Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-4, April 2025
DOI: 10.53469/jrse.2025.07(04).03

Research on Small Target Detection Algorithm for Aerial Images based on Improved YOLOv11

Dong Wang, Shixingyu Wang, Yuntian Jiang

Chongqing University of Technology, Chongqing 400054, China

Abstract: ***At present, UAV aerial photography has a good application prospect in agricultural production and disaster response. The application of drones can greatly improve work efficiency and decision-making accuracy. However, due to the inherent characteristics of drone aerial images, such as high image density, small target size, complex background, etc. In order to solve these problems, this paper proposes a small target detection algorithm for UAV aerial photography based on the improved YOLOv11n. Firstly, the FADC module was introduced into the backbone network to optimize the feature extraction process. Then, a small target detection layer was introduced into the algorithm to improve the detection performance of small targets in aerial images. Secondly, the scale sequence feature fusion network ASF-YOLO was used to replace the PANet network to improve the speed and accuracy of target detection. Then, Wise IoU is used to replace CIoU to speed up the network convergence speed and improve the regression accuracy. The algorithm was evaluated on the VisDrone-2019 dataset. Compared with YOLOV11n, the algorithm is improved by 5.7% and 4.3% in mAP@50 and mAP@0.5:0.95, respectively. Experiments show that compared with YOLOV11n, the performance of the algorithm on small targets is greatly improved.***

Keywords: Drone imagery, Small target detection, YOLOv11, Multi-scale feature fusion**.**

1. Introduction

With the rapid advancement of drone technology, drone aerial photography has been widely used in various fields such as natural disaster detection, traffic safety monitoring, search and rescue, and agricultural and forestry management. This technology greatly reduces labor costs, improves monitoring efficiency, and achieves better management and service. However, compared with the application scenarios of traditional object detection algorithms, the images captured by drones face many challenges such as a wide range of target scales, diverse angle changes, and complex backgrounds [1]. These factors significantly affect the accuracy and recall of target detection results.

At present, object detection algorithms can be divided into two categories: two-stage detection algorithms and one-stage detection algorithms. Two stage detection algorithms, including convolutional neural networks [2,3], CNN [4], and R-CNN [5], generate candidate boxes containing potential targets, and then use region classifiers to predict them. Single stage detection algorithms, such as SSD [6] and YOLO [7-13] series, directly classify and predict targets at each position on the feature map, thereby improving detection speed and practicality. Academically, the definition of small targets can be divided into two categories: relative scale and absolute scale. The former defines small targets based on their proportion in the entire image. Chen et al. [14] defined a small target as follows: when the ratio of the bounding box area to the image area is between 0.08% and 0.58%, it can be considered a small target. The latter defines small targets based on their absolute pixel size, defining them as targets with a resolution of less than 32 pixels on each side.

Based on the above definition, most targets in drone aerial images can be defined as small targets. However, small object detection is a highly challenging task. Lim et al. [15] proposed FA-SSD, which improves the detection of small targets by integrating feature information from F-SSD and A-SSD network structures. However, due to the two-stage detection algorithm used by FA-SSD, the detection speed is relatively slow. Liu et al. [16] improved the accuracy and generalization ability of the algorithm by introducing a shallow feature extraction network in the P1 layer and integrating FPN and PAN shallow features. Yang et al. [17] aimed to improve small object detection by enhancing feature information through the addition of scSE attention mechanism module and small object detection layer. However, there are still issues with missed and false detections of small targets. Zhang et al. [18] proposed a new object detection network dclinet, which utilizes dense pruning and local attention techniques to enhance the feature representation of small targets. They further incorporated bottleneck attention mechanism (BAM) into the network, greatly improving detection accuracy. Jin et al. [19] proposed a scale aware network that can accurately determine the scale of predefined anchor points. This network can effectively narrow down the scale search range, reduce the risk of overfitting, and improve the detection speed and accuracy of aerial images. Liu et al. [20] constructed the SPPCSPG module and introduced the shuffle attention (SA) mechanism into YOLOv5s, implementing a new lightweight network that greatly improves detection efficiency. The above algorithm model significantly improves the performance of small object detection. However, there is still a lot of room for improvement in terms of detection efficiency.

This article improves the YOLOv11n object detection algorithm by introducing the FADC module into the backbone network to extract richer small target feature information; Secondly, based on the SSFF module and the introduction of p2 detection head in the neck network, the feature fusion process is improved to enhance the multi-scale processing capability of the model; Finally, the Wise IoU mechanism is introduced to provide a gain allocation strategy, focusing on ordinary quality anchor boxes to improve the network's generalization ability.

2. Algorithm

This article addresses the issue of low accuracy in small target detection and improves YOLOv11n based on the characteristics of small targets. The algorithm network

structure is shown in Figure 1.

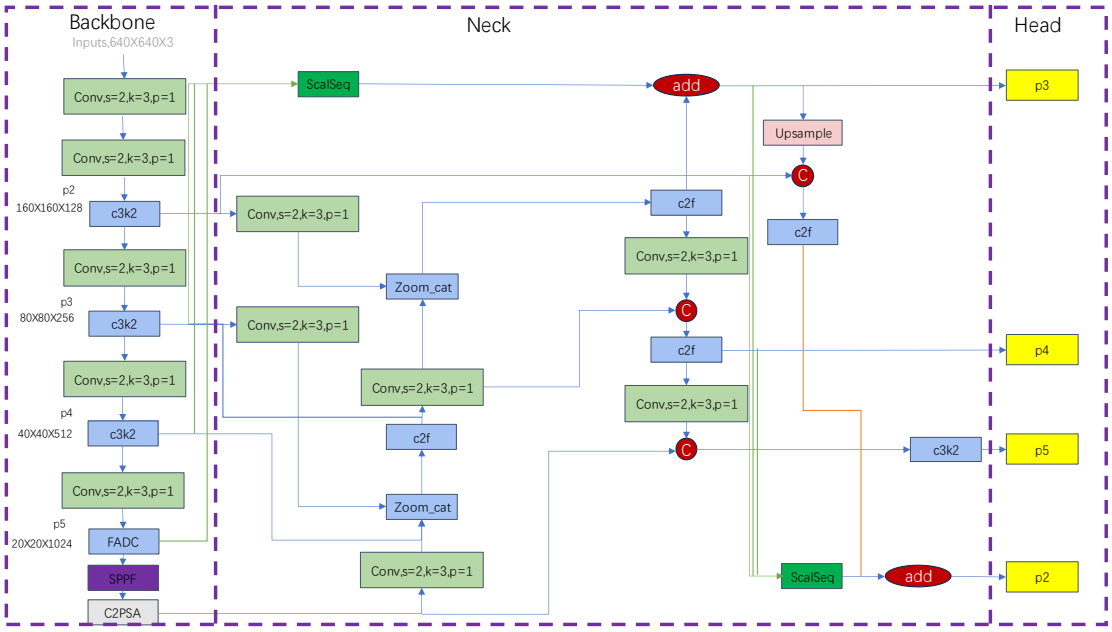


Figure 1: Algorithm network structure diagram in this article

The improvement focus of this article is on the backbone network and Neck section:

(1) Introducing FADC module into the backbone network helps the model better capture detailed information such as edges and textures of objects in the image while maintaining parameter quantity.

(2) This article adds a Conv module before the input of p2 and p3 feature maps. One function of this module is to enhance the fusion of feature information, and the second function is to adjust the parameters of the Conv module to achieve scale uniformity in the input of feature maps, which facilitates subsequent processing;

(3) In Neck, C2f+Conv module is used to replace the CSP module in ASF algorithm. The main reason is that the CSP module uses C3 structure (used by YOLOv5 algorithm). In this paper, C2f structure is used to replace C3 module, which can improve network performance and estimation accuracy. In addition, C2f module can be further improved in the future to increase network flexibility;

(4) Introducing the Wise IoU mechanism [21], the detection accuracy of the algorithm is further improved by adaptively adjusting the weight coefficients.

2.1 FADC Feature Extraction Module

A significant improvement in YOLOv11 on the backbone network is the introduction of the C3k2 block, which replaces the C2f module used in YOLOv8. The C3k2 block is a higher computational efficiency implementation for the bottleneck of Cross Phase Partial Processing (CSP). It uses two smaller convolutions instead of one larger convolution like YOLOv8. The "k2" in C3k2 represents a smaller convolution kernel size, which helps to achieve faster processing speed while maintaining performance. Meanwhile, the C3k2 module utilizes context aware mechanisms to analyze input feature maps, taking into account contextual information from multiple locations. This is achieved by utilizing convolutional layers, pooling layers, and other operations to capture different levels of information in the image. By introducing residual connections, the model's ability to understand the relationships between features has been improved, enhancing its feature extraction capability and detection accuracy. These operations expand the receptive domain of neural networks, enabling them to more comprehensively understand input images and thus improve performance. However, the fixed size convolution kernel used may contain interfering background features when extracting features from the target edge. This may lead to incorrect detection frames, which have a negative impact on accuracy and recall, thereby reducing overall detection accuracy. To address this issue, we propose the FADC (Frequency Adaptive Dilated Convolution) method. FADC adjusts the expansion rate based on the different frequencies present in the image. This method allows for the creation of extended convolutions with different receptive field sizes, enhancing the distinction between target edges and complex backgrounds. The structure of FADC is shown in Figure 2.

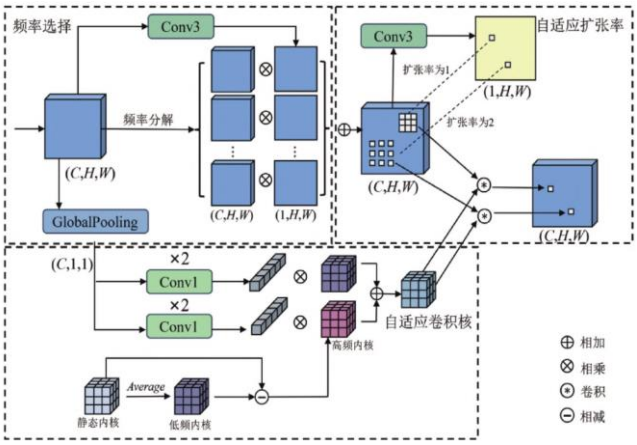


Figure 2: FADC Structure Diagram

2.2 Feature Fusion Module based on Attention Scale Sequence Fusion and Head Detection Head

The YOLOv11n model utilizes PANet as the neck for feature fusion, which is a pyramid shaped structure that fuses feature map information from bottom to top. Although this network can to some extent solve the problem of large scale differences in detecting images, it sometimes misses the feature information extraction of small targets because it mainly focuses on extracting features from deep layers, and is relatively weak in extracting shallow feature information.

ASF-YOLO (Attention Scale Sequence Fusion YOLO) is an improved object detection model based on the YOLO framework, which combines spatial and scale features to achieve accurate and fast cell instance segmentation. By introducing multiple innovative modules based on the YOLO segmentation framework, such as Scale Sequence Feature Fusion (SSFF) module, Triple Feature Encoder (TFE) module, and Channel and Position Attention Mechanism (CPAM), ASF-YOLO significantly improves the performance of the model in handling small, dense, and overlapping objects. These modules work together on different parts of the network, enhancing multi-scale information extraction capabilities, capturing detailed information of small objects, and focusing on information rich channels and features related to the spatial position of small objects, thereby improving the accuracy of detection and segmentation. As an improved object detection model based on the YOLO framework, it not only performs well in the field of cell instance segmentation, but also has technical characteristics such as multi-scale information extraction, feature fusion and detail enhancement, and attention mechanism, which make it highly compatible with the needs of small object detection in drone aerial photography. In the detection of small targets in drone aerial photography, there may be significant differences in the size and scale of the targets. For example, from small vehicles to large trucks, from distant pedestrians to nearby obstacles. To enhance the processing capability of the model for the above scenarios, this paper introduces the SSFF module and improves the feature fusion process in the neck network.

In order to solve the multi-scale problem of drone aerial images, existing literature adopts a feature pyramid structure for feature fusion, usually combining pyramid features through summation or concatenation [23]. However, various feature pyramid networks must effectively utilize the correlation between all pyramid feature maps. The SSFF scale sequence feature fusion module combines the high-dimensional information of good deep feature mapping with the detailed information of shallow feature mapping, where the size of the image changes during downsampling, but the scale invariant features remain unchanged. Scale space is constructed along the axis of an image, representing not only a single scale but also the range of scales that an object may have. Although blurry images may lose details, the structural features of the image can be preserved. The scaled image input into SSFF can be obtained from equations (1) and (2):

$$F_\sigma(w,h) = G_\sigma(w \cdot h) \times f(w,h) \quad (1)$$

$$G_\sigma(w \cdot h) = \frac{1}{2\pi\sigma^2} e^{-(w^2+h^2)/2\sigma^2} \quad (2)$$

Among them, f (w, h) represents the two-dimensional input image with width w and height h. F σ (w, h) is generated by applying a smoothing process through a series of convolutions using a two-dimensional Gaussian filter G σ (w, h). Here, σ represents the scale parameter that represents the standard deviation of the two-dimensional Gaussian filter used in the convolution process.

As shown in Figure 3, the SSFF module combines feature maps of different network depths and uses three-dimensional convolution to extract cross layer spatial features, enhancing feature expression. These are further processed through BN and SiLU to optimize and introduce nonlinearity, followed by extrusion operations to reduce size.

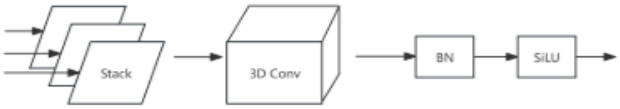


Figure 3: SSFF Structure Diagram

At the same time, in order to further enhance the network's ability to recognize small targets, we have adopted a strategy of increasing the neck structure of the P2 feature map output. Taking an input image with a size of 640 × 640 pixels as an example, the corresponding P2 feature map size is 160 × 160 pixels. Under this configuration, each feature map element corresponds to a receptive field of 4 × 4 pixels in the input image, which facilitates the detection of small targets and provides useful information to other levels during feature fusion. In addition, it enhances the understanding of context and reduces false positives and missed detections. The schematic diagram of the P2 small object detection layer is shown in Figure 4.

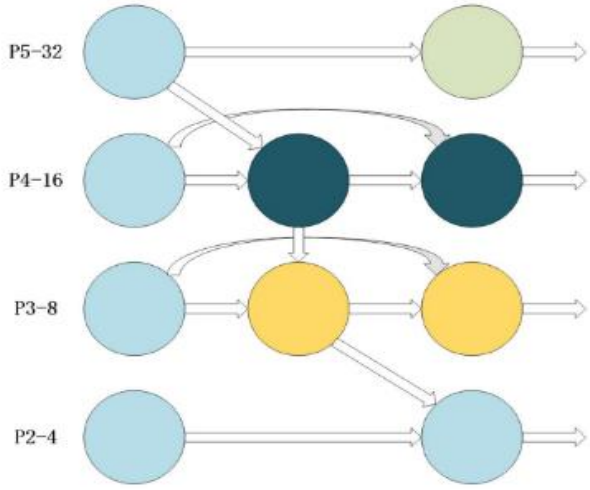


Figure 4: Schematic diagram of P2 small target detection layer

2.3 Optimizing the Positioning Loss Function

For the task of detecting small targets in drone aerial photography, this paper optimizes the positioning loss function of YOLOv11 and proposes a dynamic loss mechanism based on Wise IoU:

YOLOv11 adopts the CIoU + DFL regression loss combination, where CIoU introduces an aspect ratio penalty term (Equation 4-6) based on DIoU. However, it has drawbacks such as complex computation, insufficient sample differentiation, and ineffective aspect ratio penalty. DFL: By modeling class distribution, class imbalance can be alleviated (Equation 7), but the improvement in positioning accuracy is

limited. Therefore, this article introduces Wise IoU (WIoU). In terms of computational speed, the additional computational cost of WIoU mainly lies in the calculation of the focusing coefficient and the average statistics of IoU loss. Under the same experimental conditions, due to the absence of aspect ratio calculations, WIoU has a faster calculation speed than CIoU, with a calculation time of 87.2% of CIoU. In terms of performance improvement, WIoU not only considers area, centroid distance, and overlap area, but also introduces a dynamic non monotonic focusing mechanism. When the annotation quality of the dataset is poor, WIoU performs better relative to other bounding box losses. The weight calculation of WIoU can better reflect the differences in appearance and structure of the target, provide better target saliency, and facilitate the processing of targets with similar features. The specific information of WIoU is as follows:

Wise IoU v1: Due to the challenge of avoiding low-quality samples in the training data, geometric metrics such as distance and aspect ratio exacerbate the punishment for low-quality samples, resulting in a decrease in the model's generalization performance. A good loss function should reduce the penalty on geometric metrics when the anchor box and target box overlap well, and intervene as little as possible in training to enhance the model's generalization ability. In WIoU v1, distance attention is constructed based on distance metrics. The definition of WIoU v1 is shown in equation (3).

$$\mathcal{L}_{WIoUv1} = R_{WIoU}\mathcal{L}_{IoU} \tag{3}$$

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

In equation (4), R_WIoU ∈ [1, e) significantly amplifies the L_IoU of a regular mass anchor box, L_IoU∈[0,1], Significantly reduced the R-WIoU of high-quality anchor boxes, and in the case of good overlap between the anchor box and the target box, the attention of the anchor box to the distance from the center point was significantly reduced.

2) Wise IoU v2: The focus loss introduces a monotonic focusing mechanism tailored for cross entropy, effectively reducing the contribution of simple examples to the loss value. This enables the model to focus on challenging examples, thereby improving classification performance. Similarly, in v2, a monotonic focusing coefficient L_IoU ^ (r ^ *) was constructed for L_WIoUv1. The definition of Wise IoU v2 is shown in equation (5).

$$\mathcal{L}_{WIoUv2} = \mathcal{L}_{IoU}^{r^*}\mathcal{L}_{WIoUv1}, r > 0 \tag{5}$$

During the model training process, the gradient gain L_WIoUv1 decreases as Liou decreases, resulting in slower convergence speed in the later stages of training. Therefore, the introduced mean is used as the normalization factor, as shown in equation (6):

$$\mathcal{L}_{WIoUv2} = \left(\frac{\mathcal{L}_{IoU}^{r^*}}{\mathcal{L}_{IoU}}\right)^r \mathcal{L}_{WIoUv1} \tag{6}$$

The term L_IoU represents a moving average with momentum m, which dynamically updates the normalization factor to maintain the overall gradient gain r = ((L_IoU ^ (r ^ *))/L_IoU) ^ r at a high level, solving the problem of slow convergence speed in the later stages of training.

3) Wise IoU v3: Introducing the concept of outliers to characterize anchor box quality, defined as shown in equation (7):

$$\beta = \frac{L_{IoU}^*}{\overline{L_{IoU}}} \in [0, +\infty) \tag{7}$$

On the basis of Wise IoU v1, Wise IoU v3 introduces a non monotonic focusing coefficient based on β, defined as equation (8). A smaller outlier means a higher quality anchor box, resulting in a smaller gradient boost assigned to it, allowing better bounding box regression to focus on anchor boxes with common quality. For anchor boxes with large outliers, allocate smaller gradient boosting to effectively prevent harmful gradients from occurring in low-quality examples.

$$\mathcal{L}_{WIoUv3} = r\mathcal{L}_{WIoUv1}, r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{8}$$

At this point, when β=δ, δ makes r=1. When the outlier of the anchor box satisfies β=C (C is a constant value), the anchor box will obtain the maximum gradient lift. Due to the dynamic nature of LIoU and the dynamic quality standards of anchor boxes, Wise IoU v3 can dynamically allocate gradient boosting based on the current situation at any given time.

Through the above comparative analysis, this study achieved significant performance improvement by using Wise IoU v3 instead of traditional CIOU in YOLOv11. Wise IoU v3 adopts a dynamic non monotonic mechanism to evaluate the quality of anchor boxes, making the model more focused on anchor boxes of ordinary quality, thereby improving the object localization ability of the model. For the task of detecting small targets in drone aerial photography, the high proportion of small targets increases the difficulty of detection. Wise-IoUV3 can dynamically optimize the loss weight of small targets to improve the detection performance of the model.

3. Experimental Results and Analysis

3.1 Dataset and Experimental Environment

VisDrone 2019 is a drone aerial visual dataset collected by the AISKYEYE team from the Machine Learning and Data Mining Laboratory at Tianjin University, as shown in Figure 6. This dataset includes 10 categories: pedestrians, crowds, bicycles, cars, trucks, tricycles, sun shading tricycles, buses, and motorcycles, as well as many useful scenarios such as weather, terrain, and time. This dataset includes 6471 training set images, 548 validation set images, and 1610 test set images. In the dataset, image sizes range from 2000 × 1500 to 480 × 360. Due to being taken from a drone perspective, there are significant differences in shooting angle, image content, background, environmental illumination, and other aspects compared to images taken by ground personnel such as MS-COCO and VOC2012. Figure 5 shows an example image of VisDrone2019: (a) aerial view of urban roads at night under low light conditions; (b) Residential area images under high-intensity daytime lighting conditions; (c) Urban road intersections under glare conditions; (d) Panoramic view of city squares under cloudy conditions; (e) Aerial view of urban road areas under cloudy conditions; (f) Daytime low altitude sports field scene. The scenes composed of dataset images are very diverse, including streets, squares, parks, schools,

residential communities, etc. The lighting conditions for images include well lit daytime, poorly lit nighttime, cloudy, strong light, and glare conditions. The object types annotated in the image include 10 types: pedestrian, person, bicycle, car, truck, truck, tricycle, sunshade tricycle, bus, and electric motor.

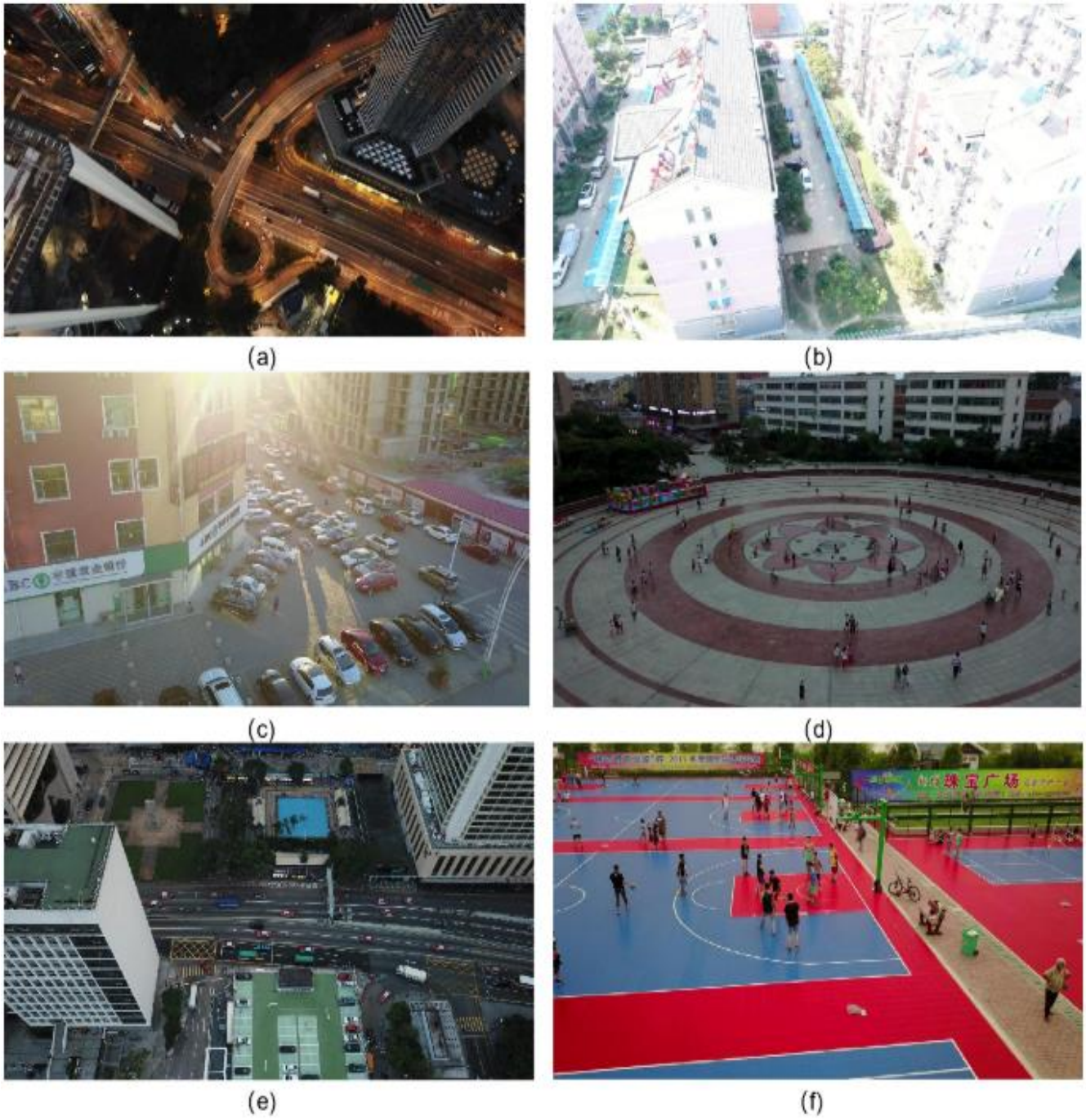


Figure 5: Example image of VisDrone 2019 dataset

The experimental equipment used in this article is RTX 3080ti 12GB. To ensure the effectiveness and fairness of the experiment, the hyperparameters of the experiment are uniformly set. The input image size is 640 × 640, IoU=0.5, The number of training rounds is 200, with an initial learning rate of 0.01 and a termination learning rate of 0.2, respectively. The SGD optimizer is used, and the batch_2 is 16. This article analyzes the performance of the algorithm model from two aspects: classification accuracy and model size. Select precision P and average precision mean mAP as evaluation indicators for classification accuracy, and model parameter Par and computational complexity O as evaluation indicators for model size.

3.2 Experimental Indicators

The experiment evaluated the proposed method from two aspects: detection performance and model parameter size. The experimental indicators include precision (P), recall (R), average precision (AP), mean average precision (mAP), and million parameters of network parameter size (M).

Precision (P) is the proportion of correctly predicted targets to all detected targets. Calculate through equation (9), where TP represents the correct prediction target and FP represents the incorrect prediction target.

$$P = \frac{TP}{(TP+FP)} \quad (9)$$

The recall rate (R) is the proportion of correctly detected targets among all existing targets. Calculate through equation (10), where FN represents a target that exists but has not been correctly detected.

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

Average precision (AP) represents the area enclosed by the curve composed of precision and recall. Calculate through equation (11). Our metrics include three different average accuracy metrics: AP0.5, AP0.75, and AP0.95. For AP0.5, in order to evaluate the bounding box prediction as true, the intersection of the joint score (IoU) between the predicted bounding box and the annotated bounding box must be greater than 0.50. For AP0.75, A bounding box prediction with an IoU score higher than 0.75 is considered correct. For AP0.95, we calculate the average accuracy values of different IoU scores within the range of 0.50:0.05:0.95, and then take the average of these calculated average accuracy values.

$$AP = \int_0^1 p(r)\, dr \quad (11)$$

Mean Precision (mAP) is the average precision of all types of samples, calculated using Equation (12). Our metrics include three different average accuracy metrics: mAP0.5, mAP0.75, and mAP0.95.

$$mAP = \frac{1}{K}\sum_{i=1}^{k} AP_i \quad (12)$$

3.3 Ablation Experiment

This article conducted 5 sets of ablation experiments, and the experimental results are shown in the table. MAP50 represents the average detection accuracy value of all categories when IoU=0.5, and mAP50-95 represents the average detection accuracy of all categories under 10 different IoUs with increasing IoUs from 0.5 to 0.95 at a step size of 0.05. The first set of experiments as a benchmark model showed poor performance in detecting small targets; The second experiment used the ASF-P2 structure proposed in this article. Due to the increase in small target information output by the backbone network, the accuracy of object detection was significantly improved. On the test set, the detection accuracy P and mAP50 increased by 4.7% and 3.9%, respectively; The third experiment added Wise IoU to the network, which resulted in a certain improvement in accuracy; The fourth experiment replaced the c3k2 module in the network with the FADC module. The experimental results showed that compared to the second experiment, the parameter calculation was reduced and the detection accuracy was improved; Finally, in the fifth experiment, the above three improvement methods were combined to improve the target detection accuracy by 6.7% compared to the baseline model. mAP50 and mAP50-95 improved by 5.7% and 4.3% respectively, with a parameter size of 2.5 × 106 and a computational complexity of 11.9 × 109. This indicates that the algorithm constructed in this article can significantly improve the accuracy of small object detection. The performance of the experimental results on the Visdrone 2019 test set is shown in Table 1.

Table 1: Results of ablation experiments

number	model	P/%	mAP_{50}/%	mAP_{50-95}/%	Par/ 10^6	O/ 10^9
1	YOLOv11n	42.1	32.6	18.8	2.6	6.3
2	YOLOv11n+ ASF-P2	46.8	36.5	21.6	3.4	12.0
3	YOLOv11n+ ASF-P2+ Wise IoU	47.1	37.1	22.0	3.4	12.0
4	YOLOv11n+ ASF-P2+ FADC	48.3	38.0	22.6	2.5	11.9
5	YOLOv11n+ ASF-P2+ FADC+ Wise IoU	48.8	38.3	23.1	2.5	11.9

3.4 Comparative Experiment

Our experimental results were compared with those of other methods published on this dataset over the years, including fast-R-CNN, RetinaNet, cascade-R-CNN, YOLOv4, TPH-YOLOv5, etc. Testing was conducted using the VisDrone2019 dataset in the same experimental environment to compare the mAP50 (%), mAP50-95 (%), and FPS of the algorithm. Due to limitations in experimental equipment, some experimental data were obtained by referencing other literature. The experimental results are shown in Table 2.

Table 2: Performance of Various Models in Comparative Experiments

method	mAP50/%	mAP50-95/%	FPS
faster - R-CNN	21.8	15.1	15
RetinaNet	13.9	9.6	4
cascade-R-CNN	23.2	16.5	6
YOLOv4	30.7	15.9	35
TPH-YOLOv5	37.3	20.8	32
YOLOv8	33.0	18.6	119.4
YOLOv10	34.7	18.9	125.9
Ours	38.3	23.1	130

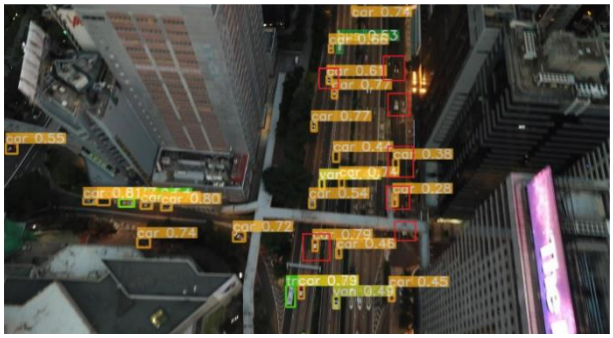
From Table 2, it can be seen that the mAP50 and mAP50-95 of our algorithm reached 38.3% and 23.1%, respectively, surpassing all YOLO series models and some improved models based on R-CNN that participated in the comparison. This significant performance improvement fully demonstrates that the algorithm proposed in this paper has higher accuracy in small target detection tasks in drone aerial photography.

3.5 Performance Analysis

Below is an analysis of the differences in object detection results between our algorithm and YOLOv11n in images of different scenes, lighting conditions, shooting positions, and target types.

As shown in Figure 6, in the middle right part of the image, there are 5 cars driving on the street, marked with red boxes. The first car had just arrived at the foot of the pedestrian overpass, and two bright lights could be seen shining on the ground. For these five vehicles, the algorithm in this article detected two of them, while YOLOv11n did not detect any of them. In the middle left part of the image, there are two cars driving on the road, marked with red boxes. The algorithm in this article detected two highly stable vehicles. YOLOv111n did not detect any vehicles.

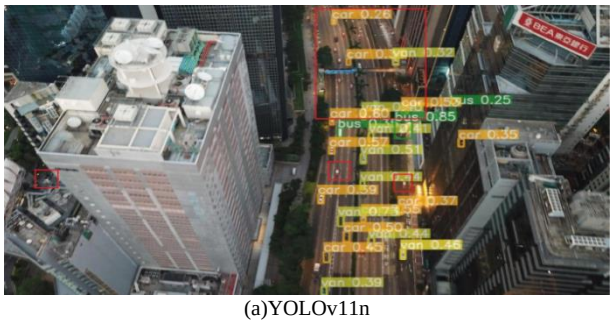
(a) YOLOv11n



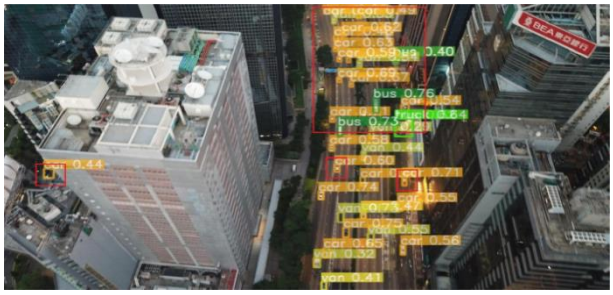
(b) This article's algorithm

Figure 6: Comparison effect of detection under insufficient lighting conditions

As shown in Figure 7, the upper part of the picture is surrounded by a large red box, which is a main road with many vehicles driving on it. Due to the high altitude and distance of aerial photography, the size of the car in the picture is relatively small. YOLOv11n only detected three cars in this area of the image. And the algorithm in this article detected most of the targets. In the center of the screen, there are two cars driving on the auxiliary road. YOLOv111n also did not detect these two cars; However, the algorithm in this article can detect these two vehicles. On the left side of the image, there is a car driving on a small road behind a building. For targets exposed in building gaps, YOLOv11n did not detect it, but our algorithm detected cars.



(a)YOLOv11n



(b) This article's algorithm

Figure 7: Comparison effect of detection under normal lighting conditions

As shown in Figure 8, in a low light nighttime environment, there are two motorcycles and two cars driving under the overpass in the red box on the left side of the image. This algorithm discovered a car, two motorcycles, and pedestrians on the motorcycles. YOLOv11n detected a car and a motorcycle.

(a)YOLOv11n

(b) This article's algorithm

Figure 8: Comparison effect of detection under insufficient lighting conditions

4. Conclusion

In response to the inherent characteristics of drone aerial images, such as high image density, small target size, and complex background, which result in low detection accuracy, this paper proposes an improved YOLOv11n based drone aerial small target detection algorithm. To enhance target background discrimination, we integrated the FADC module into the backbone of YOLOv11n. This module effectively enhances the backbone feature extraction of small targets, making it easier to distinguish them from the background. At the same time, ASF structure was introduced into the neck network to optimize the feature extraction and fusion process. It enhances the scale, spatial, and task perception capabilities of the model, and finally adopts the Wise IoU loss function and dynamic sample allocation strategy to reduce the model's focus on extreme samples and improve overall performance. The experimental results show that the proposed algorithm has significantly improved detection performance compared to other object detection algorithms. In the future, the model will be designed to be lightweight while ensuring its accuracy, and attempts will be made to deploy it to mobile devices to expand the application scope of the algorithm.

References

[1] Wu, X.; Li, W.; Hong, D.; Tao, R.; Du, Q. Deep Learning for Unmanned Aerial Vehicle-Based Object

Detection and Tracking: Asurvey. IEEE Geosci. Remote Sens. Mag. 2021, 10, 91–124. [CrossRef]

[2] Ahmed, S.; Kamal, U.; Hasan, K. DFR-TSD: A Deep Learning Based Framework for Robust Traffic Sign Detection under Challenging Weather Conditions. IEEE Trans. Intell. Transp. Syst. 2020, 23, 5150–5162. [CrossRef]

[3] Cao, J.; Zhang, J.; Jin, X. A Traffic-Sign Detection Algorithm Based on Improved Sparse R-cnn. IEEE Access 2021, 9, 122774–122788. [CrossRef]

[4] Shin, H.-C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.J.; Summers, R.M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans. Med. Imaging 2016, 35, 1285–1298. [CrossRef] [PubMed]

[5] Girshick, R.B.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014;pp. 580–587.

[6] Wei, L.; Dragomir, A.; Dumitru, E.; Christian, S.; Scott, R.; Cheng-Yang, F.; Berg, A.C. SSD: Single shot multibox detector. InProceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016.

[7] Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.

[8] Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.

[9] Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. arXiv 2018, arXiv:1804.02767.

[10] Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv 2020, arXiv:2004.10934.

[11] Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detectionon drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.

[12] Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. arXiv 2022, arXiv: abs/2209.02976.

[13] Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M.J.A. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv 2022, arXiv: abs/2207.02696.

[14] Chen, C.; Liu, M.Y.; Tuzel, O.; Xiao, J. R-CNN for small object detection. In Proceedings of the Computer Vision–ACCV 2016: 13thAsian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Springer: Cham, Switzerland, 2016.

[15] Lim, J.-S.; Astrid, M.; Yoon, H.; Lee, S.-I. Small object detection using context and attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju, Republic of Korea, 13–16 April 2019; pp. 181–186.

[16] Liu, H.; Duan, X.; Chen, H.; Lou, H.; Deng, L. DBF-YOLO: UAV Small Targets Detection Based on Shallow Feature Fusion. IEEJTrans. Electr. Electron. Eng. 2023, 18, 605–612. [CrossRef]

[17] Yang, R.; Li, W.; Shang, X.; Zhu, D.; Man, X. KPE-YOLOv5: An Improved Small Target Detection Algorithm Based on YOLOv5.Electronics 2023, 12, 817. [CrossRef]

[18] Zhang, X.; Feng, Y.; Zhang, S.; Wang, N.; Mei, S. Finding Nonrigid Tiny Person With Densely Cropped and Local Attention Object Detector Networks in Low-Altitude Aerial Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2022, 15, 4371–4385. [CrossRef]

[19] Jin, R.; Lin, D. Adaptive Anchor for Fast Object Detection in Aerial Image. IEEE Geosci. Remote Sens. Lett. 2020, 17, 839–843. [CrossRef]

[20] Liu, P.; Wang, Q.; Zhang, H.; Mi, J.; Liu, Y. A Lightweight Object Detection Algorithm for Remote Sensing Images Based onAttention Mechanism and YOLOv5s. Remote Sens. 2023, 15, 2429. [CrossRef]

[21] Tong, Z., Chen, Y., Xu, Z., et al. (2023) Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechan-ism. https://doi.org/10.48550/arXiv.2301.10051

[22] CHEN L, GU L, FU Y. Frequency-Adaptive Dilated Convolution for Semantic Segmentation[J]. ArXiv Preprint, 2024, ArXiv: 2403. 05369.

[23] Y. Quan, D. Zhang, L. Zhang, and J. Tang, ‘‘Centralized featurepyramid for object detection,’’ IEEE Trans. Image Process., vol. 32, pp. 4341–4354, 2023, doi: 10.1109/TIP.2023.3297408.

[24] Zhang, H.; Wang, Y.; Dayoub, F.; Sünderhauf, N. VarifocalNet: An IoU-aware Dense Object Detector. In Proceedings of the 2021IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.