International Journal of Computer Applications Technology and Research
Volume 14–Issue 04, 137 – 141, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1404.1011

YOLOv9 Algorithm Improvement for Small Object Detection in UAV Aerial Imagery

Guoliang Xiong
College of Communication Engineering, Chengdu University of Information Technology, Chengdu, China

Yuxiang Gao
College of Communication Engineering, Chengdu University of Information Technology, Chengdu, China

Yuxuan Liao
College of Communication Engineering, Chengdu University of Information Technology, Chengdu, China

Abstract: This paper addresses the challenges of small object detection in UAV aerial imagery by proposing an improved YOLOv9 object detection model. The core innovations are twofold: first, the introduction of a Global Attention Mechanism (GAM), which enhances the model's perception of small object features through channel and spatial dual-path attention processing; second, the adoption of a Bidirectional Feature Pyramid Network (BiFPN), which implements bidirectional feature fusion from top-down and bottom-up perspectives, effectively improving the interaction efficiency of features at different scales. Experiments on the VisDrone2019 dataset demonstrate that, compared to the baseline YOLOv9, the proposed model improves the mAP@0.5 metric by 1.6%, while reducing the parameter count by 2.8×10^6. Visual comparisons show that the model exhibits superior detection capabilities in complex environments such as aerial multi-scale small objects, dense crowds, and night scenes, effectively addressing the problems of missed detections and false detections of small objects.

Keywords: aerial imagery ; object detection; feature fusion; attention mechanism; bidirectional feature pyramid network

1. Introduction

In recent years, with the rapid development of Unmanned Aerial Vehicle (UAV) technology, drones have been increasingly applied across military, civilian, and commercial sectors. Particularly in industries such as agriculture, power grid inspection, and urban surveillance, UAVs have become essential tools for modern monitoring and detection due to their lightweight, flexible, and efficient characteristics. However, object detection in UAV aerial imagery, especially small object detection, still faces significant challenges. This is mainly because UAV aerial images are captured from high angles, resulting in small-scale target objects and complex, variable backgrounds. Small targets such as pedestrians and bicycles are often difficult for conventional object detection algorithms to accurately identify due to their small scale or background interference.

Object detection, as a critical task in computer vision, aims to automatically recognize and locate specific objects in images. Traditional object detection algorithms typically rely on manually designed features and classifiers, but they adapt poorly to small objects and struggle with effective recognition in complex backgrounds. While deep learning-based algorithms have made some progress, they still face issues such as high computational complexity, insufficient detection accuracy, and poor real-time performance.

Currently, deep learning object detection algorithms are mainly divided into two categories: two-stage algorithms and one-stage algorithms. Two-stage algorithms, such as R-CNN[1], Fast R-CNN[2], and Faster R-CNN[3], though highly accurate, have slower detection speeds due to multiple feature extraction requirements, making it difficult to meet real-time demands. In contrast, one-stage methods like the YOLO series[4] offer advantages in detection speed and real-time performance but may lack in accuracy. To improve small object detection accuracy, many researchers have optimized existing YOLO models. For instance, to enhance network sensitivity to small objects, many studies have introduced multi-scale feature fusion and attention mechanisms, such as Cao et al.[10], who improved small object detection accuracy by adding multi-scale feature fusion modules, and Yi et al.[11], who enhanced small object feature extraction capabilities through attention mechanisms. However, although these methods effectively improve detection accuracy, they often come with increased computational complexity and decreased real-time performance. Additionally, complex feature pyramid and network structure optimization methods also risk overfitting and insufficient generalization ability. Zhang et al.[12] improved feature fusion based on YOLOv8 using the PAFPN structure, yet this structure still fails to effectively solve the missed detection problem when facing complex backgrounds and small objects. How to improve real-time performance while ensuring accuracy, and effectively solve the small object detection problem, remains a challenging issue that urgently needs to be addressed in the field of object detection.

2. Related Technologies and Theories

YOLOv9 consists of a Backbone, Neck, and Head network. The Backbone is responsible for feature extraction, extracting key information from images for use by subsequent networks. The Neck network is positioned between the Backbone and Head networks, using features extracted by the Backbone for feature fusion. The Head network then utilizes these fused features for object recognition. The input image size is 640×640×3. The Silence operation module, located in the first layer of the network, is a special module that performs no computation and produces output identical to the input, facilitating auxiliary branch calls to the original image input module.

The CBS module consists of Conv (Convolutional), BN (Batch Normalization), and SiLU (Sigmoid-Weighted Linear Unit), used to extract local features from images. The RepNCSPELAN4 module combines CSPNet and ELAN (Efficient Layer Aggregation Network) to enhance feature extraction capabilities by processing long-distance dependencies and global contextual information.

International Journal of Computer Applications Technology and Research
Volume 14–Issue 04, 137 – 141, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1404.1011

RepNCSPELAN4 mainly consists of convolution modules and RepNCSP. RepNCSP is composed of convolution modules and multiple RepNBottleneck modules, while RepNBottleneck is a reparameterized neck bottom module with a residual structure, the number of which is determined by the model's width factor. RepNBottleneck consists of RepConV and convolution blocks, where RepConV is a reparameterized convolution unit composed of convolution layers and the SiLU activation function.

The CBLiner module combines Conv, BN, and SiLU, and splits features obtained after one convolution into 1-N features, enabling reversible connections to enhance the network's information flow. SPPELAN combines SPP's spatial pyramid pooling concept with ELAN's global context information capture characteristics, performing pooling operations at different scales to capture feature information at various scales. The CBFuse module aims to fuse features from different layers or branches. Multi-level auxiliary information typically serves as an additional branch attached to YOLOv9's backbone network. This branch connects to multiple levels of the feature pyramid, allowing it to receive and transmit feature information at multiple scales.

3. Improved Methods

The improved model proposed in this paper consists of three parts: Backbone, Neck, and Head, as shown in Figure 1. The Backbone is responsible for feature information extraction; the Neck enhances the features extracted by the backbone network, introduces multi-scale information, and improves the model's perception ability for small targets; the Head generates the final output for object detection. The main innovations of this paper are: introducing the Global Attention Mechanism (GAM) in the Backbone, and adopting the Bidirectional Feature Pyramid Network (BiFPN) in the Neck part.

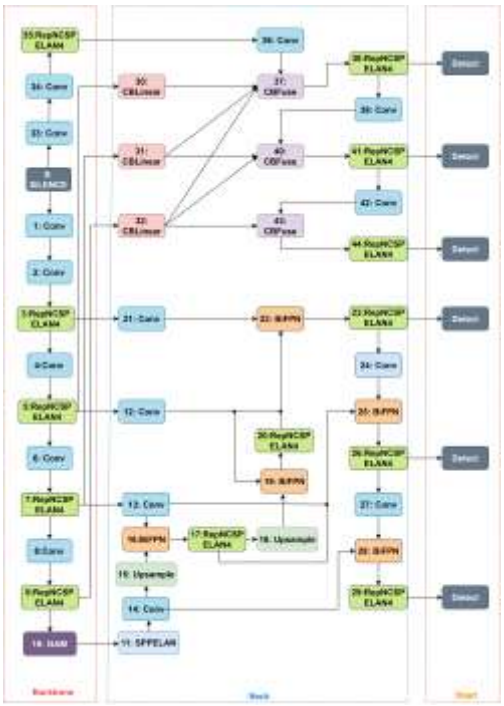

Figure 1 Network structure of improved model

3.1 Introduction of Global Attention Mechanism

In convolutional neural networks, the attention mechanism effectively enhances feature expression capabilities by dynamically adjusting channel weights. The Global Attention Mechanism (GAM) is a widely applied attention mechanism in the fields of computer vision and deep learning. The global attention mechanism consists of a channel sub-attention module and a spatial sub-attention module, as shown in Figure 2. Each sub-module processes the input feature map in its specific dimension.

The channel sub-attention module utilizes a three-dimensional arrangement to preserve information. This module is processed through a Multi-Layer Perceptron (MLP), aiming to strengthen the interaction of spatial information across different dimensions, thereby enhancing feature representation capabilities. The spatial sub-attention module primarily focuses on information in the spatial dimension. This module uses two convolutional layers to fuse spatial information. Through these convolutional layers, the spatial sub-attention module can weight features at different positions, capturing dependencies between spatial locations, thus better focusing on the importance of local regions.

Given that the input feature map is F_1, the intermediate state F_2 and output feature F_3 are defined as shown in Equation (1) and Equation (2):

$$F_2 = M_C F_1 \otimes F_1 \quad (1)$$

$$F_3 = M_S F_2 \otimes F_2 \quad (2)$$

Mc is the channel attention feature map, Ms is the spatial attention feature map, and ⊗ represents element-wise multiplication.

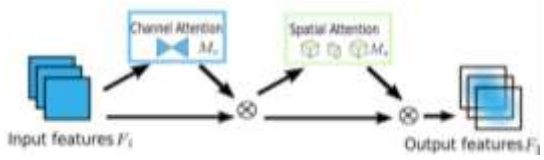

Figure 2 The structure figure of Global Attention Mechanism

3.2 Bidirectional Feature Pyramid Network

The Neck layer of the network has a significant impact on detection performance. The Bidirectional Feature Pyramid Network (BiFPN) is a new method of feature fusion, where "bidirectional" refers to feature fusion from top-down and bottom-up. Its concept involves efficient bidirectional cross-scale connections and weighted feature fusion, which means performing top-down feature fusion first, followed by bottom-up feature fusion on the basis of path enhancement.

Figure 3 shows four typical design forms of feature fusion networks, which to some extent also represent the development process of feature fusion networks. The earliest Neck part directly predicted from high-level pyramid features extracted from the Backbone, but this structure did not

International Journal of Computer Applications Technology and Research
Volume 14–Issue 04, 137 – 141, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1404.1011

perform feature fusion, resulting in relatively low accuracy. Tsung et al. [13] proposed the Feature Pyramid Network (FPN) to improve this. As shown in figure (a), FPN performs top-down feature fusion, and the feature layers with higher semantic information obtained after fusion are used for prediction, but it is limited by unidirectional information flow. The Path Aggregation Network (PAN) [14] solved this problem, as shown in figure (b), by adding an additional bottom-up path on the basis of FPN to realize bottom-up feature fusion, thus allowing bottom-layer position information to be sent to the prediction feature layer. NAS-FPN (Neural Architecture Search, NAS) [15]uses neural network search (NAS) to find irregular feature network topologies, as shown in figure (c), and then repeatedly applies the same block, but the searched network is irregular, difficult to explain and modify, and using NAS technology is very time and effort consuming.

This research uses the BiFPN structure as a method of feature fusion. Unlike the above three methods based on adjacent feature fusion, BiFPN adopts bidirectional cross-scale and weighted feature fusion to achieve higher-level feature fusion, as shown in figure (d). To better connect BiFPN with the YOLOv9 network structure, a convolutional layer with a size of 3×3, a stride of 2, and 256 channels is added when the Backbone inputs feature layers into the BiFPN.

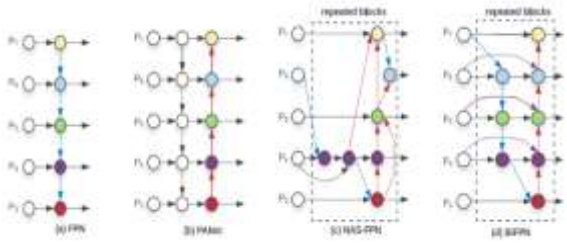

Figure 3 Typical feature fusion network

4. Experimental Results and Analysis

All training and validation experiments were implemented on a computer equipped with an NVIDIA GeForce RTX4060Ti with 12GB of video memory. The basic environment includes Python version 3.12, PyTorch version 1.13.1, and CUDA version 11.8. Uniform parameter settings were adopted in the experiments, as shown in Table 1.

Table 1 Experimental parameter setting

Parameter	Experimental Setting
Input image size	640×640 pixels
Initial learning rate (lr0)	0.01
Final learning rate (lrf)	0.01
Batch size	8
Learning rate momentum	0.937
Number of threads (workers)	8

4.1 Dataset Introduction

To demonstrate that the improved model has better universality, the VisDrone2019[16] dataset was selected as the experimental validation object. The VisDrone2019 dataset was collected using different drone models under various scenarios, weather conditions, and lighting conditions. It has been manually annotated with over 2.6 million bounding boxes or points of interest, mainly including ten categories: pedestrians, people, bicycles, cars, vans, trucks, tricycles, awning-tricycles, buses, and motorcycles. The distribution of instances across categories is shown in Figure 4. This dataset features complex backgrounds, imbalanced category distributions, crowded and dense targets, and contains numerous small objects, making it widely used for training and evaluating small object detection algorithms.

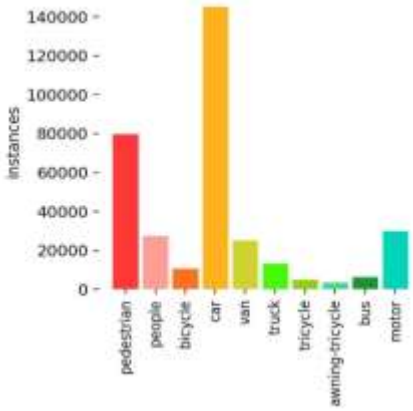

Figure 4 VisDrone2019 dataset

4.2 Evaluation Metrics

The evaluation metrics selected for this experiment include Precision (P), Recall (R), and mean Average Precision (mAP). The calculation formulas for Precision, Recall, and mAP are shown in Equations (3)~(6):

$$P = \frac{TP}{TP + FP} \times 100\% \quad (3)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

$$AP = \int_0^1 P(x)dx \quad (5)$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP(i) \quad (6)$$

TP (True Positive) represents correctly classified positive samples, FP (False Positive) represents incorrectly classified negative samples, FN (False Negative) represents incorrectly classified positive samples. n represents the number of detection target categories. mAP@0.5 reflects the overall accuracy of the algorithm, specifically representing the average precision of all detection target categories when the IOU threshold is set to 0.5.

International Journal of Computer Applications Technology and Research
Volume 14–Issue 04, 137 – 141, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1404.1011

4.3 Ablation Experiments

To verify the effectiveness of the improved algorithm based on the YOLOv9 baseline model, ablation experiments were conducted by adding modules individually and in different combinations. The detailed experimental results are shown in the figure, where A represents the use of the GAM attention mechanism, and B represents the use of BiFPN. The experimental results in Table 2 indicate that, first, after adding the GAM attention mechanism, mAP@0.5 increased by 0.8%, demonstrating that the GAM module can effectively enhance the model's focus on important features, thereby improving overall detection performance. Then, after adding the BiFPN module, mAP@0.5 increased by 1.2%, and this module improved the model's performance in multi-scale feature fusion, especially significantly enhancing the detection capability for small objects.

When the two modules were combined, the model's performance was further optimized. The GAM+BiFPN combination increased mAP@0.5 to 52.2%, with precision and recall also reaching optimal values. Additionally, the parameter count was reduced by 2.8×10^6 compared to the baseline model, improving computational efficiency and proving that the combination of multiple modules can effectively enhance YOLOv9's object detection performance.

Table 2 Results of ablation experiment

Model	GAM	BiFPN	P/%	R/%	mAP@0.5/%	Params/10^7
v9			59.8	48.1	50.6	6.08
A	☑		60.2	48.5	51.4	6.09
B		☑	61.4	48.7	51.8	5.82
C	☑	☑	62.6	49.3	52.2	5.80

4.4 Experimental Results and Analysis on VisDrone

To validate the detection effectiveness of the proposed algorithm in real-world scenarios, we selected challenging images from the VisDrone2019 test set for visualization. The detection effect comparison is shown in Figure 5, with the YOLOv9 algorithm on the left and the improved algorithm on the right. From the first and second rows of images, it can be observed that our algorithm demonstrates superior capability in detecting smaller targets in aerial multi-scale small objects and dense crowd scenarios, significantly reducing the miss detection rate and enhancing the model's anti-interference ability and feature extraction capability in complex environments. In the third row showing a nighttime high-altitude small object scenario, the improved algorithm not only successfully detected motorcycles at the intersection but also precisely identified more extremely small targets at long distances. These results indicate that the improved algorithm exhibits superior performance when facing complex scenes, extremely small targets, dense crowds, and large target detection scenarios, with significantly improved precision and greatly reduced false detection and miss detection rates.

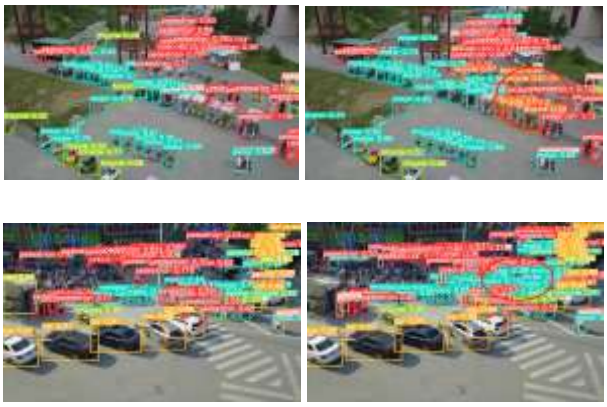

Figure 5 Comparison of VisDrone2019 detection performance

5. Conclusion

This paper proposes an improved algorithm based on global attention mechanism and bidirectional feature pyramid network, which effectively enhances the detection performance of small targets in UAV aerial images. By introducing GAM, the network's ability to focus on important feature channels and spatial positions is strengthened; adopting BiFPN achieves efficient interaction and fusion of features at different scales, further improving the detection capability for small targets. Experimental results validate the effectiveness of the proposed method, demonstrating superior performance especially when processing high-density small targets and complex environmental scenarios.

Future work will further explore how to reduce the computational complexity of the model and improve inference speed, making the algorithm more suitable for real-time application scenarios. Meanwhile, we will also investigate how to further optimize attention mechanisms and feature fusion strategies to better adapt to object detection tasks in various complex environments.

6. REFERENCES

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

[2] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[3] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.

[4] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[5] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Com puter Vision and Pattern Recognition, 2017: 7263-7271.

[6] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arXiv:1804.02767, 2018.

[7] BOCHKOVSKIY A, WANG C Y, LIAO H. YOLOv4: optimal speed and accuracy of object detection[J]. arXiv: 2004.10934, 2020.

[8] LI C Y, LI L, JIANG H L, et al. YOLOv6: a single-stage object detection framework for industrial applications[J]. arXiv:2209.02976, 2022.

[9] Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2024: 1-21.

[10] Cao S, Wang T, Li T, et al. UAV small target detection algorithm based on an improved YOLOv5s model[J]. Journal of Visual Communication and Image Representation, 2023, 97: 103936.

[11] Yi W, Wang B. Research on underwater small target detection algorithm based on improved YOLOv7[J]. IEEE Access, 2023, 11: 66818-66827.

[12] Zhang Z. Drone-YOLO: An efficient neural network method for target detection in drone images[J]. Drones, 2023, 7(8): 526.

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

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

[15] Elsken T, Metzen J H, Hutter F. Neural architecture search: A survey[J]. Journal of Machine Learning Research, 2019, 20(55): 1-21.

[16] Du D, Zhu P, Wen L, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results[C]//Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019: 0-0.