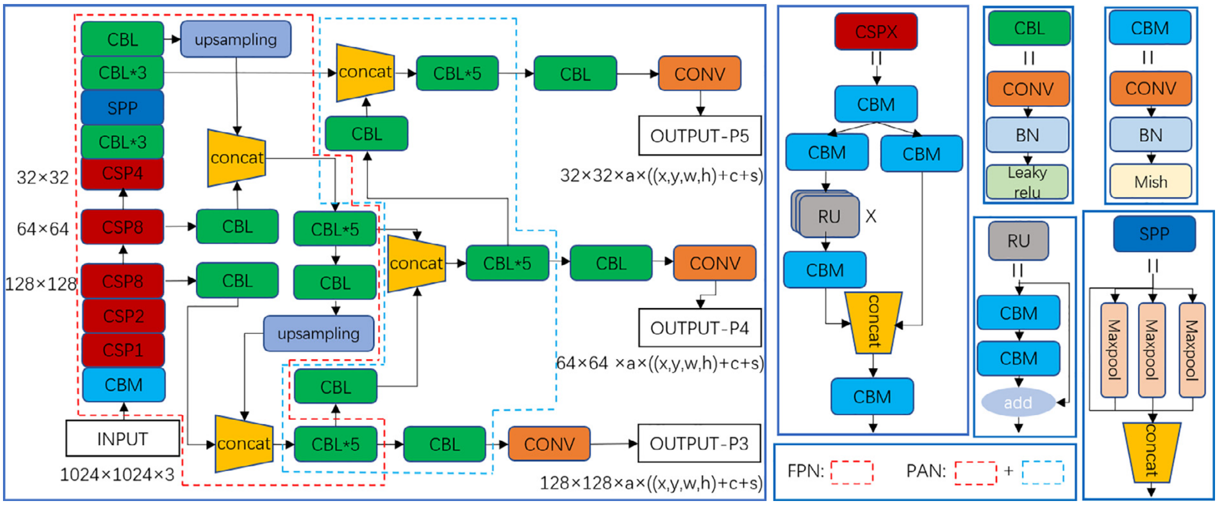


Fig. 2. The YOLO v4 network architecture. The backbone of the network uses CSPDarknet53 [16], and the neck of the network uses the PAN [22] structure. CBL is composed of a CONV, a BN and a leaky relu activation function, where CONV means CONV3 × 3, BN indicates batch normalization. CBM is composed of a CONV, a BN and a mish activation functions [37]. SPP cascades different scale pooling result feature maps obtained by the four 1 × 1, 5 × 5, 9 × 9, 13 × 13 maximum pooling operations to realize the multi-scale feature fusion. RU is composed of two CBM and a cross stage feature map addition. CSPX is composed of five CBM, X RUs and the cascade of cross stage feature maps. The output parameters a, (x, y, w, h), c, and s represent the anchor number, predicted boxes, class number, and score of each grid respectively.

Then, 1000 points are randomly generated in a circle whose center is set at (3, 3) and radius is set as 3. Anchor boxes with 7 areas (0.5, 0.67, 0.75, 1, 1.33, 1.5 and 2) and 7 aspect ratios (1:4, 1:3, 1:2, 1:1, 2:1, 3:1 and 4:1) are generated with each point randomly generated above as the center. The whole simulation experiment contains $1000 \times 7 \times 7 \times 7$ regression cases, including nonoverlapping, large aspect ratios and other difficult cases, in order to make the experimental results more representative.

Real dataset used in this study are the large scale remote sensing object detection dataset DOTA [5], including 2806 remote sensing images with sizes ranging from 800 × 800 to 4000 × 4000. This dataset is divided into 1411 training data, 458 validation data and 937 test data. There are totally 15 object categories in the DOTA: baseball diamond (BD), bridge (BR), basketball court (BC), ground track field (GTF), helicopter (HC), harbor (HA), large vehicle (LV), plane (PL), roundabout (RA), small vehicle (SV), storage tank (ST), soccer ball field (SBF), swimming pool (SP), ship (SH), and tennis court (TC). We cut the training set, validation set and test set into 1024 × 1024 tiles with an overlap rate of 50%, 20% and 20% respectively and finally obtain 29457, 5297, 10,833 images for network input. We patch 0 values for images whose sizes are less than 1024 × 1024. Finally, in the discussion section, we further use the MS COCO [2] dataset to evaluate the robustness of our proposed MIOU loss on other object detection networks.

4.1.2. Experiment settings

Algorithm 1 Simulation Experiments

```
Input: $\{\{B_{n,s}\}_{s=1}^{S}\}_{n=1}^{N}$ is the set of anchor boxes at N = 1000
   randomly distributed points within
   the circular region with a center (3, 3) and a radius 3, and
   S = 7 × 7means numbers of anchor boxes combined in 7
   areas and 7 aspect ratios. $\{B_i^{gt}\}_{i=1}^{7}$ is the set of target boxes
   that are fixed at (3, 3) with area 1, and 7 different aspect
   ratios.
Output: Regression error $E \in R^{T \times N}$, where T is the maximal
   iteration.
1:  $(\mathrm{E},\ \mathrm{T}) \leftarrow (0, 200)$
2:  for n = 1 to N do
3:    for s = 1 to S do
4:      for i = 1 to 7 do
5:        for t = 1 to T do
6:          if $t \leq 0.8T$ then $\eta = 0.1$
7:          else if $t \leq 0.9T$ then $\eta = 0.01$
8:          else $\eta = 0.001$
9:          end
10:         $\nabla B_{n,s}^{t-1} = \frac{\partial L\left(B_{n,s}^{t-1}, B_i^{gt}\right)}{\partial B_{n,s}^{t-1}}$
11:         $B_{n,s}^{t} = B_{n,s}^{t-1} + \eta \nabla B_{n,s}^{t-1}$
12:         $\mathrm{E}(\mathrm{t},\mathrm{n}) = \mathrm{E}(\mathrm{t},\mathrm{n}) + \left|B_{n,s}^{t} - B_i^{gt}\right|$
13:       end
14:     end
15:   end
16: end
17: return E
```

All experiments are implemented on PyTorch. For the simulation experiment, the process is shown in Algorithm 1. For all cases, a total of 200 iterations are set. We train the model with simulation experiments on a CPU i7-9700k, adopting a stochastic gradient descent (SGD) optimizer with an initial learning rate 0.1 and decaying it by a factor of 0.1 at 160th iteration and 180th iteration. When the iteration is t, the predicted box B_i^t can be expressed as:

$$B_i^t = B_i^{t-1} + \eta \nabla B_i^{t-1} \tag{33}$$

where B_i^{t-1} represents the value of the predicted box B_i when the iteration is t-1, η represents the learning rate, and ∇B_i^{t-1} represents the gradient of loss relative to B_i when the iteration is t-1. In line 12 of Algorithm 1, the error between a predicted box and the corresponding target box in each iteration is saved. The final output E represents the sum of errors in each iteration for all cases.

In order to explore the reliability of the MIOU loss on CNNs with different network complexity, we adopt the strategy in [41] to divide YOLO v4 into YOLO v4 s, YOLO v4 m, YOLO v4 l, and YOLO v4 x, and the numbers of their parameters are 0.2, 0.5, 1.0, and 1.8 times that of YOLO v4 respectively.

For the DOTA object detection experiments, all MIOU loss-based YOLO v4 run no more than 300 epochs. These models are trained on one Nvidia RTX 3090 GPU, with a stochastic gradient descent (SGD) optimizer and a cosine decay learning rate scheduling strategy with an initial learning rate 0.1. The warm-up steps are maximum of 3 epochs and 1000 iterations. The momentum and weight decay are set as 0.937 and 0.005 respectively.

4.2. Evaluation metrics

In this paper, we use two metrics PASCAL VOC [1] and MS COCO [2]. PASCAL VOC metric includes precision, recall, AP, mAP and other parameters:

$$precision = \frac{TP}{TP + FP} \tag{34}$$

$$recall = \frac{TP}{TP + FN} \tag{35}$$

where TP, FP, FN indicates true positive, false positive, false negative, respectively. When the IOU of the predicted box and target box exceeds 0.5, the predicted box is called positive, otherwise negative

$$AP = \int_0^1 p(r)d_r \tag{36}$$

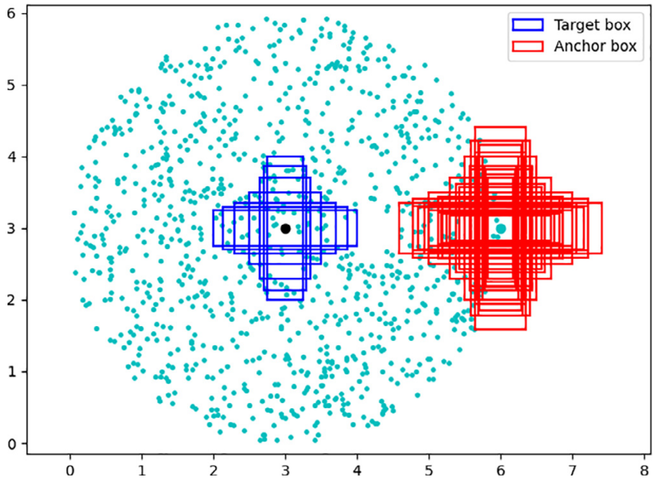


Fig. 3. Simulated data.

where p is precision, r is recall, and AP is average precision representing the area under the precision-recall curve in the coordinate system with recall on the horizontal axis and precision on the vertical axis

$$mAP_{voc} = \frac{\sum_{c=1}^{n} AP_c}{n} \tag{37}$$

where c represents category, n represents the total number of categories, and mAP_{voc}(mean average precision) represents the average value of AP for multiple categories.

The MS COCO metric includes precision, recall, mAP_{50}, mAP_{75}, mAP_{coco}, mAP_s, mAP_m, mAP_l. mAP_{50} is equivalent to mAP_{voc}, indicating mAP when IOU > 0.5. mAP_{75} indicates mAP when IOU > 0.75. mAP_{coco} indicates average value of mAP under the threshold of 10 IOU (0.50:0.05:0.95), which can better evaluate the positioning capability of an object detection network. mAP_s means mAP for small targets with areas less than 32 × 32, and mAP_m means mAP for medium targets with areas greater than 32 × 32 and less than 96 × 96, while mAP_l means mAP of large targets with areas greater than 96 × 96.

This paper uses frames per second (FPS) to evaluate the speed of a network.

4.3. Simulation experiments

Fig. 4(a) represents the regression error sum curves of different loss functions. From the Fig. 4(a), we can draw the conclusion that IOU, DIOU, CIOU, EIOU are in line with the conclusions of Zhang et al. [15] and Zheng et al. [14]. The reason why GIOU and CIOU have nearly the same effect is that the maximum distance between an anchor and the corresponding target in our case is set as no more than 3, and the number of sampling points in Zheng et al. [14] is 3000, while ours is 1000, which reduces many difficulties and weakens the main advantages of CIOU over GIOU. At the same time, there are some differences in parameter settings. MIOU-E and MIOU-C proposed in this paper can reach the minimum regression error. Although MIOU-E decreases faster at the early stage, their final results are very close.

Fig. 4(b) is the result of ablation studies on different distance items of the IOU-based loss function. From the curve, we can see that DIOU-M decreases faster than DIOU at the early stage, but DIOU exceeds at the middle and late stage, and finally the two IOU-based losses achieve the same effect. In addition, we can get similar results from the curves of EIOU to EIOU-M and the curves of CIOU to CIOU-M. Finally, we guess that the IOU-based loss with Manhattan distance is easier to converge than the IOU-based Loss with Euclidean distance at the early stage, but the derivative of Manhattan distance is constant and will fluctuate near the stable value at the later stage, which is made up by the Euclidean distance item. Combining the Manhattan distance item and the Euclidean distance item together can improve both speed and accuracy at the same time.

In Fig. 4(d), CIOU adds the Manhattan distance item to obtain MIOU-C*, which is superior to CIOU and CIOU-M in terms of speed and accuracy in the whole training period, verifying our conjecture. Fig. 4(c) is the result of ablation studies on the MIOU loss function whether the denominator of the distance item is derived. In MIOU-C*, denominator refers to the denominator of the Manhattan distance item and the Euclidean distance item. In MIOU-E*, denominator refers to the denominator of the Manhattan distance item, the Euclidean distance item and the aspect item. According to the proposed method, we believe that the denominator, representing the attribute of minimum enclosing rectangle of both anchor and target, will become smaller and smaller with the regression process, resulting in larger and larger loss function and counteracting the loss decline. Therefore, it is used as a coefficient and does not participate in the backpropagation process. According to the experimental results, it can be seen that the decline speed of the error curve of MIOU-C is significantly higher than that of MIOU-C*, and the final error of MIOU-C is smaller. MIOU and MIOU-E* also show the same law, which proves our proposed method.

4.4. Remote sensing object detection results with the proposed method

4.4.1. Ablation studies on IOU-based loss

In order to verify its effectiveness of the proposed method in real data, we apply different IOU-based losses to YOLO v4 s, use the DOTA's training dataset to train the network, and the DOTA's validation dataset to verify its accuracy. According to Table 1, we can see that the performance of CIOU and EIOU in YOLO v4 s is better than that of GIOU. After adding the ND strategy, all indices of CIOU are improved, where mAP_{coco} is increased from 49.1% to 49.6%. The result proves that the denominator of the normalized distance term will indeed play a role in combating the loss reduction. Setting the denominator as a coefficient so that it does not participate in backpropagation can effectively curb its negative effect. After adding MD to CIOU, the mAP_{coco} has increased from 49.1% to 49.4%, which proves that adding Manhattan distance can effectively make up for the instability of CIOU in the early stage of training and improve the accuracy of the network.

After adding ND and MD to CIOU simultaneously, the mAP_{coco} is increased from 49.1% to 49.7%, and the precision is increased by 2.5% to 64.8%, which shows that ND and MD can improve the performance of CIOU at the same time, and MIOU-C is more suitable to be as an object detection loss function than CIOU. For CIOU, compared with just adding ND, the application of both two strategies led to the increase of mAP_{coco} but the decrease of mAP_{50} and mAP_{75}. According to the principle of mAP_{coco}, the latter can obtain better mAPs with larger IOU thresholds, which shows that MD can promote the regression accuracy of the box.

After adding ND and MD to EIOU respectively, all indicators of EIOU are improved, which verifies the effectiveness of these two strategies again. After adding ND and MD to obtain MIOU-E version at the same time, compared with the original EIOU, the mAP_{coco} is increased by 0.5%, and the mAP_{50} is increased by 0.7% and the recall is increased by 3.7%, indicating that MIOU-E is more suitable to be as loss function of bounding box regression task. For EIOU, compared with just adding MD, the application of both strategies led to 0.4 increase for mAP_{50} but 0.6 decrease for mAP_{75}. Due to the same mAP_{coco}, we found that adding ND based on the MD strategy did not achieve significant improvement, but it would not hinder MD.

In summary, through the above comparative experiments, we verify that the ND and MD strategies have a certain improvement effect on both CIOU and EIOU, and improve the accuracy on the DOTA verification set.

4.4.2. Ablation studies on YOLOv4

In order to verify the effectiveness of the proposed MIOU loss function on different complexity of deep learning networks, we apply CIOU and MIOU to YOLO v4 s, YOLO v4 m and YOLO v4 l. Then we use training data set and validation data set of DOTA to train and verify the networks respectively.

As shown in Table 2, MIOU performs better than CIOU on all the three complexities, and FPS does not decrease with the shifting of loss functions, indicating that MIOU can increase the object detection accuracy without affecting the network inference speed. In the case of YOLO v4 s, precision, recall, mAP_{75}, and mAP_{coco} are improved 2.5%, 0.1%, 0.6%, and 0.6% respectively for MIOU, while mAP_{50} remains unchanged. For YOLO v4 l, precision, recall, mAP_{50}, mAP_{75}, and mAP_{coco} have achieved overall improvement of 1.3%, 0.6%, 0.5%, 0.1%, and 0.3% respectively for MIOU. In the case

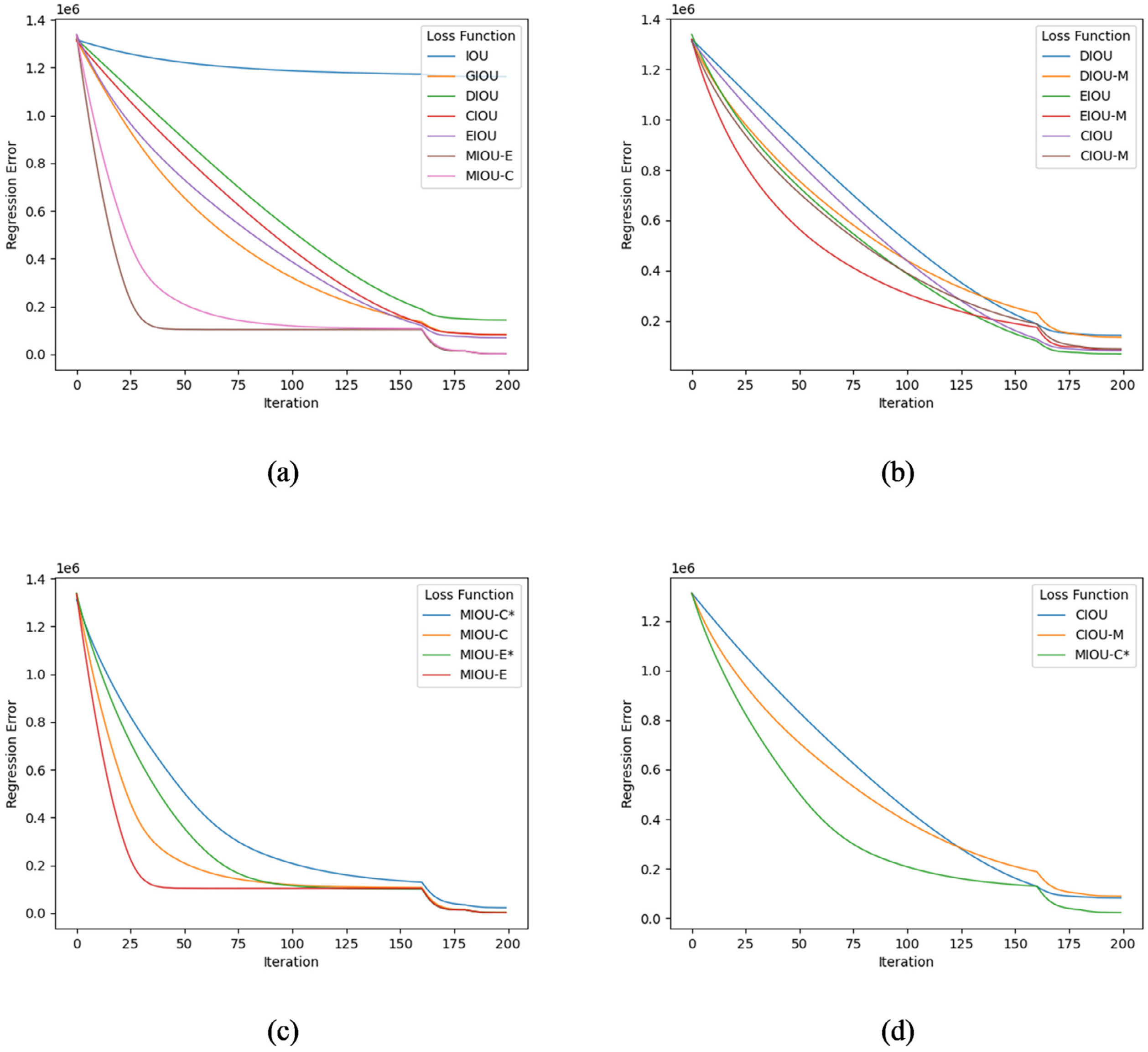


Fig. 4. Simulation experimental results: (a) Regression error sum curves of different loss functions. (b) Ablation studies on different distance item of IOU-based loss function. M indicates using the Manhattan-Distance item instead of the Euclidean-Distance item. (c) Ablation studies on MIOU loss function whether the denominator of the distance item is derived. (*) indicates denominator of the distance item of MIOU is derived. (d) Ablation studies on MIOU loss function about effect of Manhattan-Distance.

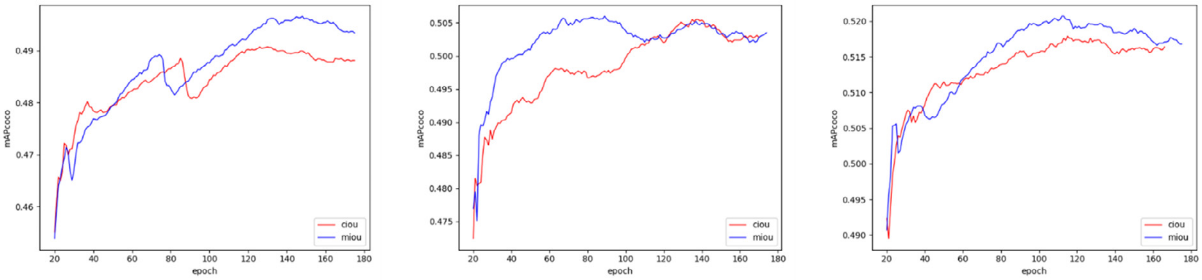
Table 1
Ablation Studies on IOU-based Loss, ND indicates that the denominator of all distance terms in the loss function does not participate in the derivation, and MD indicates that Manhattan distance is added to the loss function.

Method	ND	MD	Precision	Recall	mAP_{50}	mAP_{75}	mAP_{coco}
GIOU	–	–	61.7	79.6	73.6	52.8	48.5
CIOU-M	–	✓	61.2	78.9	73.9	53.2	48.8
CIOU	–	–	62.3	79.3	74.6	53.8	49.1
	✓	–	63.2	**79.8**	**75.2**	**54.7**	49.6
	–	✓	63.6	79.4	74.4	54.2	49.4
	✓	✓	**64.8**	79.4	74.6	54.4	**49.7**
EIOU-M	–	✓	61.7	80.2	74.1	53.0	48.6
EIOU	–	–	63.0	79.3	74.1	53.7	48.9
	✓	–	**66.3**	79.0	74.6	53.8	49.0
	–	✓	64.5	79.0	74.4	**54.3**	49.4
	✓	✓	63.7	**83.0**	**74.8**	53.7	**49.4**

of YOLO v4 m, MIOU achieved a slight increase in the overall index mAP_{coco} and a slight decrease in Precision and mAP_{75}. From Fig. 5 (b), we can find that starting from the 20th epoch, the network accuracy improvement speed of the MIOU-based YOLO v4 m is higher than that of the CIOU-based YOLO v4 m, and there is an obvious gap at the 40th epoch. At the 80th epoch, the MIOU-

Table 2
Ablation Studies on YOLO v4 with three complexities.

Complexity	Methods	Precision	Recall	mAP_{50}	mAP_{75}	mAP_{coco}	FPS
s	CIOU	62.3	79.3	74.6	53.8	49.1	82
	MIOU – C	**64.8**	**79.4**	74.6	**54.4**	**49.7**	82
m	CIOU	**61.5**	81.3	75.7	**56.1**	50.5	69
	MIOU – C	61.1	**81.4**	**75.8**	55.9	**50.6**	69
l	CIOU	59.3	82.4	76.6	57.7	51.8	55
	MIOU – C	**60.7**	**83.0**	**77.1**	**57.8**	**52.1**	55



(a)

(b)

(c)

Fig. 5. Ablation studies on volumed YOLOv4. (a), (b) and (c) show the accuracy comparison of MIOU and CIOU on the validation set at different stages of training when the networks are YOLO v4 s, YOLO v4 m and YOLO v4 l, respectively.

based YOLO v4 m network is close to fitting, and mAP_{coco} reaches the highest value of 50.6%. But at the same time, the mAP_{coco} of the CIOU-based YOLO v4 m is only about 49.7%, and it didn't reach the peak value of 50.5% until about the 140th epoch, indicating that although the final performance of MIOU on YOLO v4 m is only slightly better than that of CIOU, its network fitting speed is much higher than that of CIOU. At the same time, it is also verified that MIOU can well alleviate the instability of CIOU in the early stage of training, thus accelerating the network's convergence. According to Fig. 5\ (a), both MIOU and CIOU show large fluctuation at the 80th and 90th epoch, respectively. We speculate that it is a short-term instability, mainly due too few network parameters in YOLO v4 s and complex training data. Then they are fitted at the 130th epoch. Finally, MIOU performs better than CIOU. According to Fig. 5(c), the accuracy of MIOU starts to distance from CIOU at the 60th epoch, and reaches the maximum value at the 110th epoch. Combined with Table 2, it can be concluded that MIOU performs better than CIOU on YOLO v4 l. To sum up, experimental results prove that MIOU is effective on different complexity of deep learning networks.

4.4.3. Ablation studies on multi-class remote sensing object detection

In order to examine the effectiveness of MIOU loss for multi-class remote sensing object detection, we use the DOTA training set to train the network and the DOTA validation set to verify the accuracy, apply the CIOU loss and the MIOU loss to YOLO v4 with different complexity, and obtain the accuracy comparison results for different categories. As shown in Table 3, in the case of YOLO v4 s, mAP_{voc} of MIOU is increased by 0.3%, and accuracy is improved for 9 categories, is unchanged for 1 category, and is decreased for 5 categories, compared with CIOU. In the case of YOLO v4 m, mAP_{voc} of MIOU is increased by 0.5%, and accuracy is improved for 10 categories and is decreased for 5 categories, compared with CIOU. In the case of YOLO v4 l, mAP_{voc} of MIOU is increased by 0.7%, and accuracy is improved for 9 categories, is unchanged for 2 categories, and is decreased for 4 categories, compared with CIOU.

To sum up, we can see that MIOU loss can improve the overall accuracy of multi-class remote sensing object detection. It can improve the accuracy in most categories but cause accuracy loss in a few categories, which shows that MIOU loss is effective in improving the regression of most categories yet is not robust enough for a few categories. This may be because objects with different aspect ratios and scales have different requirements for the loss function. At the same time, we observe that the categories of accuracy decline are different in different complexity networks, indicating that the amount of network parameters is also one of the main reasons affecting category loss. Furthermore, For BD, BR, GTF, BC, SBF, and RA, we can find that the accuracy of MIOU-C decreases by 0.4–2.2% compared with CIOU in networks with different complexity. However, in these categories, MIOU-C performs better in at least one complexity. According to Fig. 6, we can find that the DOTA dataset has a very serious class imbalance, and the numbers of samples in the six controversial categories are very small. The lack of samples leads to underfitting and instability of training, which cannot highlight the performance of MIOU-C. But looking at the whole dataset, MIOU-C still performs better than CIOU.

4.5. Comparison with State-of-the-Arts

In order to verify the reliability of the MIOU loss-based volumed YOLO v4 in speed and accuracy, we compared it with state-of-the-art one-stage object detection algorithm with the DOTA test set. As shown in Table 4, mAP_{voc} and FPS of ASSD are 76.0% and 21% respectively, reaching the highest accuracy and speed in the one-stage network, and achieving the current optimal balance of speed and accuracy in the remote sensing object detection dataset. Compared with ASSD, mAP_{voc} of the proposed MIOU loss-based YOLO v4 l is 0.3% higher, and the speed reaches 55 FPS, almost three times the speed of ASSD, which has good practical value. Compared with ASSD, mAP_{voc} of the proposed MIOU loss-based YOLO v4 x is 0.7% higher, and the FPS is 35, nearly twice higher, which reaches the current optimal value in speed and accuracy at the same time, realizing a new optimal balance between speed and accuracy. Fig. 7

Table 3
Ablation studies on multi-class remote sensing object detection.

Complexity	Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP_{voc}
s	CIOU	95.5	**76.1**	**51.6**	**57.3**	67.2	**86.0**	91.8	95.7	71.2	**90.7**	51.4	73.1	82.8	69.8	71.1	75.4
	MIOU-C	**95.7**	75.2	50.3	56.3	**67.5**	85.6	91.8	**96.4**	**72.7**	90.4	**54.3**	**73.2**	**82.9**	**69.9**	**73.1**	**75.7**
m	CIOU	95.8	**79.3**	**55.8**	56.1	**70.2**	86.0	91.6	96.0	72.6	92.0	**56.9**	**75.8**	83.9	65.0	73.4	76.7
	MIOU-C	**96.2**	78.9	54.0	**58.4**	70.0	**86.2**	**91.9**	**96.1**	**75.2**	**92.2**	56.0	73.7	**84.8**	**69.2**	**75.1**	**77.2**
l	CIOU	**96.4**	77.8	54.6	57.8	71.3	86.6	92.0	96.1	**81.1**	**93.0**	**53.6**	73.5	84.6	70.6	73.7	77.5
	MIOU-C	96.2	**80.2**	**55.3**	57.8	**72.5**	**87.5**	92.0	**96.2**	78.9	92.8	52.9	**74.5**	**84.7**	**72.1**	**78.6**	**78.2**

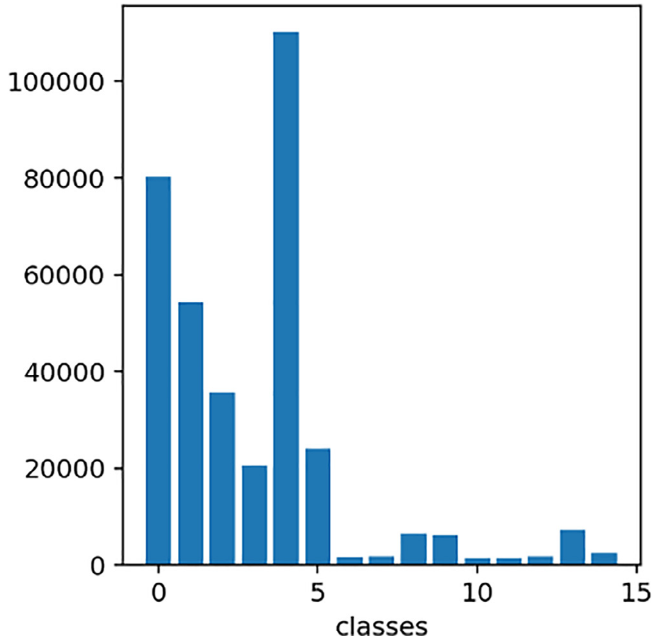


Fig. 6. Number of instances per category on all 1024 × 1024 tiles of DOTA dataset, 0–14 indicates SV, LV, PL, ST, SH, HA, GTF, SBF, TC, SP, BD, RA, BC, BR, HC respectively.

depicts detection examples of our proposed method in image tiles of the DOTA dataset. Fig. 8 shows detection examples of our proposed method in large scenarios of the DOTA dataset. The experimental results in two images prove the effectiveness and practical value of the proposed method in remote sensing object detection task again, which will be put into the engineering application of remote sensing object detection in the future.

4.6. Incorporations with classical algorithms on DOTA and MS COCO

It has been proved that MIOU can effectively improve the performance of YOLO v4 in remote sensing object detection tasks, but whether it can be applied to the mainstream object detection networks is still a question worth exploring. In order to verify the universality and robustness of the MIOU loss, this paper selects representative two-stage networks Faster R-CNN and Mask R-CNN, and one-stage network RetinaNet to further test the performance of MIOU on the remote sensing object detection dataset DOTA and natural image object detection dataset MS COCO.

According to Table 5, we can see that for Faster R-CNN, when the MIOU-E loss is used to replace the original loss function, mAP_{coco} and mAP_{50} are increased by 1.0% and 2.5% respectively. Compared with CIOU and EIOU, MIOU-C and MIOU-E have achieved some advancements, which proves that the proposed method can promote the improvement of remote sensing object detection on Faster R-CNN without increasing calculation cost. Mask R-CNN also shows a similar trend, which proves that the proposed MIOU loss function can be effectively applied to the general two-stage object detection task. For the one-stage network RetinaNet, mAP_{coco}, mAP_{50}, and mAP_{75} increased by 2.2%, 1.3%, and 3% respectively, realizing greater accuracy improvement. It shows that the MIOU loss function has a better effect on the one-stage object detection network in the remote sensing tasks and confirms again that MIOU has an obvious effect on the optimization of the regression process of bounding boxes.

In addition, the relative improvements brought by the MIOU in Table 4 and Table 5 are different, which might be caused by different evaluation rules. The size of a remote sensing image is much larger than that of a natural image. For example, the number of some image pixels in DOTA dataset are 13,000×13,000, while the number of some image pixels in MS COCO dataset are 400×600. Therefore, remote sensing images are generally segmented into tiles for training and inference. The inference results on the tiles will be restored to the original large-scale image. When a large-scale image is segmented into tiles, many truncated targets will be generated, and the inference results of these targets in tiles are considered to be the correct detection results. When the tiles are restored to the original large-scale image, these truncated targets will be regarded as wrong detection results. The accuracy statistics of Table 5 are based on the statistical results of tiles, and the accuracy statistics of Table 4 are based on the statistical

Table 4
Comparison with State of the Art one-stage methods. BL v1 means CIOU loss-based YOLO v4 l. Ours v1 means MIOU loss-based YOLO v4 l. BL v2 means CIOU loss-based YOLO v4 x. Ours v2 means MIOU loss-based YOLO v4 x.

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP_{voc}	FPS
RetinaNet [8]	ResNet101	78.2	53.4	26.4	42.3	63.6	52.6	73.2	87.2	44.6	58.0	18.0	51.0	43.4	56.6	7.4	50.4	14
YOLO v3 [19]	DarkNet53	79.0	77.1	33.9	**68.1**	52.8	52.2	49.8	89.9	74.8	59.2	55.5	49.0	61.5	55.9	41.7	60.0	13
DSSD [42]	Hour104	**91.1**	71.8	54.6	66.4	79.0	77.2	87.5	87.6	52.1	69.7	38.0	**72.6**	75.4	59.4	28.9	67.4	9
DYOLO [43]	DarkNet19	86.6	71.4	54.6	52.5	79.2	80.6	87.8	82.2	54.1	75.0	51.0	69.2	66.4	59.2	51.3	68.1	17
RFBNet [44]	VGG16	88.0	**84.5**	38.5	66.3	72.1	73.3	81.7	90.6	84.5	64.7	55.9	56.4	70.0	75.2	60.6	70.8	14
FMSSD [45]	VGG16	89.1	81.5	48.2	67.9	69.2	73.6	76.9	90.7	82.7	73.3	52.7	67.5	72.4	80.6	60.2	72.4	16
EFR [46]	VGG16	88.4	83.9	45.8	67.2	76.8	77.2	85.4	90.8	**85.6**	75.8	54.6	60.8	71.4	77.9	60.9	73.5	11
ASSD [29]	VGG16	89.3	82.0	52.1	62.6	79.8	81.8	**88.2**	**90.9**	82.0	82.9	**59.8**	65.3	77.1	**82.0**	64.5	76.0	21
BL v1	CSPDarkNet53	88.7	83.3	**58.0**	59.9	80.5	84.7	87.9	90.5	85.6	86.3	51.0	63.2	81.4	80.3	58.3	75.9	55
Ours v1	CSPDarkNet53	88.9	82.5	57.1	60.0	81.2	85.2	88.0	90.5	84.2	**87.0**	50.2	63.8	81.2	80.6	64.1	76.3	**55**
BL v2	CSPDarkNet53	88.6	83.1	57.6	59.1	80.8	84.9	88.0	90.3	84.3	86.8	50.5	65.7	**81.5**	81.0	64.6	76.4	35
Ours v2	CSPDarkNet53	88.7	84.2	56.4	61.3	**81.3**	**85.5**	88.1	90.3	84.5	86.6	52.6	63.0	81.2	80.8	**65.7**	**76.7**	35

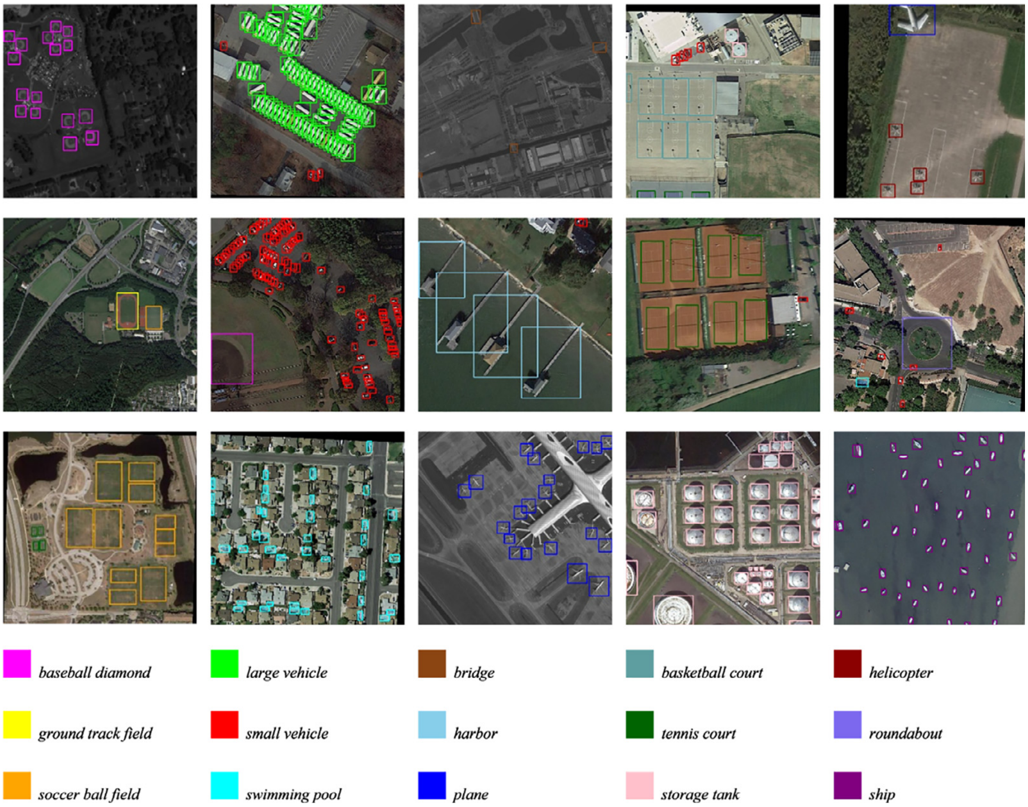


Fig. 7. Detection examples of our proposed method in image tiles on DOTA dataset.

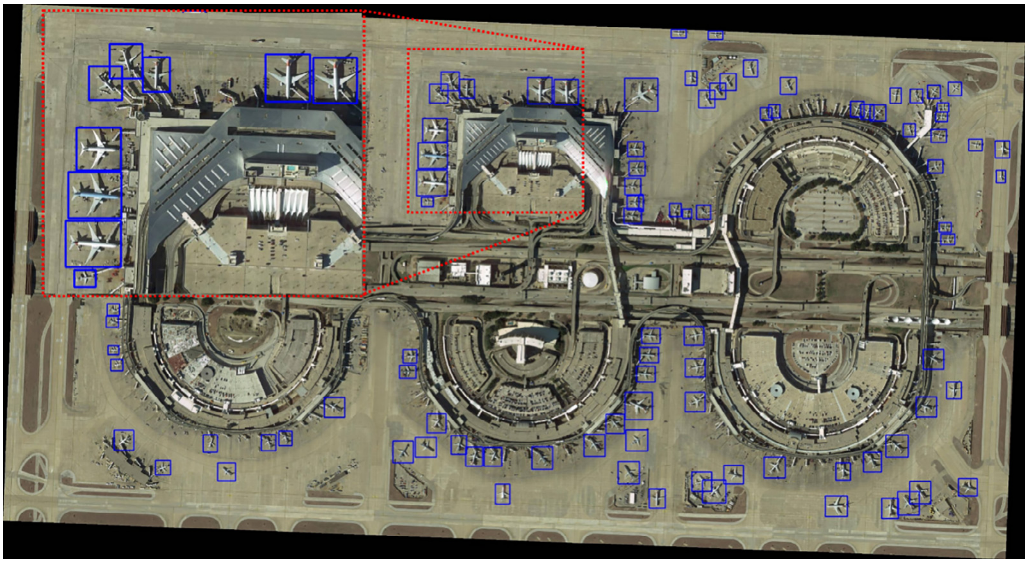


Fig. 8. Detection examples of our proposed method in large scenarios on DOTA dataset.

results of the original image, so the accuracy improvement in Table 4 is not as much as that in Table 5.

Table 6 indicates the MIOU loss incorporated into classical algorithms Faster R-CNN, Mask R-CNN, and RetinaNet on MS COCO. In Faster R-CNN, MIOU-C is 1.1% higher than CIOU's mAP$_{coco}$, MIOU-E is 0.8% larger than EIOU's mAP$_{coco}$, and MIOU-E is 1.2% larger than the baseline's mAP$_{coco}$. This sufficiently proves the effectiveness of MIOU in Faster R-CNN. We also found similar trends in Mask R-CNN and RetinaNet, which proves that the MIOU loss can effectively improve the detection ability of deep learning networks in natural scenes. In summary, we confirmed that the MIOU loss has strong robustness in remote sensing object detection tasks as well as natural object detection tasks, and it can be easily incorporated into mainstream object detection networks to improve their performance.

5. Conclusion

This paper summarizes and explains the shortcomings of the existing regression loss functions in deep learning object detection

Table 5
MIOU loss incorporated into Classical Algorithms on DOTA. Baseline means the original loss.

Method	Backbone	Loss	mAP$_{coco}$	mAP$_{50}$	mAP$_{75}$	mAP$_{s}$	mAP$_{m}$	mAP$_{l}$
Faster R-CNN	ResNet-50-FPN	Baseline	33.9	53.4	37.3	16.3	32.5	45.7
		CIOU	34.3	54.9	37.4	17.2	32.8	44.6
		MIOU-C	34.7	55.7	37.4	16.8	33.7	**46.5**
		EIOU	34.2	54.8	37.4	17.0	32.7	45.2
		MIOU-E	**34.9**	**55.9**	**37.8**	**17.2**	**34.1**	45.2
Mask R-CNN	ResNet-50-FPN	Baseline	33.9	53.6	36.7	16.4	33.2	45.3
		CIOU	34.2	54.3	37.2	15.8	33.7	46.3
		MIOU-C	34.8	55.6	37.7	**17.4**	33.6	46.7
		EIOU	34.4	55.1	37.8	16.0	33.6	45.8
		MIOU-E	**36.1**	**57.8**	**38.9**	17.3	**35.8**	**47.5**
RetinaNet	ResNet-50-FPN	Baseline	34.4	56.9	35.8	14.6	35.1	47.3
		CIOU	34.5	56.9	36.2	15.5	35.7	46.4
		MIOU-C	**36.6**	**58.2**	**38.8**	**17.1**	37.3	**48.8**
		EIOU	35.1	57.5	37.3	15.1	36.6	46.9
		MIOU-E	35.5	58.1	37.1	15.7	**37.3**	47.7

Table 6
MIOU loss incorporated into Classical Algorithms on MS COCO. Baseline means the original loss.

Method	Backbone	Loss	mAP$_{coco}$	mAP$_{50}$	mAP$_{75}$	mAP$_{s}$	mAP$_{m}$	mAP$_{l}$
Faster R-CNN	ResNet-50-FPN	Baseline	24.9	37.4	28.3	7.6	24.8	43.2
		CIOU	24.2	36.8	26.6	8.2	23.6	40.8
		MIOU-C	25.3	37.4	28.5	8.0	24.7	43.5
		EIOU	25.3	37.9	27.9	9.0	24.5	43.5
		MIOU-E	**26.1**	**38.8**	**29.2**	**9.0**	**26.3**	**44.1**
Mask R-CNN	ResNet-50-FPN	Baseline	25.3	37.4	28.4	8.2	24.9	44.4
		CIOU	25.4	37.4	28.4	8.0	24.7	44.0
		MIOU-C	26.6	**39.4**	29.7	8.8	**26.8**	44.9
		EIOU	25.7	38.2	28.6	8.8	25.7	42.5
		MIOU-E	**26.8**	39.2	**29.9**	**9.0**	26.3	**45.8**
RetinaNet	ResNet-50-FPN	Baseline	33.7	**53.1**	35.8	18.1	37.6	44.2
		CIOU	34.3	52.9	36.4	18.3	38.1	44.7
		MIOU-C	34.5	52.1	**36.9**	**18.9**	37.8	**45.5**
		EIOU	34.2	52.9	36.3	18.0	37.9	45.1
		MIOU-E	**34.7**	53.0	36.9	18.8	**38.6**	45.2

task and puts forward the MIOU loss with faster regression speed and higher accuracy. By setting up simulation experiments, we prove that the MIOU loss has the advantages of regression speed and accuracy compared with other IOU-based losses. In order to explore the effect of the MIOU loss embedded in CNN in real remote sensing datasets, we propose MIOU loss-based YOLOv4, which proves the effectiveness of MIOU loss and obtains high-precision and high-speed remote sensing object detection results. We also incorporate the MIOU loss into many classical object detection networks on the natural image object detection dataset MS COCO to verify its universality. The experimental results show that the MIOU loss is a general object detection regression loss function, which can be easily applied to most object detection networks to improve performance. However, our current research mainly focuses on two-dimensional horizontal bounding boxes. In the future, we will pay attention to the regression of rotating bounding boxes and three-dimensional bounding boxes, and upgrade the existing version.

CRediT authorship contribution statement

Yanyun Shen: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Feizhao Zhang:** Conceptualization, Formal analysis, Software. **Di Liu:** Writing – review & editing. **Weihua Pu:** Writing – review & editing. **Qingling Zhang:** Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2017YFB0504204), the Shenzhen Science and Technology Innovation Project (No. JSGG20191129145212206), and supported in part by the National Natural Science Foundation of China (General Program, No. 42071351), the National Key Research and Development Program of China (No. 2020YFA0608501), and the Western Talents (No. 2018XBYJRC004)

References

[1] M. Everingham, L. Van Gool, C.K.I. Williams, et al., The Pascal visual object classes (VOC) challenge, International journal of computer vision. 88 (2) (2010) 303–338.

[2] T Y Lin, M Maire, S Belongie, et al., Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[3] R. Girshick, J. Donahue, T. Darrell et al., Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, pp. 580-587.

[4] J. Redmon, S. Divvala, R. Girshick et al., You only look once: Unified, real-time object detection, Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 779-788.

[5] G.-S. Xia, X. Bai, J. Ding et al., DOTA: A large-scale dataset for object detection in aerial images, Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 3974-3983.

[6] K. Li, G. Wan, G. Cheng, et al., Object detection in optical remote sensing images: A survey and a new benchmark, ISPRS J. Photogramm. Remote Sens. 159 (2020) 296–307.

[7] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105.

[8] T.-Y. Lin, P. Goyal, R. Girshick et al., Focal loss for dense object detection, Proceedings of the IEEE international conference on computer vision. 2017, pp. 2980-2988.

[9] R. Girshick, Fast r-cnn, Proceedings of the IEEE international conference on computer vision. 2015, pp. 1440-1448.

[10] K. He, G. Gkioxari, P. Dollár et al., Mask r-cnn, Proceedings of the IEEE international conference on computer vision. 2017, pp. 2961-2969.

[11] S. Ren, K. He, R. Girshick, et al., Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015) 91–99.

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

[13] H. Rezatofighi, N. Tsoi, J. Gwak et al., Generalized intersection over union: A metric and a loss for bounding box regression, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, pp. 658-666.

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

[15] Y.-F. Zhang, W. Ren, Z. Zhang *et al.*, Focal and efficient IOU loss for accurate bounding box regression, *arXiv preprint arXiv:2101.08158*, 2021.

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

[17] P. Sermanet, D. Eigen, X. Zhang *et al.*, Overfeat: Integrated recognition, localization and detection using convolutional networks, *arXiv preprint arXiv:1312.6229*, 2013.

[18] J. Redmon, and A. Farhadi, YOLO9000: better, faster, stronger, Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 7263-7271.

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

[20] W. Liu, D. Anguelov, D. Erhan, et al., Ssd: Single shot multibox detector, European conference on computer vision, Springer, 2016, pp. 21–37.

[21] T.-Y. Lin, P. Dollár, R. Girshick et al., Feature pyramid networks for object detection, Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 2117-2125.

[22] S. Liu, L. Qi, H. Qin et al., ''Path aggregation network for instance segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 8759-8768.

[23] K. He, X. Zhang, S. Ren, et al., Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.

[24] J. Dai, Y. Li, K. He et al., R-fcn: Object detection via region-based fully convolutional networks, Advances in neural information processing systems. 2016, pp. 379-387.

[25] A. Van Etten, You only look twice: Rapid multi-scale object detection in satellite imagery, *arXiv preprint arXiv:1805.09512*, 2018.

[26] C. Li, R. Cong, C. Guo, et al., A parallel down-up fusion network for salient object detection in optical remote sensing images, Neurocomputing 415 (2020) 411–420.

[27] L. Chen, C. Liu, F. Chang, et al., Adaptive multi-level feature fusion and attention-based network for arbitrary-oriented object detection in remote sensing imagery, Neurocomputing 451 (2021) 67–80.

[28] X. Yang, J. Yang, J. Yan, et al., Scrdet: Towards more robust detection for small, cluttered and rotated objects, Proceedings of the IEEE/CVF International Conference on Computer Vision (2019) 8232–8241.

[29] T. Xu, X. Sun, W. Diao, et al., ASSD: feature aligned single-shot detection for multiscale objects in aerial imagery, IEEE Trans. Geosci. Remote Sens. (2021).

[30] H. Zhang, H. Chang, B. Ma, et al., Dynamic R-CNN: Towards high quality object detection via dynamic training, European conference on computer vision, Springer, 2020, pp. 260–275.

[31] J. Pang, K. Chen, J. Shi, et al., Libra r-cnn: Towards balanced learning for object detection, European conference on computer vision, Springer, 2020, pp. 821–830.

[32] S. Zhang, C. Chi, Y. Yao et al., Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, pp. 9759-9768.

[33] D. Zhou, J. Fang, X. Song, et al., Iou loss for 2d/3d object detection, 2019 International Conference on 3D Vision (3DV), IEEE, 2019, pp. 85–94.

[34] Q. Meng, W. Wang, T. Zhou, et al., Towards a weakly supervised framework for 3d point cloud object detection and annotation, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[35] J. Yin, J. Shen, X. Gao, et al., Graph neural network and spatiotemporal transformer attention for 3D video object detection from point clouds, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[36] X. Dong, J. Shen, W. Wang, et al., Dynamical hyperparameter optimization via deep reinforcement learning in tracking, IEEE Trans. Pattern Anal. Mach. Intell. 43 (5) (2019) 1515–1529.

[37] D. Misra, Mish: A self regularized non-monotonic neural activation function, *arXiv preprint arXiv:1908.08681,* vol. 4, pp. 2, 2019.

[38] K. He, X. Zhang, S. Ren et al., Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 770-778.

[39] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu et al., CSPNet: A new backbone that can enhance learning capability of CNN, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020, pp. 390-391.

[40] G. Ghiasi, T.-Y. Lin, and Q. V. Le, Dropblock: A regularization method for convolutional networks, *arXiv preprint arXiv:1810.12890*, 2018.

[41] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, ''Scaled-yolov4: Scaling cross stage partial network, Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. 2021, pp. 13029-13038.

[42] C.-Y. Fu, W. Liu, A. Ranga *et al.*, Dssd: Deconvolutional single shot detector, *arXiv preprint arXiv:1701.06659*, 2017.

[43] O. Acatay, L. Sommer, A. Schumann, et al., Comprehensive evaluation of deep learning based detection methods for vehicle detection in aerial imagery, IEEE, 2018, pp. 1–6.

[44] S. Liu, and D. Huang, Receptive field block net for accurate and fast object detection, Proceedings of the European conference on computer vision (ECCV). 2018, pp. 385-400.

[45] P. Wang, X. Sun, W. Diao, et al., FMSSD: feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 58 (5) (2020) 3377–3390.

[46] K. Fu, Z. Chen, Y. Zhang *et al.*, Enhanced feature representation in detection for optical remote sensing images, *Remote Sensing,* 11(18) (2019).

Yanyun Shen received the M.E. degree in surveying and mapping engineering from the School of Remote Sensing and Information Engineering, WuHan University, Wuhan, China, in 2019. He is currently pursuing the Ph. D. degree in Aerospace Science and Technology at the School of Aeronautics and Astronautics, Sun Yat-sen University, Shenzhen Campus, Shenzhen, China. His research interests include computer vision, remote sensing image processing and object detection.

Feizhao Zhang received the B.E. degree in automotive engineering from Chongqing University, Chongqing, China, in 2019. He is pursuing the M.E. degree in Aerospace Science and Technology with Sun Yat-sen University, Shenzhen Campus, Shenzhen, China. His research involves computer vision, remote sensing image processing, including object detection and blind super-resolution.

Di Liu received the M.E. degree in surveying and mapping engineering from the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China, in 2019. He is currently pursuing the Ph.D. degree in Aerospace Science and Technology at the School of Aeronautics and Astronautics, Sun Yat-sen University, Shenzhen Campus, Shenzhen, China. His research interests include Nightlight Remote Sensing, Synthetic Aperture Radar (SAR) and Artificial intelligence(AI) Applications of Multi-source Remote Sensing.

Weihua Pu, senior engineer, obtained a M.E. degree from Shenzhen University, Shenzhen, China. He is currently the product manager of Shenzhen Aerospace Dongfanghong Satellite Co., Ltd. and the director of Guangdong intelligent satellite engineering research center. His research interests include satellite integrated electronic system design and on-board computer design.

Qingling Zhang received the Ph.D. degree in geography from Boston University, Boston, MA, USA, in 2009. From 2008 to 2014, he was a Research Scientist with the School of Forestry and Environmental Studies, Yale University, New Haven, CT, USA. From 2014 to 2018, he has been a Professor with the Shenzhen Institutes of Advanced Technology, as well as the Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences. Since 2018, he is working as a Full Professor at the School of Aeronautics and Astronautics, Sun Yat-sen University, Shenzhen Campus, Shenzhen, China. He is an expert in urbanization, land cover land use change, and nighttime light remote sensing. His research interests include crop monitoring and yield forecasting, global environmental change, urban remote sensing, land cover land use change, food security, GIS and spatial analysis, and remote sensing data mining based on cloud computing. Dr. Zhang was a recipient of the One Hundred Person Project of the Chinese Academy of Sciences Award for Excellence in 2015.