Research on UAV Target Detection Based on Improved YOLOv11

Guangdong Ren, Junyi Wu, Weizheng Wang

School of Safety Engineering, North China Institute of Science and Technology, Yanjiao, China
Email: rgd2497788737@163.com

Abstract

In response to the challenges of small object detection in UAV aerial photography, such as complex backgrounds, tiny targets, dense targets, and edge deployment, the YOLOv11n model was improved. Specifically, an EfficiBackbone module was designed for the backbone part, the C3K2 was improved using the RipViT block in the Neck part, and the original detection head was replaced with a dynamic detection head. The improved YOLOv11 network was thus completed. Experimental results show that the model has significantly improved mAP@0.5 and mAP@0.5:0.95 on the VisDrone2019 dataset, proving the effectiveness of the model.

Keywords

YOLOv11, Aerial Images, Small Target Detection, Single-Object Detection, Dynamic Detection Head

1. Introduction

With the rapid development of intelligent and automated technologies, the combination of drones and image processing techniques has provided strong support for real-time monitoring and precise analysis of vast areas, helping decision-makers to obtain detailed data in a timely manner. In UAV aerial photography tasks [1], image quality and target detection are affected by environmental changes (such as overexposure in strong light, low contrast in dim light, and extreme weather) and target characteristics (such as dynamic target movement, size variation, changes in perspective and altitude). Meanwhile, the uneven distribution of targets and occlusion phenomena also increase the complexity of feature extraction and recognition, reducing detection accuracy. In the field of deep learning, object detection algorithms are typically divided into two categories: Two-stage detection algorithms [2] and One-stage detection algorithms [3]. Two-stage algo-

rithms first generate candidate region proposals, then perform deep feature extraction, classification, and bounding box regression on them, ultimately achieving target recognition and localization. Typical Two-stage algorithms include SPPnet, Faster R-CNN, FPN, etc. In contrast, One-stage algorithms skip the candidate region generation step and directly complete the detection task by predicting target categories and locations, which is more efficient. Typical One-stage algorithms include SSD, RetinaNet, CenterNet, and the YOLO series of algorithms.

Dong [4] analyzed five reasons for the low detection accuracy of small objects and summarized small object detection methods from aspects such as multi-scale feature fusion, distortion evaluation metrics, super-resolution reconstruction, and lightweight network models. Liang [5] proposed a UAV image object detection algorithm based on the improved YOLOv7, which enhances the detection performance of small objects by adding a small object detection layer and incorporating a multi-information flow fusion attention mechanism. Liang [6] proposed the TS-YOLO object detection algorithm, an improved version of YOLOv8, with an efficient feature extraction module in the backbone network and a dual cross-scale weighted feature fusion structure in the neck network to enhance feature representation. Li [7] designed a receptive field convolution block attention module to optimize the YOLOv8 backbone network, alleviating the problem of spatial information sparsity caused by down sampling and improving feature extraction efficiency. Additionally, by fully utilizing large-scale features, increasing multi-scale feature fusion, and employing dynamic upsampling, a pyramid neck network that balances spatial and semantic information fusion was designed to effectively enhance the feature information of small objects.

This paper, based on the YOLOv11 algorithm, proposes an enhanced UAV image object detection algorithm through improvements to specific modules. The specific enhancements are as follows: In the backbone component, the EfficiBackbone module was designed, significantly improving the model's global perception and long-range dependency modeling capabilities; meanwhile, the C3k2 was optimized using the RipViT block, efficiently fusing features from different levels and further increasing the efficiency of feature fusion. In the detection head component, the Dynamic Head replaced the traditional detection head, enabling the model to more flexibly handle complex scenes and diverse targets, and significantly enhancing the detection capabilities for small, dense, and occluded targets. Through these improvements, the algorithm proposed in this paper has demonstrated greater robustness and higher accuracy in UAV image object detection.

2. Improved YOLOv11 Object Detection Algorithm

YOLOv11 [8] is the latest object detection algorithm released by Ultralytics on September 30, 2024. Compared to previous models in the YOLO series, YOLOv11 has achieved significant improvements in both accuracy and speed, as shown in **Figure 1**. The model mainly consists of a backbone network, a neck network, and a detection head.
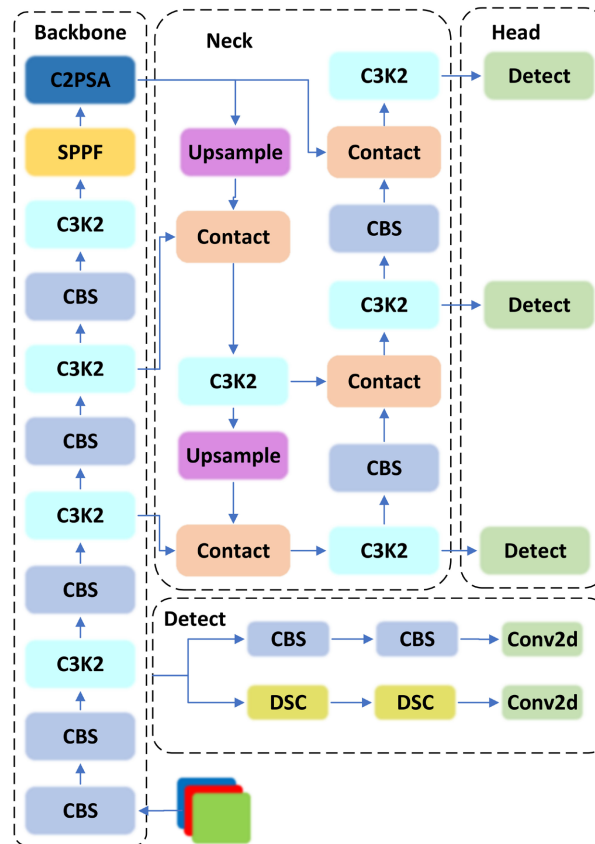
Figure 1. Structure of YOLOv11.

YOLOv11 uses an improved version of CSPDarknet53 [9] as its backbone network, generating feature maps of different scales through five downsamplings. On this basis, the backbone network replaces the original C2f module with the C3K2 module and introduces the CBS module (Convolution, Batch Normalization, and SiLU activation function) for processing, while also enhancing feature diversity using the Spatial Pyramid Pooling Fast (SPFF) module. The C2PSA module incorporates a design with a pyramid slice attention mechanism, further improving feature extraction capabilities. The neck network adopts a PAN-FPN structure [10], fusing shallow location information with deep semantic information through a bottom-up path, effectively addressing the shortcomings of object localization in the FPN structure.

The detection head of YOLOv11 uses a decoupled structure, with separate branches for predicting category and location information, and selects appropriate loss functions based on the task: Binary Cross-Entropy Loss (BCE Loss) is used for classification tasks.

Although the YOLOv11 detector demonstrates relatively high detection accuracy on general datasets, drone imagery, characterized by long shooting distances and wide fields of view, often contains a high proportion of small targets, complex backgrounds, dense target distributions, and frequent target occlusions. These factors lead to missed and false detections of small, distant targets by the YOLOv11

model. To address the limitations of the original YOLOv11n model in understanding complex, long-range relationships in drone-captured image data, we optimized the backbone network, improved the feature fusion mechanism, and incorporated an efficient detection head. Taking into account both model performance and resource consumption, the improved network structure is shown in **Figure 2**.
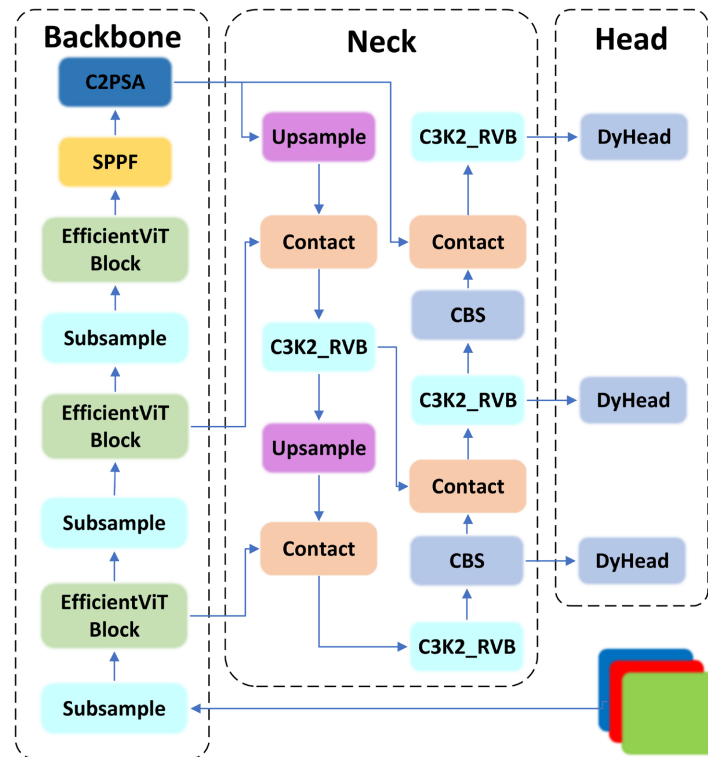
Figure 2. The design of the improved network structure.

2.1. Designing EffiBackbone

The EffiBackbone, by combining the strengths of Transformer and convolutional networks, is able to more effectively capture global information during the feature extraction phase. This enhances detection accuracy in scenes with complex backgrounds or dense targets, and has shown considerable potential particularly in the detection of small and distant targets. For the issues of complex backgrounds and multi-target detection in object recognition tasks, the EffiBackbone utilizes the self-attention mechanism to model the long-range dependencies in images, enabling the model to learn the relationships between distant pixels.

As shown in **Figure 3**, in this study, we constructed the backbone part of the model using three-layer EfficientViT Blocks [11]. The cascaded group attention mechanism in the structure of this module is capable of jointly capturing local and global features, enhancing the ability of feature representation. The specific calculation process is as follows:

$$X_{ij} = \text{Attention}\left(X_{ij}W_{ij}^Q, X_{ij}W_{ij}^K, X_{ij}W_{ij}^V \right), \tag{1}$$

$$X_{i+1} = \text{Concat}\left[X_{ij} \right]_{j\ =\ 1:h} W_i^P \tag{2}$$

X_{ij} denotes the j-th attention head of the input features, where $X_i = \left[X_{i1}, X_{i2}, \cdots, X_{ih} \right]$ and $1 \le j \le h$. $W_{ij}^Q, W_{ij}^K, W_{ij}^V$ are the projection layers that split the input features into different subspaces. W_i^P is a linear layer that projects the concatenated output features back to the dimension consistent with the input. The computation of feature splitting can effectively save computational costs. In the cascaded attention structure, the Q, K, V layers can learn richer information feature projections. The output of the previous attention head is passed as input to the next head through a cascaded approach, enabling self-attention to simultaneously capture both local and global relationships. The cascaded design can increase network depth and enhance model capacity without introducing additional parameters.
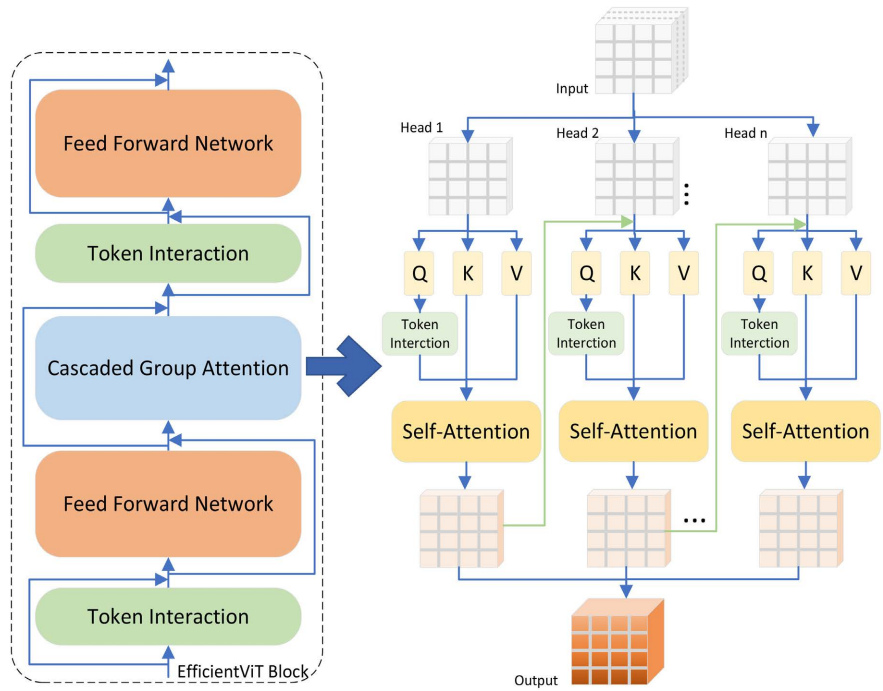
Figure 3. Structure of EfficiBackbone.

In the EfficiBackbone module, the number of attention heads is a key hyperparameter. We conducted experiments with varying numbers of attention heads, and the results are shown in **Table 1**. The experiments demonstrated that as the number of attention heads increased, both the mAP@0.5 and mAP@0.5:0.95 metrics of the model improved, but the computational load also increased. When the number of attention heads was set to 3, the model achieved a good balance between performance and computational load.

Moreover, each attention layer in the model is connected to a feed-forward neural network (FFN) layer. The FFN layer introduces non-linear transformations to

further enhance the model's expression ability, helping it better understand local and global structures in images.

Table 1. The effect of detection heads with different numbers.

Number of attention heads	mAP@0.5	mAP@0.5:0.95	GFLOPs
1	0.287	0.164	6.7
2	0.317	0.180	7. 2
3	0.358	0.209	7.9
4	0.360	0.211	11.6

2.2. Improved C3K2 Module

In YOLOv11, the C3k2 module faces certain challenges in feature extraction capabilities, especially in the task of small object detection in drone imagery, where its performance is limited. To improve the accuracy of small object detection, the RepViT Block [12] has been introduced and applied to the C3k2 module, forming the improved C3k2-RVB module (as shown in **Figure 4**). This module combines a spatial attention mechanism, which can adaptively adjust the size of the receptive field, thereby enhancing the network's processing capabilities for features at different scales.
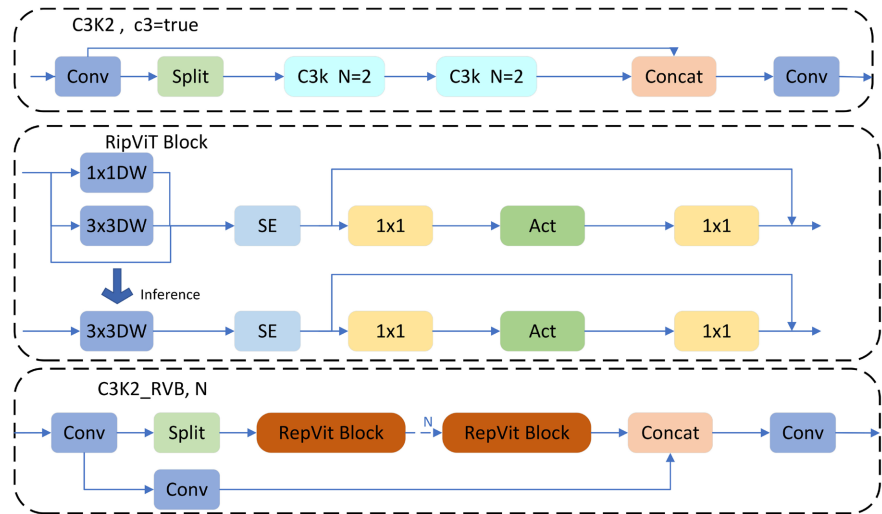
Figure 4. Structure of C3K2, RipViT Block, C3K2 RVB.

The RepViT Block is an efficient neural network module that combines convolutional networks and self-attention mechanisms. It introduces structural re-parameterization technology, using a multi-branch structure (such as convolution and skip connections) during the training phase to enhance feature extraction capabilities, and simplifying to a single-branch structure during the inference phase to significantly improve computational efficiency. This design fully combines the following two advantages:

- Convolutional networks: Good at capturing local features such as edges, tex-

tures, and small objects.

- Self-attention mechanisms: Modeling dependencies between distant pixels, capturing global information, and performing well in understanding complex backgrounds and relationships between objects.

In the C3k2-RVB module, two parallel depthwise separable convolution layers (3×3 DW and 1×1 DW) provide a multi-branch structure during the training phase to enrich feature extraction capabilities. During the inference phase, these branches are simplified to a single 3×3 DW convolution layer through re-parameterization technology, effectively reducing computational overhead. Additionally, a self-attention module (SE) is introduced to enhance global feature modeling capabilities, while a 1×1 convolution layer is used to adjust the number of feature channels and fuse information. Finally, during the inference phase, the structure is simplified to a 3×3 DW convolution layer followed by an SE module, and then through two 1×1 convolution layers to improve efficiency and feature fusion capabilities.

This improved design enhances the model's detection accuracy for small targets in complex scenarios while reducing computation costs, making it more suitable for efficient inference in resource-constrained environments.

2.3. Improved Detection Head

YOLOv11's detection head faces challenges in handling multi-size targets, particularly in balancing the detection performance of small and large targets. Additionally, accurately locating targets becomes more difficult in complex backgrounds or when targets are occluded, The original detection head lacked the necessary flexibility to meet the diverse needs of tasks, limiting its dynamic learning ability, especially in small target detection tasks for UAVs.

To overcome these limitations, this study introduces the Dynamic Head to replace the traditional detection head, The Dynamic Head [13] adjusts the prediction strategy dynamically, significantly improving small target detection capabilities and enhancing the model's adaptability to multi-scale targets, as shown in **Figure 5**. Additionally, it optimizes the feature fusion process, improving the combination of features from different levels. The computational process can be formalized as follows:

$$W(F) = \pi_C \left(\pi_S \left(\pi_L (F) \cdot F \right) \cdot F \right) \cdot F$$

Given a three-dimensional feature tensor $F \in R^{L \times S \times C}$, π_C, π_S, π_L, attention functions are applied across channel, spatial, and horizontal dimensions. These attention mechanisms are applied in sequence to the detection head, and through multiple layers, the feature representation ability is further enhanced.

This method not only reduces computational complexity, improving the model's real-time performance, but also significantly enhances the model's robustness and generalization ability in complex scenes and environmental changes. Additionally, it maintains high computational efficiency, which is crucial for applications that require rapid response, such as UAV small target detection.
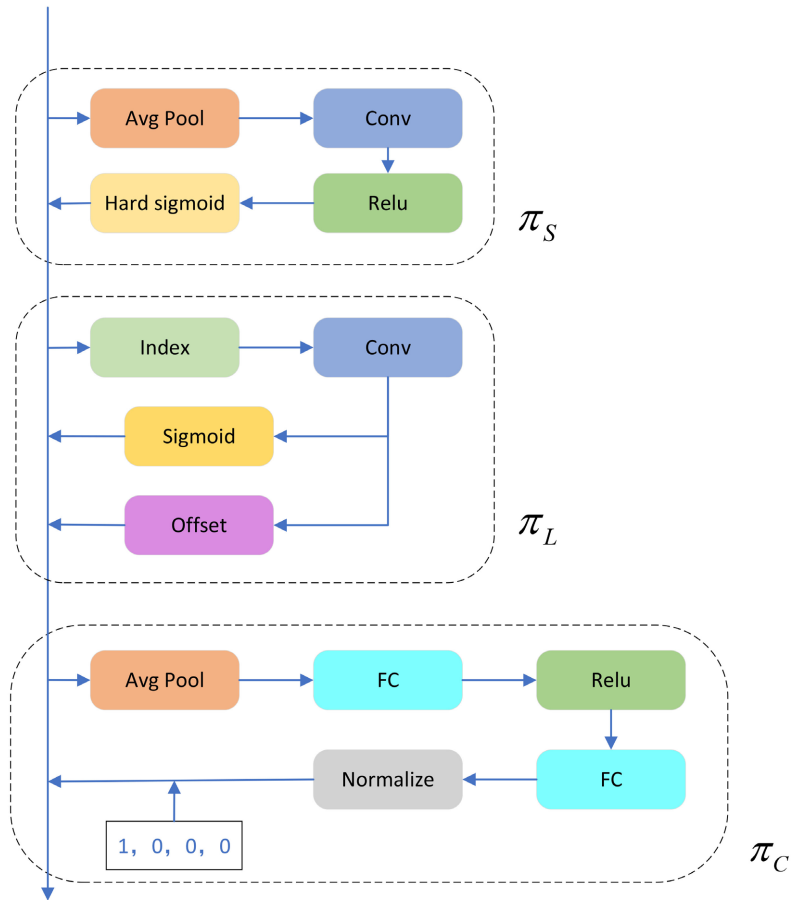
Figure 5. Structure of Dynamic Head (Dy Head).

3. Results and Discussion

3.1. Datasets and Experimental Environment

The VisDrone 2019 dataset [14], collected by the AlSKYEYE team from the Machine Learning and Data Mining Laboratory of Tianjin University, is a benchmark dataset for UAV vision tasks. It contains 288 video clips (261,908 frames) and 10,209 static images, all captured by different UAV camera models. The dataset covers diverse scenes from 14 cities in China, including urban and rural environments, with various targets such as pedestrians, vehicles, and bicycles, and densities ranging from sparse to crowded. The data was collected under different weather and lighting conditions and carefully labeled, with over 2.6 million target bounding boxes. It is one of the recognized benchmark datasets for remote sensing target detection models.

The experimental platform used Ubuntu20.04 as the operating system, equipped with an Intel(R) Core(TM) i7-13700 CPU @ 2.90 GHz, 32 GB of RAM, and an Nvidia GeForce RTX 4060Ti (16 GB) GPU. The PyTorch framework version was 2.4.0 + cu119 with Python version 3.8.19. The batch size was set to 16, image size to 640×640, and epochs to 300. No pre-trained model weights were used, and all experiments were trained with consistent hyperparameters.

3.2. Model Evaluation Metrics

In UAV small target detection tasks, false positives and false negatives are prominent issues. The mAP@0.5 and mAP@0.5:0.95 are commonly used as the main evaluation standards. mAP combines the model's precision and recall, making it widely regarded as a comprehensive performance metric.

1) Mean Average Precision (mAP)

mAP@0.5 [15] represents the average detection precision for all categories at an loU threshold of 0.5. mAP@0.5:0.95 calculates the average precision at different loU thresholds ranging from 0.5 to 0.95. In object detection tasks, the higher the mAP value, the better the model performance. The formula is as follows:

$$mAP = \frac{1}{N} \sum_{i=0}^{N} AP_i$$

where AP, is the average precision for class *i*, and *N* is the total number of classes.

2) Precision

Precision is defined as the ratio of correctly detected objects (True Positives, TP) to the total number of detected objects. Its calculation formula is:

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

where TP represents true positives (correctly detected objects), and FP represents false positives (incorrectly detected objects).

3) Recall

Recall is defined as the ratio of the number of correctly detected objects to the total number of actual objects present. The calculation formula is:

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

4) Giga Floating Point Operations Per Second (GFLOPs)

GFLOPs [16] represent the number of floating-point operations executed by the model per second. It is typically used to evaluate the model's computational-complexity and execution efficiency.

3.3. Comparative and Ablation Experiments

Table 2 shows the detection results of popular algorithms in recent years on the VisDrone validation set, Considering the limitations of UAV hardware, the model needs to ensure that the number of parameters and computational cost are kept as low as possible while maintaining high accuracy. From **Table 1**, it can be seen that the improved algorithm ranks among the top in terms of detection performance, while having a lower computational cost, its mAP@0.5 and mAP@0.5:0.95 reached39.7% and 23.6%, respectively, which are improvements of 6.5% and 4.3% compared to the baseline YOLOv11n model. It also outperforms algorithms such as uYOLOv5sYOLOv8s, and YOLOv10s. The improved model has a 4% lower mAP@0.5:0.95 compared to the ACAM-YOLO model, but it is ten times more efficient in terms of computational efficiency. This experiment further demon-

strates that the improved model effectively balances performance, parameters, and computational cost, and outperforms the most common algorithms, indicating its good design value.

Table 2. Experimental results on VisDrone2019-val.

Method	P	R	mAP@0.5	mAP@0.5:0.95	GFLOPs
Faster R-CNN	0.313	0.346	0.289	0.122	350
SSD	0.203	0.338	0.220	0.109	61.3
YOLOv5s [17]	0.430	0.325	0.320	0.185	4.7
YOLOv6s	0.404	0.404	0.298	0.174	8.6
YOLOv8s [18]	0.441	0.327	0.327	0.190	5.6
YOLOv11n	0.457	0.329	0.332	0.193	6.7
ACAM-YOLO [19]	--	--	0.495	0.225	130.8
Ours	0.493	0.373	0.397	0.236	12.6

As shown in Table 3, to comprehensively verify the performance improvement of each modification module, ablation experiments were conducted on the proposed improved model using the VisDrone2019 dataset. These experiments were based on the YOLOv11n model, progressively integrating the improvements to quantify the contribution of each module to detection accuracy and efficiency. The experimental results indicate that each modification module significantly enhances the model's ability to detect small targets.

Table 3. Ablation experiment result in VisDrone2019-val.

Baseline	EfficientViT	DyHead	C3K2_RVB	P	R	mAP@0.5	mAP@0.5:0.95	GFLOPs
√				0.457	0.329	0.332	0.193	6.7
√	√			0.464	0.464	0.358	0.209	7.9
√	√	√		0.475	0.366	0.375	0.223	9.1
√	√	√	√	0.493	0.373	0.397	0.236	12.6

4. Conclusions

To enhance the detection performance of small objects in UAV aerial images, an improved model based on YOLOv11n has been proposed. This model introduces EfficientViT to replace the original CSPDarknet53. EfficientViT models long-range dependencies in images through self-attention mechanisms, enabling the model to learn the relationships between distant pixels. During the feature extraction phase, EfficientViT can more effectively capture global information, reducing the loss of detail information and thus improving the flexibility and computational efficiency of feature extraction. This improvement enhances detection accuracy in complex background or dense target scenarios, particularly in the detection of small and distant objects. Furthermore, to address the shortcomings of the C3k2 module in small object feature extraction, the RipViT Block has been

proposed for its improvement. This module combines a spatial attention mechanism, which can adaptively adjust the size of the receptive field, thereby enhancing the network's perception of features at different scales. Based on this, the DyHead structure was further introduced, integrating scale-aware, spatial-aware, and task-aware attention mechanisms to further enhance the representation capability of small object features.

Through these innovative improvements, the proposed model has significantly enhanced detection performance in small object detection tasks of UAV aerial images. Experimental results on the VisDrone dataset show that the improved model achieved 39.7% on the mAP@0.5 metric, outperforming most mainstream models and significantly higher in accuracy than YOLOv11n. Meanwhile, the computational load of the improved model has slightly increased, with a computation amount of 12.6, which proves the effectiveness and high efficiency of the model in UAV small object detection tasks. Looking ahead to future research, we will continue to optimize the model structure to further reduce its complexity, aiming to achieve more efficient deployment and application of UAV small object detection on edge devices.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Mohsan, S.A.H., Khan, M.A., Noor, F., Ullah, I. and Alsharif, M.H. (2022) Towards the Unmanned Aerial Vehicles (UAVs): A Comprehensive Review. *Drones*, **6**, Article No. 147. https://doi.org/10.3390/drones6060147

[2] Zhou, X., Koltun, V. and Krähenbühl, P. (2021) Probabilistic Two-Stage Detection.

[3] Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D. and Luo, J. (2019). A Fast and Accurate One-Stage Approach to Visual Grounding. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 4682-4692. https://doi.org/10.1109/iccv.2019.00478

[4] Gang D, Weicheng X, Xiaolong H, *et al.* (2023) Review of Small Object Detection Algorithms Based on Deep Learning. *Journal of Computer Engineering & Applications*, **59**.

[5] Zeng, Y., Zhang, T., He, W. and Zhang, Z. (2023) Yolov7-UAV: An Unmanned Aerial Vehicle Image Object Detection Algorithm Based on Improved Yolov7. *Electronics*, **12**, Article No. 3141. https://doi.org/10.3390/electronics12143141

[6] Li, Y., Fan, Q., Huang, H., Han, Z. and Gu, Q. (2023) A Modified Yolov8 Detection Network for UAV Aerial Image Recognition. *Drones*, **7**, Article No. 304. https://doi.org/10.3390/drones7050304

[7] Li, Y., Li, Q., Pan, J., Zhou, Y., Zhu, H., Wei, H., *et al.* (2024) SOD-YOLO: Small-Object-Detection Algorithm Based on Improved YOLOv8 for UAV Images. *Remote Sensing*, **16**, Article No. 3057. https://doi.org/10.3390/rs16163057

[8] Khanam, R. and Hussain, M. (2024) Yolov11: An Overview of the Key Architectural Enhancements.

[9] Mahasin, M. and Dewi, I.A. (2022) Comparison of CSPDarkNet53, CSPResNeXt-50,

and EfficientNet-B0 Backbones on YOLO v4 as Object Detector. *International Journal of Engineering, Science and Information Technology*, **2**, 64-72.
https://doi.org/10.52088/ijesty.v2i3.291

[10] Wang, G., Chen, Y., An, P., Hong, H., Hu, J. and Huang, T. (2023) UAV-Yolov8: A Small-Object-Detection Model Based on Improved Yolov8 for UAV Aerial Photography Scenarios. *Sensors*, **23**, Article No. 7190. https://doi.org/10.3390/s23167190

[11] Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H. and Yuan, Y. (2023) EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 14420-14430. https://doi.org/10.1109/cvpr52729.2023.01386

[12] Wang, A., Chen, H., Lin, Z., Han, J. and Ding, G. (2024) Rep ViT: Revisiting Mobile CNN From ViT Perspective. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 15909-15920.
https://doi.org/10.1109/cvpr52733.2024.01506

[13] Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., *et al.* (2021) Dynamic Head: Unifying Object Detection Heads with Attentions. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 7369-7378.
https://doi.org/10.1109/cvpr46437.2021.00729

[14] Du, D., Zhu, P., Wen, L., *et al.* (2019) VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

[15] Yue, Y., Finley, T., Radlinski, F. and Joachims, T. (2007) A Support Vector Method for Optimizing Average Precision. *Proceedings of the* 30*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, 23-27 July 2007, 271-278. https://doi.org/10.1145/1277741.1277790

[16] Lee, Y., Waterman, A., Avizienis, R., Cook, H., Sun, C., Stojanovic, V., *et al.* (2014) A 45 nm 1.3 GHz 16.7 Double-Precision GFLOPS/W RISC-V Processor with Vector Accelerators. ESSCIRC 2014 40th European Solid State Circuits Conference (*ESSCIRC*), Venice Lido, 22-26 September 2014, 199-202.
https://doi.org/10.1109/esscirc.2014.6942056

[17] Jocher, G., Stoken, A., Chaurasia, A., *et al.* (2021) Ultralytics/Yolov5: v6.0-YOLOv5n "Nano" Models, Roboflow Integration, TensorFlow Export, OpenCV DNN Support. Zenodo.

[18] Chu, X., Zheng, A., Zhang, X. and Sun, J. (2020) Detection in Crowded Scenes: One Proposal, Multiple Predictions. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 12211-12220.
https://doi.org/10.1109/cvpr42600.2020.01223

[19] Li, Z., Wang, Z. and He, Y. (2023) Aerial Photography Dense Small Target Detection Algorithm Based on Adaptive Collaborative Attention Mechanism. *J Aeronaut*, **10**, 1-12.