Automation and Machine Learning (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2025.060203
ISSN 2516-5003 Vol. 6 Num. 1

Small Target Detection Algorithm for UAV Aerial Photography Based on YOLOv11n

Lingsheng Liang[1,a], Zijian Dong[1,b,*], Zhaojin Huangfu[1,c], Xinrui Chen[1,d]

[1]School of Electronic Engineering, Jiangsu Ocean University, Lianyungang, Jiangsu, 22000, China
[a]2023220611@jou.edu.cn, [b]1995000021@jou.edu.cn, [c]2023220624@jou.edu.cn, [d]2023220620@jou.edu.cn
**Corresponding author*

Keywords: YOLOv11n, Small Target Detection, Recalibrated Feature Pyramid, Focal-DIoU

Abstract: To address the challenges of small target detection in aerial images captured by unmanned aerial vehicles (UAVs), such as complex backgrounds, dense targets, large scale variations, and mobile deployment, this paper proposes an improved algorithm, RRF-YOLOv11n, based on the YOLOv11n model. Firstly, a convolutional layer C3K2-RVB-EMA is constructed by integrating RepViTBlock and an efficient multi-scale attention module (EMA), enhancing the model's feature extraction capability for multi-scale targets in complex backgrounds, especially for significantly deformed small targets. Secondly, to deal with the situation where small targets are more numerous in UAV aerial images, a new small target detection layer P2 (PredictionLayer2) is added and the large target detection layer P5 is removed, effectively improving the capture accuracy of small target features while reducing redundant computations in the large target detection layer. Thirdly, a Re-Calibration FPN is introduced to replace the traditional pyramid, recalibrating the boundary and semantic information in features and enhancing the weight of important features. Finally, a Focaler-DIoU loss function combining Focal Loss and DIoU is proposed, optimizing the accuracy and convergence speed of bounding box regression and solving the sample imbalance problem in small target detection. Experimental results show that RRF-YOLOv11n outperforms the original YOLOv11n model by 6.9% in the mAP50 metric on the Vis-Drone2019 dataset, reaching 41.2%, and enhances the detection accuracy of small targets in UAV aerial images. Compared with other advanced target detection algorithms, this algorithm demonstrates superior performance in both detection accuracy and speed.

1. Introduction

In recent years, UAV detection technology has been widely applied in various fields such as military reconnaissance, logistics and distribution, agricultural plant protection, environmental monitoring, and disaster response. However, in practical applications, UAV detection technology faces numerous challenges, particularly in the detection of small targets. These small targets are typically characterized by weak features, complex backgrounds, dense distribution, and diverse scales

posing difficulties for accurate detection. Therefore, research on UAV small target detection algorithms has attracted significant attention.

In the field of deep learning, target detection mainly includes one-stage algorithms such as the YOLO series[1], SSD[2], and RetinaNet[3], as well as two-stage algorithms like RCNN[4] and Faster R-CNN[5] To address the low accuracy of small tar-get detection, Hou[6] et al. improved the YOLOv8 backbone network by using four feature detection heads to enhance the detection rate of small targets, and designed the ConvSPD convolution module and BiFormer attention to strengthen the ability to capture shallow detail features of small targets. However, due to the small receptive field of shallow features, they cannot fully cover small targets in some complex scenarios. Moreover, the complex structure of the ConvSPD convolution may increase the computational load and parameter count of the model. Li[7] et al. designed a dilated feature pyramid convolution module to replace the original SPPF layer, which strengthens the extraction of detailed features of UAV small targets. They also introduced the CSPOK module and GGBD convolution to improve the global feature extraction capability and multi-scale feature fusion capability. Peng[8] et al. enhanced the representational capability of multi-scale structural feature maps by adding an additional contextual semantic enhancement module, which improved the detection capability of small targets, but the false detection rate for similar targets remains high. Zhao[9]et al. improved the detection performance in complex backgrounds through the DynamicHead with multiple attention mechanisms, increased multi-scale detection to enhance the extraction of small and medium targets, and integrated DenseNet to strengthen feature transmission and prevent overfitting. Zhou[10] et al. replaced the back-bone network CSPDarknet with the lightweight Mo-bileNet-V3, reducing the number of model parameters and improving inference speed. Despite extensive research efforts to improve the accuracy of small target detection, existing methods still exhibit significant shortcomings in performance under complex backgrounds. To address these deficiencies, this paper proposes an improved algorithm based on YOLOv11n, namely RRF-YOLOv11n. Through innovative design, this algorithm not only improves the detection accuracy of small targets but also significantly reduces model parameters and optimizes the use of computational resources, providing an efficient and accurate solution for small target detection in UAV aerial photography.

2. The proposed method

2.1 Enhanced YOLO Model

YOLOv11, the new-generation computer vision model launched by Ultralytics, has achieved significant improvements in both speed and accuracy compared to previous models in the YOLO series. For example, its backbone network adopts the more efficient C3K2 module to replace the original C2f module, enabling the selection of appropriate feature extraction methods according to different needs and scenarios. This improvement enhances the flexibility and adaptability of the model, allowing it to better handle image data of varying complexities. It also introduces the C2PSA module, which, through the multi-head attention mechanism and feedforward neural network, can selectively focus on important parts of the input features, suppress unimportant information, and enhance multi-scale feature extraction capabilities. The neck network uses a PAN-FPN structure, which enhances the fusion of shallow position information and deep semantic information through bottom-up path enhancement, thereby strengthening the target localization ability. The detection head of YOLOv11 adopts a decoupled design and DWConv operation, reducing model parameters and computational overhead.

In general, through improvements in architectural optimization, performance enhancement, and adaptability improvement, YOLOv11 has become a more advanced,efficient and adaptable computer vision model.

However, for the problem of missed detection and false detection caused by the characteristics of small target detection in UAV aerial images, such as complex background, dense targets, large scale changes, and indistinct features, the effect of using YOLOv11n is not ideal, and there is still much room for improvement. Based on YOLOv11n, this paper proposes a model RRF-YOLOv11n suitable for small target detection in UAV aerial images, as shown in Fig. 1.

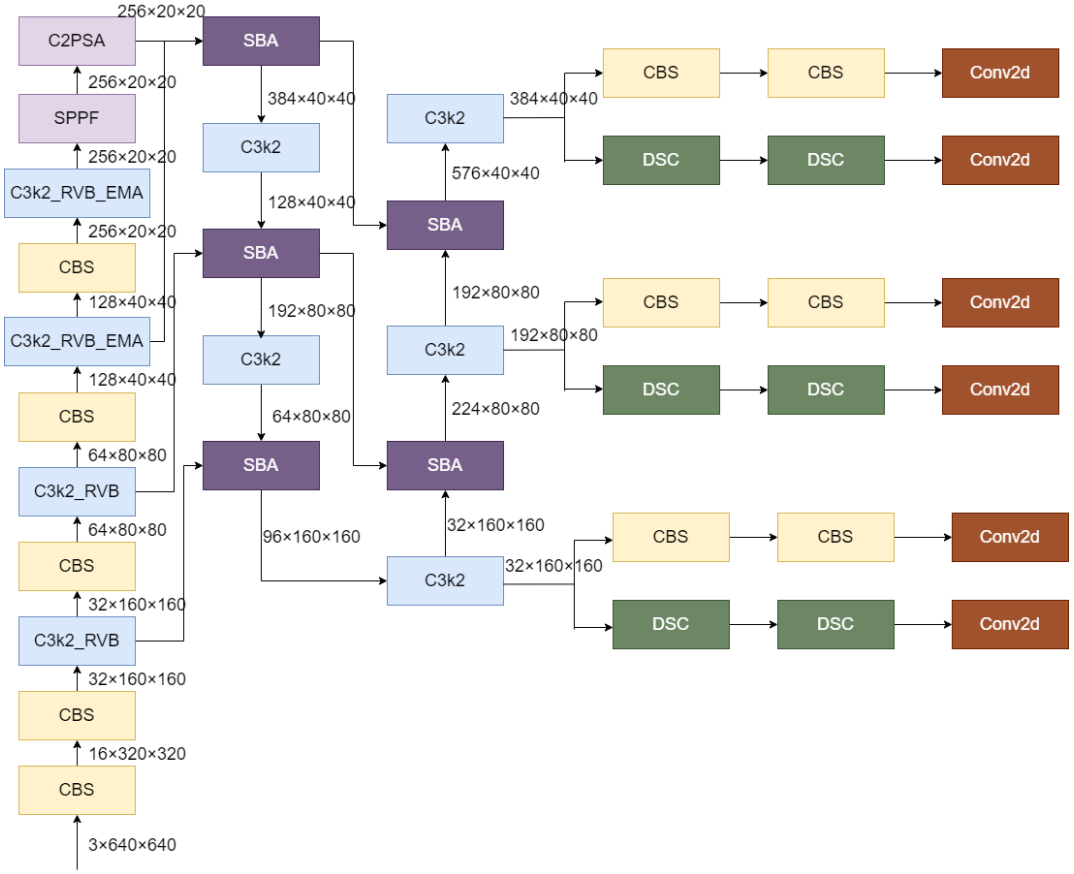


Figure 1 Structure diagram of the RRF-YOLOv11n network.

2.2 Structure of C3K2-RVB-EMA

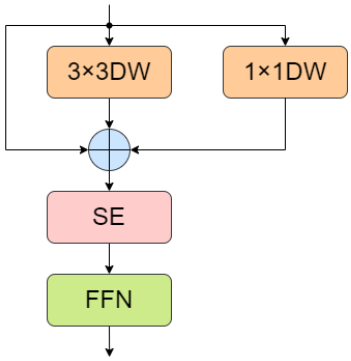


Figure 2 The structure of RepViTBlock.

In recent years, the RepViT-SAM model optimized for high-resolution visual tasks has been proposed. Its core component, RepViTBlock, integrates depthwise separable convolution with feedforward neural networks to construct an efficient feature extraction architecture. As shown in

Figure 2, this module adopts an early convolutional backbone design, achieving a fourfold down-sampling of input features through two sets of 2-strided convolutions. This strategy inherits the advantages of traditional CNN feature extraction, enabling the model to quickly capture basic visual features such as edges and textures in the initial stage, while significantly reducing computational complexity.

It is worth noting that the RepViTBlock incorporates an SE component for channel attention within its structure, which can dynamically allocate weights to feature channels: enhancing the representation power of key features while effectively suppressing irrelevant and redundant information. This ultimately helps the model significantly improve its performance in selecting effective features in complex scenarios. Additionally, thanks to its pure convolutional network architecture, the model demonstrates excellent real-time performance when processing high-resolution images, making it particularly suitable for applications where inference speed is a critical factor.

In the backbone of the YOLOv11n model, the original C3K2 module adopts the Bottleneck structure for feature extraction and computational efficiency optimization. However, when dealing with small targets in drone aerial images, it has insufficient feature extraction capabilities for small targets, limited multi-scale feature fusion ability, poor robustness against complex backgrounds, and low computational efficiency and accuracy. Therefore, this paper draws on the idea of RepViTBlock and designs the C3K2-RVB module to improve the C3K2 module. The C3K2-RVB module replaces the Bottleneck module in C3K2 with RepViTBlock; this module reduces the computational load and parameter quantity by using depthwise separable convolution and structural reparameterization, and effectively extracts multi-scale features by combining the depthwise down-sampling module and multi-stage design, significantly improving the accuracy rate. Further, the EMA attention mechanism is integrated into the RepViTBlock of C3K2-RVB to form the C3K2-RVB-EMA module (structure shown in Figure 3). The EMA attention can dynamically regulate the channel and spatial position weights of the feature map, enhancing the focus on small target features, weakening background interference, and weighted fusion of multi-scale features, thereby improving the model's perception efficiency for multi-scale targets.

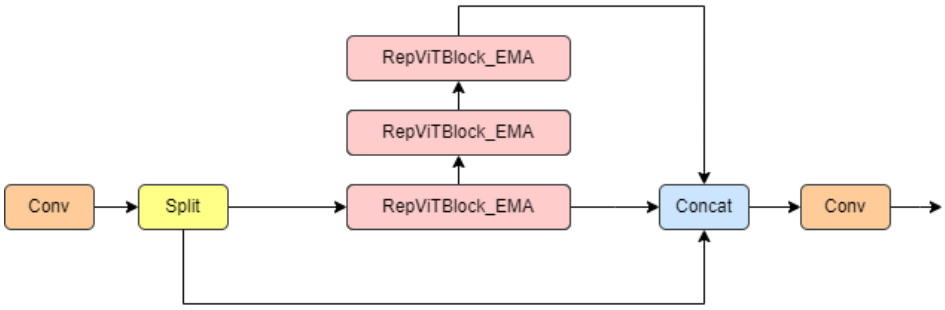
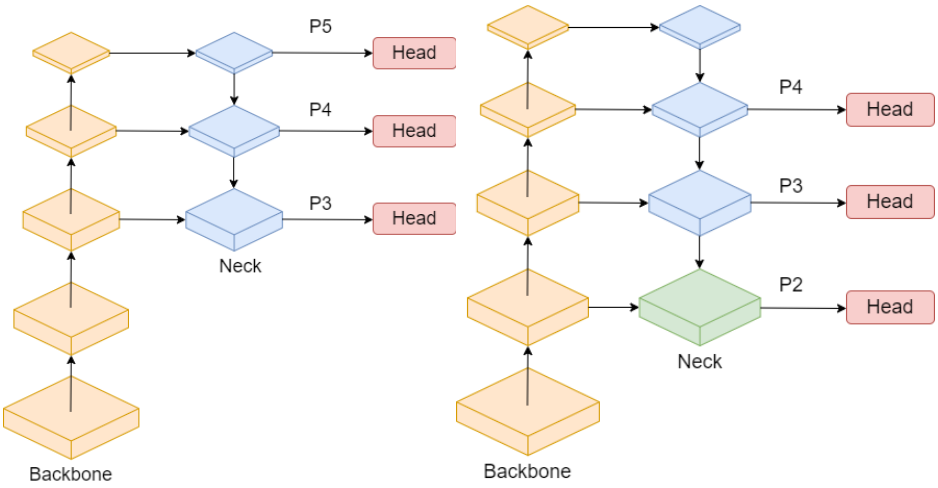


Figure 3 The structure of C3K2-RVB-EMA.

2.3 Add a small target detection layer

The YOLOv11n detection layer (Figure 4(a)) is in a three-head form, consisting of P3, P4, and P5 layers, which correspond to resolutions of 80×80, 40×40, and 20×20 respectively for detecting targets of different scales. Among them, the P5 layer has a lower resolution and rich semantic information, and is mostly used for large target detection. However, due to its large pixel receptive field, it is prone to losing details when processing small targets; while the local features of small targets (such as vehicles and pedestrians) in unmanned aerial vehicle remote sensing images are crucial for detection, the P5 layer cannot effectively express these features.

To address this issue, this paper proposes an improvement: deleting the P5 large target detection layer and adding a dedicated small target detection layer P2 (Figure 4(b)). The resolution of the P2 layer is increased to 160×160, with a smaller pixel receptive field, which can retain more details of small targets and thereby improve the accuracy of small target detection; at the same time, this modification can reduce redundant computations in feature fusion of the model and further enhance the ability of multi-scale feature extraction.



(a) Original three-detector head structure. (b) Improve the structure of the three detection heads.

Figure 4 Improvement of the detection head structure.

The introduced small target detection layer has enhanced the processing capability of the YOLOv11n model for unmanned aerial vehicle (UAV) aerial photography tasks with small targets. It effectively reduces the omission and false detection of small targets, improves the detection accuracy, and ensures the balance of computational efficiency. Through optimized design, this model demonstrates higher performance when dealing with small-sized targets in complex scenarios, while maintaining overall computational efficiency.

2.4 Structure of Re-CalibrationFPN

The Neck structure of YOLOv11n adopts a design combining FPN and PANet. The core idea is to fuse multi-scale feature maps through top-down, bottom-up and lateral concatenation paths to generate feature maps (such as P3, P4, P5) rich in semantic and detailed information for target detection. However, after multiple sampling in this structure, the feature details are easily lost, resulting in insufficient feature capture for small aerial targets and causing missed detections and false detections; moreover, the direct fusion of low-level and high-level features is prone to redundancy and inconsistency, and without feature re-calibration and boundary enhancement mechanisms, it is unable to dynamically adjust feature weights, weakening the expression of important features, and affecting the detection effect of small targets and blurred boundaries.

To solve the above problems, this paper replaces the original FPN with Re-CalibrationFPN: a new SBA module is added to selectively aggregate and re-calibrate the boundaries and semantic information, enhancing the weight of important features, to refine the target contour and calibrate the target position; at the same time,a new recalibration attention unit (RAU) block is introduced, adaptively extracting complementary information from two inputs (Fs and Fb) shallow and deep information is input into two RAU blocks in different ways to make up for the lack of high-level spatial boundary information and low-level semantic information. Finally, after 3×3 convolution, the two RAU blocks are concatenated to output, achieving robust combination of different features and

rough feature refinement, as shown in Figure 5.

The calculation formula of RAU is as follows:

$$T_1' = W_\theta(T_1), T_2' = W_\emptyset(T_2) \tag{1}$$

$$PAU(T_1, T_2) = T_1' \odot T_1 + T_2' \odot T_2 \odot \left(\odot\left(T_1'\right)\right) + T_1 \tag{2}$$

Among these, T1 and T2 are input features. Two linear mappings and sigmoid functions $W_\theta(\cdot)$ and $W_\theta(\cdot)$ are applied to the input features to reduce the channel dimension of the input features to 32, resulting in feature maps T_1' and T_2'. $\odot$ denotes element-wise multiplication, and $\odot(\cdot)$is the inverse operation performed by subtracting the feature T_1', which refines the imprecise and coarse estimates into accurate and complete prediction maps.

Finally, using a convolution operation with a kernel size of 1×1 as the linear mapping process, the output of the SBA module can be calculated by the following formula:

$$Z = C_{3\times3}\left(Concat\left(PAU(F^S, F^b), PAU(F^b, F^s)\right)\right) \tag{3}$$

Among them, $C_{3\times3}(\cdot)$ is a 3×3 convolution operation with batch normalization and ReLU activation layers. F^S and F^b are features containing deep semantic information and shallow boundary information respectively. Concat(·) is the concatenation operation along the channel dimension, and Z is the output of the SBA module.

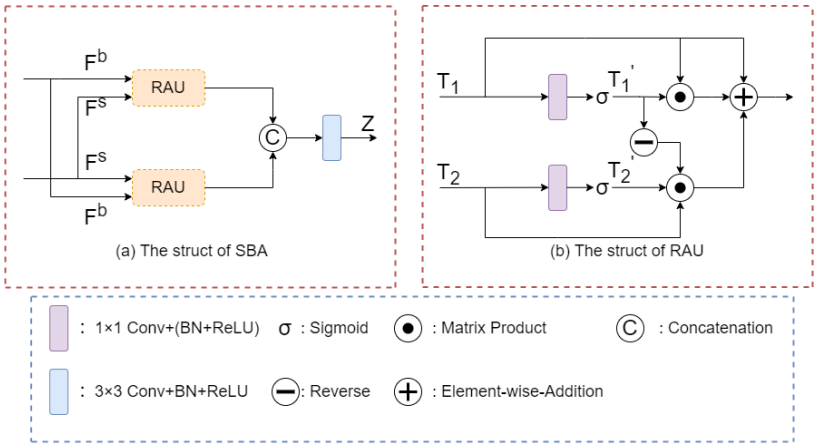


Figure 5 The struct of SBA and RAU.

2.5 Focaler-DIoU Loss Function

The traditional CIoU loss function in YOLOv11n takes into account the center distance, aspect ratio, and intersection-over-union (IoU) of bounding boxes. However, when dealing with small targets, the aspect ratio penalty term is prone to introducing additional noise, resulting in inaccurate regression; and it has low sensitivity to small target detection and is difficult to locate, with limited effectiveness in complex backgrounds, target overlaps, and scenarios with dense small targets.

To address these issues, this paper designs a Focaler-DIoU loss function based on FocalLoss and DIoU improvement: In this function, FocalLoss uses a regulation factor $(1-pt)\gamma$ to reduce the weight of easily classified samples and increase the weight of difficultly classified samples, thereby alleviating the problem of class imbalance; DIoU adds a penalty term on top of IoU, minimizing the normalized distance between the center points of the predicted box and the target box, accelerating the convergence speed of bounding boxes, optimizing the positioning effect, and reducing positioning errors caused by the small size or occlusion of small targets. The core of the Focaler-DIoU loss function lies in integrating the mechanism of FocalLoss into DIoU, using the regulation factor to

enhance the attention to small targets.

Its calculation formula is:

$$FocalLoss(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{4}$$

$$L_{DIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} \tag{5}$$

$$L_{Focaler-DIoU} = (1-IoU)^{\gamma} \cdot L_{DIoU} \tag{6}$$

Here, p_t represents the predicted category probability by the model; α_t is the category weight used to balance positive and negative samples; γ is the adjustment factor that controls the weight distribution of difficult and easy samples; IoU is the intersection-over-union ratio between the predicted box and the target box; c is the diagonal length of the smallest enclosing rectangle that can contain both the predicted box and the real box; ρ is the Euclidean distance between the centers of the predicted box and the real box; b is the center point of the predicted box; b^{gt} is the center point of the real box.

The improved loss function combines the mechanism of FocalLoss, reducing the weight of easy-to-classify samples (such as large targets) and increasing the weight of difficult-to-classify samples (such as small targets), thereby enhancing the focus on small targets. At the same time, it retains the advantages of DIoU by minimizing the center point distance to accelerate convergence. Compared to the traditional CIoU loss function, it significantly improves the ability to detect small targets and performs better in scenarios with dense targets, severe occlusion, or class imbalance.

3. Experiments and discussion

3.1 Dataset

The UAV aerial photography small target dataset used in this experiment is VisDrone2019[11]. This dataset was collected and constructed by the AISKYEYE team of the Machine Learning and Data Mining Laboratory of Tianjin University. The dataset covers 14 different urban and rural scenes, including 10 types of targets such as pedestrians, people, bicycles, cars, vans, trucks, tricycles, covered tricycles, buses, and motorcycles. In terms of data scale, the dataset contains a total of 8,629 images, with 6,471, 548, and 1,610 images in the training set, validation set, and test set respectively. The entire dataset contains a total of 2.6 million target instance samples.

3.2 Experimental environment

The experimental environment of this paper was all conducted on the GPU. The hardware configuration and experimental parameters used are shown in Table 1 and Table 2.

Table 1 Experimental environment configuration.

Parameter	Experimental Environment
Operating System	Linux
CPU	Inteli9-13900HX
GPU	RTX4090(24GB)
Internal Storage	64GB
Python	3.8
Pytorch	2.1.0
Cuda	11.8

Table 2 Model training parameters.

Parameter	Configuration
Imagessize	640
Epoch	300
Batchsize	16
Optimizer	SGD

3.3 Evaluation index

In order to better evaluate the performance of the model, this paper uses precision (P), recall (R), mAP50, mAP50:95, parameter quantity, and computational quantity as the main evaluation criteria. mAP50 represents the average detection accuracy of the model when the IoU threshold is 0.50, while mAP50:95 represents the IoU threshold ranging from 0.50 to 0.95. The corresponding formulas are as follows:

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{9}$$

Among them, TP represents the number of samples that are actually positive and whose prediction result is also positive; FP represents the number of samples that are actually negative but whose prediction result is positive; FN represents the number of samples that are actually positive cases but whose prediction result is negative; P is the accuracy rate; R is the recall rate, AP is the average accuracy rate, and the average value of AP for all categories can be obtained as mAP.

3.4 Experimental Results and Analysis

3.4.1 Algorithm comparison

In order to comprehensively evaluate the detection performance of RRF-YOLOv11n for small targets, under the same conditions, it was compared with several current mainstream target detection algorithms. The experimental results are shown in Table 3.

Table 3 Algorithm comparison experiment on VisDrone2019.

Model	P	R	mAP50	mAP50:95	Parameters/M	GFLOPs
Faster R-CNN	0.343	0.365	0.305	0.129	63.2	370.0
SSD	0.210	0.356	0.24	0.117	12.3	63.2
YOLOv5n	0.438	0.322	0.324	0.188	2.18	5.8
YOLOv5s	0.482	0.351	0.372	0.225	7.81	18.8
YOLOv7-tiny	0.425	0.329	0.326	0.191	6.01	12.3
YOLOv8n	0.436	0.335	0.333	0.196	2.68	6.8
YOLOv8s	0.505	0.387	0.4	0.238	9.83	23.4
TPH-YOLOv5[12]	0.501	0.391	0.405	0.246	60.4	145.8
UAV-YOLOv8s[13]	0.505	0.393	0.402	0.241	10.30	-
YOLOv11n	0.45	0.344	0.343	0.201	2.58	6.3
YOLOv11s	0.51	0.398	0.407	0.246	9.41	21.3
Ours	0.508	0.391	0.412	0.251	2.45	16.5

From Table 3, it can be seen that the improved algorithm ranks at the forefront in terms of detection performance. The mAP50 and mAP50-95 values reached 41.2% and 25.1% respectively, which were 6.9% and 5% higher than those of the baseline YOLOv11n model, respectively. Secondly, the improved model outperformed YOLOv11s in terms of recall rate, mAP50, and mAP50-95, and the parameter size and computational cost were only 25% and 77% of those of YOLOv11s, respectively, demonstrating that the improved model maintained the accuracy improvement while not increasing the parameter size and computational cost. Then, compared with the first-stage SSD and the YOLO series, the detection accuracy was higher, and the parameter size was only slightly higher than YOLOv5n. Secondly, compared with the second-stage Faster R-CNN, the improved model was comprehensively ahead. Finally, when comparing with the existing algorithm literature [12] and [13], although it was slightly ahead in terms of accuracy and recall rate, the algorithm in this paper was far superior in terms of parameter size and computational cost to both [12] and [13]. In summary, the comparative experiments verified the superiority of the RRF-YOLOv11n algorithm, improved the detection accuracy for small targets based on the drone perspective, effectively balanced the model performance, parameters, and computational cost, and outperformed most common algorithms, having good design value.

3.4.2 Ablation experiment

To verify the effectiveness of the improved module proposed in this paper in enhancing model performance and the contribution of each module, an ablation experiment was conducted on the VisDrone2019 dataset. The experiment was based on YOLOv11n and gradually integrated the improved methods by module. The contributions of each module to the detection accuracy and efficiency were quantified. The specific process is as follows: first, replace the C3K2 in the network with the C3K2-RVB-EMA structure; then, improve the Neck part to the Re-CalibrationFPN structure; next, add small target detection layers and delete the P5 backbone part; finally, replace the traditional CIoU loss function with the improved Focaler-DIoU loss function. The detailed results of the ablation experiment are shown in Table 4 (Baseline is the YOLOv11n model).

Table 4 Ablation experiment.

Baseline	C3K2 -RVB	P2	Re-CalibrationFPN	Focaler-DIoU	P	R	mAP50	mAP50:95	Parameters/M	GFLOPs
√					0.45	0.344	0.343	0.201	2.58	6.3
√	√				0.456	0.346	0.351	0.211	2.44	6.1
√	√	√			0.472	0.369	0.37	0.23	1.84	9.7
√	√	√	√		0.497	0.404	0.399	0.247	2.42	16.5
√	√	√	√	√	0.508	0.391	0.412	0.251	2.45	16.5

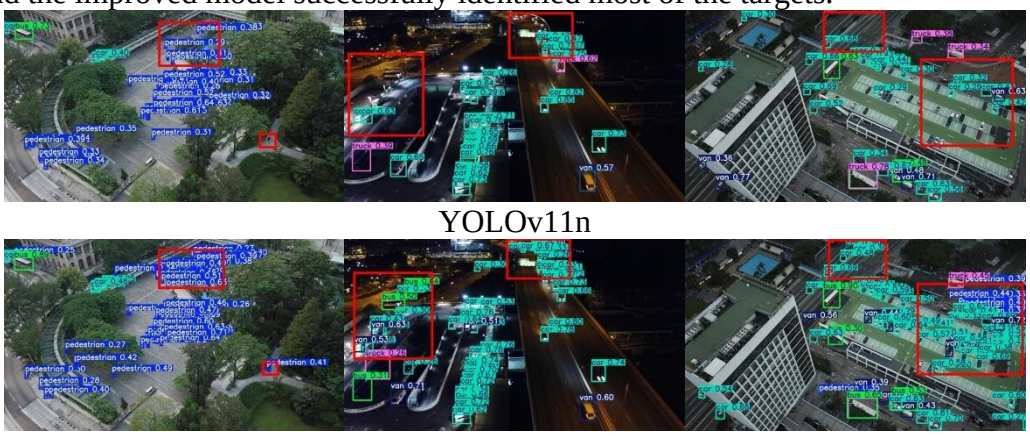
From Table 4, it can be seen that when C3K2 is replaced with the C3K2-RVB-EMA structure, mAP50 and mAP50:95 increase by 0.8% and 1% respectively. This indicates that after adding the RepViTBlock and EMA module, multi-scale features can be effectively extracted, information loss is reduced, and the weights of important features are enhanced while the weights of unimportant features are suppressed. Secondly, by adding the P2 small target detection layer and deleting the P5 large target detection layer, mAP50 and mAP50:95 increase by 2.7% and 2.9% respectively, and the parameter quantity decreases by 28.7%. This shows that adding the small target detection layer has the performance advantage in detecting small targets. Again, by introducing the Re-CalibrationFPN structure instead of the original FPN structure, mAP50 and mAP50:95 increase by 5.6% and 4.6% respectively. This result highlights the advantages of Re-CalibrationFPN in dynamic feature fusion, cross-layer information interaction, and global context modeling, significantly improving the performance of small target detection. Finally, replacing the loss function with Focaler-DIoU, in the situation where the parameter quantity and computational quantity hardly change, mAP50 and

mAP50:95 increase by 6.9% and 5% respectively. This indicates that this loss function effectively handles low-quality samples and class imbalance. From the above ablation results, it can be seen that the model accuracy keeps improving, indicating the effectiveness of the model improvement.

3.5 Visualized results and analysis

To verify the effectiveness of the improved algorithm in detecting small targets of unmanned aerial vehicles (UAVs), a visual comparison was conducted on typical complex scenarios such as dense, low-light, and high-altitude targets from the VisDrone2019 test set. The detection results are shown in Figure 6. The results indicate that the algorithm proposed in this paper can more accurately detect and locate small targets such as cars and pedestrians.

Specifically, in a dense environment (a), the RRF-YOLOv11n algorithm shows significant advantages in dense crowd detection. The red area detected by the YOLOv11n algorithm missed a large number of pedestrians. The improved model reduces the number of missed detections and can identify occluded targets. In a low-light environment (b), when facing a small vehicle group with insufficient lighting, the improved model has a low detection rate of missed areas and performs better than the benchmark method, verifying its adaptability to complex lighting conditions in dense vehicle scenes. The detection of high-altitude targets further demonstrates the advantages of the improved algorithm. In the high-altitude environment (c), YOLOv11n completely missed extremely small targets, and the improved model successfully identified most of the targets.



YOLOv11n

RRF-YOLOv11n

(a) Dense scene (b) Low-light scene (c) High-altitude scene

Figure 6 Comparison of the detection effectiveness of high-altitude vehicles under bright light conditions and low light conditions.

4. Conclusion

In response to the challenges faced by small target detection in drone aerial images, such as complex backgrounds, dense targets, and diverse scales, this paper proposes an improved algorithm RRF-YOLOv11n based on YOLOv11n. By constructing a C3K2-RVB-EMA module that integrates RepViTBlock and EMA attention, it enhances the extraction of multi-scale target features in complex backgrounds; adding a P2 small target detection layer and eliminating the redundant P5 large target layer, it improves the ability to capture the details of small targets while reducing parameters; introducing the Re-CalibrationFPN structure to dynamically re-calibrate features to optimize the fusion of boundaries and semantic information, alleviating the problem of detail loss in traditional FPN; designing the Focaler-DIoU loss function to balance sample distribution and accelerate the convergence of bounding box regression. Experiments show that this model achieves an mAP50 of

41.2% on the VisDrone2019 dataset, an improvement of 6.9% compared to the baseline, with a 5% reduction in parameters and an optimization of computational complexity to 16.5GFLOPs, achieving a balanced improvement in accuracy and efficiency.

However, the improved model still has limitations. For example, the parameter reduction is not significant after optimizing tiny targets. In the future, techniques such as pruning and distillation will be adopted to reduce model complexity and further improve performance to meet actual deployment requirements.

References

[1] Y. Liu, P. Sun, N. Wergeles, et al., A survey and performance evaluation of deep learning methods for small object detection[J]. Expert Systems with Applications, 2021, 172: 114602.

[2] Liu, W., Angelov, D., Erhan, D., et al. SSD: Single shot multibox detector. Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I, 21-37.

[3] T.Y. Ross, G. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2980-2988.

[4] R. Girshick, J. Donahue, T. Darrell, et al., Rich feature hierarchies for accurate object detection and semantic segmentation, in: 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, JUN 23-28, 2014, CVPR, Comp Vis Fdn; IEEE; IEEE Comp Soc, 2014, pp. 580-587.

[5] S.Q. Ren, K.M. He, R. Girshick, et al., Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[6] Y. Hou, Y. Wu, X.R. Kou, et al., Improved YOLOv8 algorithm for small target detection in UAV aerial images[J]. Computer Engineering and Applications, (Online) 2025, pp. 21-37.

[7] B. Li, S.L. Li, Improved YOLOv11n algorithm for UAV small target detection[J]. Computer Engineering and Applications, 2025, 61(7): 96-104.

[8] Y.F. Peng, T. Zhao, Y.K. Chen, et al., UAV small target detection algorithm based on contextual information and feature refinement[J]. Computer Engineering and Applications, 2024, 60(5): 183-190.

[9] X. Zhao, L.L. Chen, W.C. Yang, et al., DY-YOLOv5: Aerial image object detection based on multi-attention mechanism[J]. Computer Engineering and Applications, 2024, 60(7): 183-191.

[10] Q. Zhou, G.Q. Tan, S.L. Yin, et al., Improved YOLOv5s algorithm for road object detection[J]. Chinese Journal of Liquid Crystals and Displays, 2023, 38(5): 680-690.

[11] D. Du, P. Zhu, L. Wen, et al., VisDrone-DET2019: the vision meets drone object detection in image challenge results, in: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshops, Piscataway: IEEE, 2019, pp. 213-226.

[12] Q. Zhao, B. Liu, S. Lyu, et al., TPH-YOLOv5++: Boosting object detection on drone-captured scenarios with cross-layer asymmetric transformer[J]. Remote Sensing, 2023, 15(6): 1687.

[13] Y.W. Li, Y.P. Feng, W.Z. An, et al., Small target detection algorithm in UAV images based on improved YOLOv8[J]. Chinese Journal of High Technology Letters, 2024, 34(7): 765-775.