Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 3, 2023

Object Detection of UAV Aerial Image based on YOLOv8

Chen Liu, Fanrun Meng, Zhiren Zhu, Liming Zhou

College of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, China

Abstract: With the development of technology, unmanned aerial vehicles (UAVs) have shed their military uses and gradually expanded to civilian and commercial fields. With the development of drone technology, object detection technology based on deep learning has become an important research topic in the field of drone applications. Apply object detection technology to unmanned aerial vehicles to achieve object detection and recognition of ground scenes from an aerial perspective. However, in aerial images taken by drones, the detection objects are mostly small targets, and the target scale changes greatly due to the influence of aerial perspective; The image background is complex, and the target object is easily occluded. It has brought many challenges to the target detection of unmanned aerial vehicles. Conventional object detection algorithms cannot guarantee detection accuracy when applied to drones, and optimizing the target detection performance of drones has become an important research topic in the field of drone applications. We improve the WIoUv3 loss function on the basis of YOLOv8s to reduce regression localization loss during training and improve the regression accuracy of the model. The experimental results indicate that the improved model mAP@0.5 It increased by 0.6 percentage points to 40.7%.

Keywords: YOLOv8; UAV; WIoU.

1. Introduction

With the development of UAV technology, object detection technology based on deep learning has become the application field of UAV. The important research content. The traditional target detection algorithm has low precision, poor detection efficiency, and insufficient generalization and robustness. With the rise of convolutional neural networks, object detection algorithms based on deep learning have gradually replaced traditional object detection algorithms. Currently, target detection algorithms based on deep learning are mainly divided into two-stage model represented by Faster-RCNN [1] and single-stage model represented by YOLO series and SSD [2]. Since YOLO [3] algorithm was first proposed, it has been widely used in target detection tasks in various scenarios due to its requirement of fast detection speed. However, the detection accuracy of small targets needs to be improved. The development of UAV technology is becoming more and more mature, and more and more scholars pay attention to the target detection of UAV aerial images to further improve the visual perception ability of UAV. Let drones play an irreplaceable role in rescue and disaster relief, police reconnaissance, agricultural development, traffic monitoring and other fields. We have designed an improved unmanned aerial vehicle aerial photography small target detection algorithm based on YOLOv8s, replacing the original model's CIoU [4] with WIoUv3[5], reducing regression localization losses during training and improving the regression accuracy of the model.

2. YOLOv8 Object Detection Algorithm

The network structure of YOLOv8 is mainly composed of four parts: Input layer, Backbone layer, Neck layer and output detection Head layer. The network structure of YOLOv8 is shown as follows:

2.1. Input Layer

The input layer is mainly responsible for data preprocessing of the input images. The input layer of YOLOv8 inherits the Mosaic data enhancement strategy of YOLOv4[6] and YOLOv5 algorithms. The advantages of using Mosaic data enhancement strategy are as follows: on the one hand, it can enrich the data set by randomly selecting 4 images for scaling and then randomly distributing and splicing, which greatly enrichis the detection data set and effectively improves the robustness of the algorithm; On the other hand, it can reduce the GPU storage space occupied in the training process. However, if the entire training process is turned on, the final training result will be reduced, so in YOLOv8, Mosaic data enhancement is turned off for the last 10 epochs to achieve better training results.

2.2. Backbone Layer

YOLOv8 uses CSPDaknet53 as the backbone network to downsample the input feature images five times and obtain five feature maps of different proportions in turn. The backbone network is composed of CBS module, C2f module and SPPF module.

(1) CBS module: CBS module includes convolution operation (Conv), batch normalization (BN) and activation function (SiLU), which is improved from CBL module. Compared with the LeakyRelu activation function in the CBL module, the smooth and non-monotonic characteristics of the SiLU function can make the model achieve better results during deep learning training.

(2) C2f module: YOLOv8 proposed a new module C2f module to replace C3 module in the original YOLOv5 network architecture. C2f combines the idea of C3 module and ELAN module, and adopts gradient shunt connection to optimize the module structure, which enriches the information flow of the feature extraction network while maintaining lightweight, and can effectively improve the overall detection performance of the algorithm. The structure of C2f is shown as follows:

(3) SPPF module: The Spatial Pyramid Pooling Fast (SPPF) structure used in YOLOv8 is optimized on the basis of the original SPP [7] structure. The SPPF structure uses three 5×5 convolution nuclei to replace the 13×13, 9×9, 5×5 and 1×1

convolution nuclei in the original SPP structure. By using multiple small convolution nuclei in series, the calculation amount of the model is reduced, the detection rate is improved, and the detection accuracy is still close to that of the original structure. The structure diagram of SPPF module is shown as follows:

2.3. Neck Layer

The Neck layer of YOLOv8 still adopts the PANet structure, which is composed of Feature Pyramid Network (FPN)[8] and Path Aggregation Network (PAN)[9].A top-down and bottom-up network structure is constructed, and the shallow location information and deep semantic information are effectively complemented by feature fusion, and the integrity of feature information is preserved.

2.4. Head Layer

(1) Decoupling detection head:YOLOv8's detection layer uses the decoupage detection head structure, uses two separate branches for object classification and predictive bounding box regression, and uses different loss functions for both types of tasks. On the one hand, binary cross entropy loss (BCE) is used in classification tasks; On the other hand, in the predictive bounding box regression task, Distribution Focal Loss (DFL) and CIoU are used. The decoupled head structure can improve the detection accuracy and speed up the model convergence.

(2) Anchor Free: The traditional Anchor Based method relies on the hand-designed anchor frame. The size and aspect ratio of the anchor frame should be as close to the real target as possible, and its universality is poor for different data sets with big difference in target size. A large number of anchor frames generated during detection will not only increase the calculation amount, but also reduce the training and reasoning speed, and a large number of anchor frames will become negative samples because they do not reach a certain IoU with the real frame, which will lead to the problem of positive and negative sample imbalance. The Anchor Free method adopted by YOLOv8 can cope with multi-scale target changes and has good generalization ability. Because it is not affected by the number and position of anchor frames, it has better robustness when dealing with occluded and dense targets.

3. Improving the Loss Function

The Intersection over Union (IoU) is used to weigh the degree of match between the predicted border and the real marked border, that is, the ratio of the intersection between the predicted border and the real marked border and the union between them. The ideal case is complete overlap, when IoU=1.

CIoU is used as coordinate loss function in YOLOv8 network model. CIoU considers three important geometric factors, namely overlapping area, center point distance and aspect ratio. Given prediction frame B and real frame B^{gt}, CIoU loss function is defined as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{1}$$

Where b and b^{gt} represent the center point of B and B^{gt} respectively, $\rho^2(\,)$ represents the Euclidean distance between them, c is the diagonal length of the minimum external frame of the prediction box and the real box, α is the positive balance parameter, and v is the aspect ratio of the prediction box and the real box. Since the last parameter v in L_{CIoU} only reflects the difference in aspect ratio, and does not reflect the actual difference in width and height respectively and their confidence levels, it will hinder the optimization similarity of the model.

EIoU [10] divides the loss function into three parts: IoU loss L_{IoU}, distance loss L_{dis} and aspect ratio loss L_{asp}. Considering the overlap area, center point distance and true difference of length, width and side length, the fuzzy definition of aspect ratio is solved based on CIoU. The loss term of the aspect ratio is divided into the difference between the width and height of the prediction frame and the width and height of the minimum external frame, and the EIoU loss function is defined as follows:

$$L_{EIoU} = L_{IoU} + L_{dis} + L_{asp} \tag{2}$$

$$= 1 - IoU + \frac{p^2(b, b^{gt})}{(w^c)^2 + (h^c)^2} + \frac{p^2(w, w^{gt})}{(w^c)^2} + \frac{p^2(h, h^{gt})}{(h^c)^2},$$

Where, h^w and h^c are the width and height of the minimum external frame for the prediction and target frames.

SIoU[11] first introduced the vector Angle between bounding box regressions as a penalty factor. First, based on the magnitude of the Angle, between the prediction box and the actual box, the prediction box quickly moves toward the nearest axis and then returns to the real box. The degree of freedom of regression can be effectively reduced and the convergence speed of the model can be accelerated.

Several mainstream loss functions described above all adopt static focusing mechanism. WIoU not only takes aspect ratio, centroid distance and overlapping area into account, but also introduces dynamic non-monotonic focusing mechanism to calculate IoU loss in predicted category loss based on dynamic method, and uses gradient gain to evaluate anchor frame quality. It can improve the performance of the algorithm, prevent slow convergence, improve convergence accuracy, and enhance the generalization ability of the model. There are three versions of WIoU: WIoU v1, WIoU v2, and WIoU v3. WIoU v1 uses an attention-mechanism-based predictive frame loss design and introduces distance metrics. When the prediction box overlaps with the real box within a certain range, the penalty of geometric measures is reduced to obtain better generalization ability. WIoU v1 is calculated as follows:

$$L_{WIoUv\,1} = R_{WIoU} L_{IoU} \tag{3}$$

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{\left(W_g^2 + H_g^2\right)^*}\right) \tag{4}$$

$$L_{IoU} = 1 - \frac{W_i H_i}{wh + w_{gt} h_{gt} - W_i H_i} \tag{5}$$

Among them, x and y represent the horizontal and vertical coordinates of the center point of the prediction box, w and h represent the width and height of the prediction box, xgt and ygt represent the horizontal and vertical coordinates of the center point of the real box for x and y, respectively, Wgt and Hgt represent the width and height of the real box. Wg and Hg are used to represent the width and height of the

minimum enclosed box formed by the predicted box and the real box.

Based on v1 version, WIoUv2 constructs $L_{IoU}^{\gamma^*}$ monotone focusing coefficient, which can effectively reduce the weight of simple samples in the loss value, make the model pay more attention to difficult samples, and improve the classification performance. However, in the course of training, the gradient gain will decrease with the decrease of L_{IoU}, resulting in the deterioration of the convergence of the model at the later stage of training. In order to solve this problem, the average value of L_{IoU} is introduced to normalize L_{IoU}. The formula for calculating WIoU v2 is as follows:

$$L_{WIoUv2} = \left(\frac{L_{IoU}^{*}}{L_{IoU}} \right)^{\gamma} L_{WIoUv1} \tag{6}$$

WIoUv3 builds β non-monotone focusing coefficient r based on anomaly degree A, and through r, dynamically optimizes the contribution of high-quality and low-quality samples to the gradient, reduces the harmful gradient generated by low-quality samples, thereby speeding up the convergence rate of the model and improving the overall performance of the model. The calculation formula for WIoU v3 is shown as follows:

$$L_{WIoUv3} = rL_{WIoUv1} \tag{7}$$

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}}, \quad \beta = \frac{L_{IoU}^{*}}{L_{IoU}} \in [0, +\infty)$$

4. Experimental Design and Analysis of Results

4.1. Experimental Environment and Parameter Settings

The relevant configuration parameters of the experimental platform environment construction are shown in the table 1:

Table 1. Experimental environment configuration

name	configuration
CPU	Inter(R) Core (TM) i7-11700 @ 2.50GHz
GPU	NVIDIA RTX A4000
Stystem	Windows 10 64-bit
CUDA	11.7.1
Programming language	Python 3.8
Deep learning framework	Pytorch 1.13.1

4.2. Training Parameter Settings

Table 2. Training parameter Settings

Training parameter	Parameter value
optimizer	**SGD**
Learning rate	**0.01**
Momentum parameter	**0.937**
Learning rate attenuation	**0.0005**
Input size	**640×640**
Batch size	**4**
Epoch	**300**

The configuration of training parameters in the experiment is shown in the table 2, and the training parameters are consistent in all experimental processes.

4.3. Experimental Data Set

The VisDrone2019 dataset was collected by AISKYEYE team of Machine Learning and data Mining Laboratory of Tianjin University. Captured by a variety of drone cameras, it covers a wide range, including location (from 14 different cities thousands of kilometers apart in China), environment (urban and rural), objects (pedestrians, vehicles, bicycles, etc.), and density (sparse and crowded scenes). The VisDrone2019 dataset contains 10 categories of detection targets. The data set contains 6471 images of the training dataset, 548 images of the verification dataset, and 1580 images of the test dataset. The statistics of the number of labels in each category are shown in Figure1:

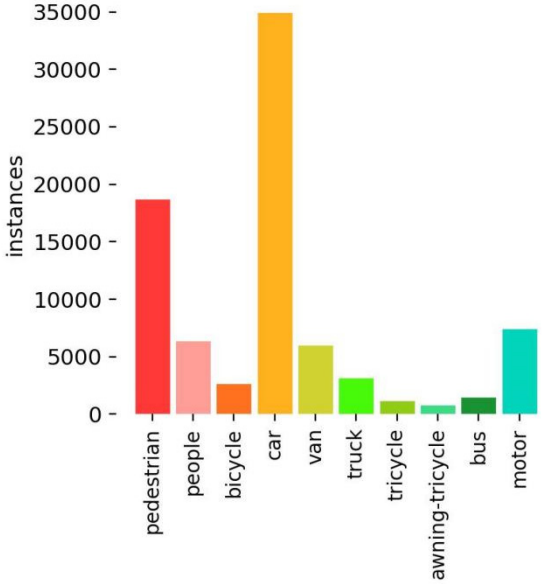


Figure 1. Label statistics of various categories in the VisDrone dataset

4.4. Evaluation Indicators

In our experiment, mean average precision (mAP), Precision (P) and Recall (R) were used to evaluate the detection performance of the model.

(1) mAP represents the average accuracy of all detection categories, calculated as follows:

$$mAP = \frac{1}{c} \sum_{i=1}^{c} AP_i \tag{8}$$

Where c represents the total number of detection categories, i represents the number of detection times, AP is the detection accuracy of a single category, and mAP is obtained by adding the average AP of detection accuracy of all categories. mAP@0.5 indicates the average accuracy when the IoU threshold is set to 0.5.

(2) P (Precision) represents the accuracy of the model detection target. The calculation formula is as follows:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{9}$$

Where TP represents the number of positive samples for correct classification, and FP represents the number of negative samples for false detection.

(3) R (Recall):

$$R = \frac{TP}{TP + FN} \times 100\% \quad (10)$$

FN indicates the number of positive samples missed.

4.5. Comparative Experiment of Different Loss Functions

In the experiment, EIoU, SIoU, WIoUv1, WIoUv2, WIoUv3 were selected to compare with CIOU in the original YOLOv8 model to verify the influence of different loss functions on the model performance. The experimental results are shown in the table 3:

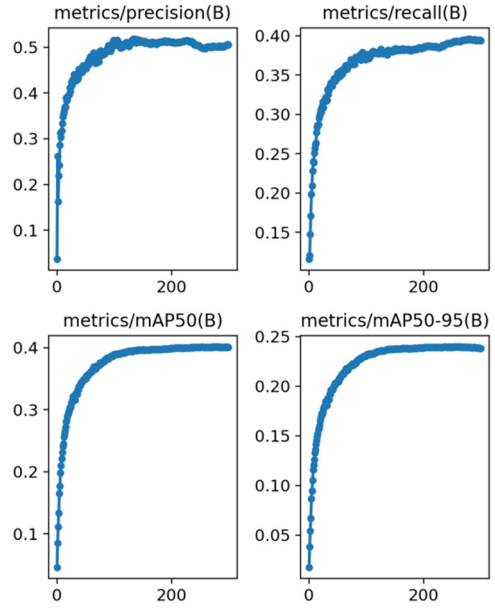
Table 3. Comparison of loss function

Loss function	mAP@0.5	P	R
CIoU	0.401	0.498	0.393
EIoU	0.401	0.518	0.391
SIoU	0.402	0.517	0.388
WIoUv1	0.396	0.504	0.386
WIoUv2	0.406	0.517	0.392
WIoUv3	0.407	0.512	0.393

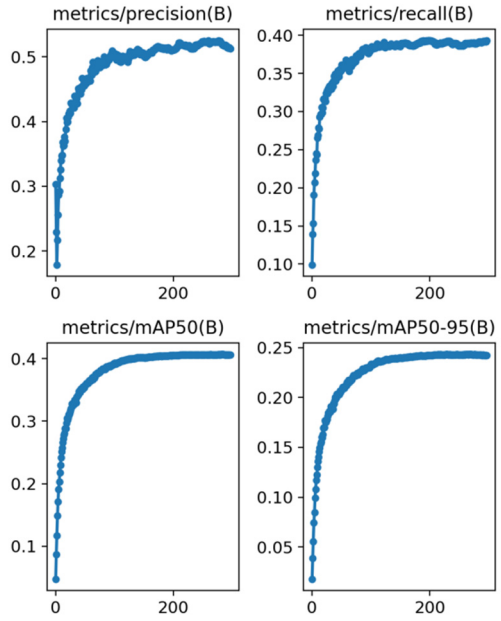
CIoU is the original loss function in YOLOv8, and mAP@0.5 in VisDrone2019 dataset is 0.401. Compared with EIoU, SIoU, WIoUv1 and WIoUv2, WIoUv3 has the most obvious improvement effect on model performance, with mAP@0.5 increasing by 0.6 percentage points. That's 40.7 percent. Table 4 shows the comparison of the precision of various classes after the improvement of WIoUv3 loss function. Figure 2 shows a comparison of the improved model detection performance.

Table 4. Comparison of improved WIoUv3 detection performance by category

Detection category	YOLOV8s	YOLOv8s+WIoU
pedestrian	0.405	0.501
people	0.319	0.38
bicycle	0.103	0.173
car	0.738	0.823
van	0.355	0.445
truck	0.283	0.354
tricycle	0.199	0.275
awning-tricycle	0.107	0.138
bus	0.43	0.542
motor	0.391	0.47
all	0.333	0.41



(a) YOLOv8s



(b) YOLOv8s-WIoUv3

Figure 2. Comparison of model detection performance before and after improvement

5. Conclusion

In this paper, an improved small target detection algorithm for UAV aerial photography based on YOLOv8s is designed. Replacing the loss function with WIoUv3 reduces the loss of regression positioning during training and improves the regression accuracy of the model. The experimental results show that the improved model mAP@0.5 increases by 0.6 percentage points, reaching 40.7%. In the follow-up study, targeted optimization of the network structure will be continued, and the number of parameters will be reduced to achieve the purpose of lightweight model, so as to better apply and deploy on edge devices with limited computing power.

References

[1] Ren S,He K,Girshick R , et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.

[2] Berg A C , Fu C Y , Szegedy C , et al. SSD: Single Shot MultiBox Detector:,10.1007/978-3-319-46448-0_2[P]. 2015.

[3] Redmon J, Divvala S ,Girshick R , et al. You Only Look Once: Unified, Real-Time Object Detection[J]. IEEE, 2016.

[4] Zheng Z ,Wang P , Ren D , et al. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation[J]. 2020.

[5] TONG Z, CHEN Y, XU Z, et al. Wise-Io U: Bounding box regression loss with dynamic focusing mechanism[J]. ar Xiv preprint ar Xiv:2301.10051, 2023

[6] Bochkovskiy A , Wang C Y , Liao H Y M .YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020. DOI: 10. 48550/arXiv.2004.10934. R. Girshick, "Fast r-cnn," Proceedings of the IEEE international conference on computer vision, vol. 12, pp. 1440-1448, 2015.

[7] Kaiming,He,Xiangyu,et al.Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015.DOI:10.1109/tpami.2015.2389824.

[8] Lin T Y , Dollar P , Girshick R , etal. Feature Pyramid Networks for Object Detection[J]. IEEE Computer Society, 2017.

[9] Wang W , Xie E , Song X , et al. Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network[J]. IEEE, 2019.

[10] Zhang Y F , Ren W , Zhang Z , et al. Focal and Efficient IoU Loss for Accurate Bounding Box Regression[J].2021.

[11] Gevorgyan Z. SIoU loss: More powerful learning for bounding box regression[J]. ar**v preprint ar**v:2205.12740, 2022.