Based On the Improved Yolov7 Unmanned Aerial Vehicle (UAV) Aerial Image Small Target Detection Algorithm

Yu Zhou

Abstract— **In recent years, the detection of small targets in fields such as drone aerial photography and industrial quality inspection has faced challenges such as sparse target pixels, large background noise, and drastic scale changes. Traditional algorithms have problems such as weak feature extraction ability, low multi-scale fusion efficiency, and high computational resource consumption, resulting in insufficient detection accuracy. Specifically, shallow features are prone to losing details, the feature pyramid is insensitive to small targets, and complex networks have difficulty balancing accuracy and real-time performance. To address these issues, this paper proposes an improved algorithm and lightweight solution based on YOLOv7, enhancing detection performance from three aspects: feature enhancement, network optimization, and loss function. This paper designs the CBFR module to dynamically adjust the fusion weights of deep and shallow features, optimizes the detection head structure and adds a small target detection layer to enhance the ability to capture local details; combines the bidirectional routing attention mechanism to focus on key feature regions, and proposes a NWD and IOU composite loss function to balance sample weights. Extensive experiments were conducted on the VisDrone dataset. The experiments show that the improved model maintains the advantages of lightweight while significantly improving detection accuracy and recall rate.**

Key words: Attention mechanism, Unmanned Aerial Vehicle (UAV) aerial images, YOLOv7, Small Object Detection,FReLU

I. INTRODUCTION

With the deep integration of unmanned aerial vehicle (UAV) technology and computer vision, the importance of aerial image target detection in fields such as disaster rescue and environmental monitoring is increasingly prominent. As the core link of the unmanned aerial vehicle (UAV) intelligent perception system, target detection needs to accurately identify small targets (such as vehicles and pedestrians) in the image and overcome challenges such as complex backgrounds and variable scales to support real-time decision-making and task planning. Unmanned aerial vehicle (UAV) aerial photography can obtain large-area and high-resolution images, providing rich data support for various tasks. When processing and analyzing these images, object detection is one of the key links. It can help identify the objects of interest from complex aerial photography scenes, thereby providing a basis for subsequent decision-making.

Manuscript received April 18, 2025
Yu Zhou,School of computer science and technology, Tiangong University, Tianjin, China

At present, object detection algorithms based on deep learning dominate in the analysis of unmanned aerial vehicle (UAV) aerial photography images, among which the YOLOv7 algorithm shows significant performance advantages. YOLOv7 adopts a more efficient network architecture and advanced training strategies, significantly enhancing the detection speed while ensuring detection accuracy, and meeting the real-time processing requirements of unmanned aerial vehicle (UAV) aerial images. For example, its optimization in multi-scale feature fusion enables the model to better capture the features of targets of different sizes and detect various types of targets more accurately in complex aerial photography scenes.

However, when dealing with aerial images taken by drones, YOLOv7 also faces some challenges. Due to the complexity of aerial image scenes, which contain a large amount of background information, and the diverse changes in the size, shape and posture of target objects, especially for some small targets and occluded targets, the detection accuracy of YOLOv7 will decline. In addition, possible factors such as illumination changes and weather influences in aerial images can also cause certain interferences to the detection performance.

In recent years, significant progress has been made in the research of small target detection algorithms for unmanned aerial vehicle (UAV) aerial images. By introducing the attention mechanism, feature pyramid network, new loss function and lightweight network structure, the detection accuracy and real-time performance of the model for small targets have been significantly improved. The development of these technologies provides strong technical support for the application of unmanned aerial vehicles in fields such as military, transportation, and disaster prevention and control. However, small object detection still faces many problems, including limited feature information due to the small size of the object, which is difficult to extract effectively; The shape and edge information is not clear and is easily confused with the background. The limited internal features lead to the loss of a large amount of context semantic information in the feature extraction stage of the network. Moreover, the imbalance of positive and negative samples during the training process increases the difficulty of model training. Coupled with the problems such as high computational overhead and complex data annotation faced by existing methods, future research needs to further optimize the algorithm performance, reduce the computational cost, and explore more efficient small object detection strategies.

This paper aims to improve the YOLOv7 algorithm based on the characteristics of unmanned aerial vehicle (UAV) aerial photography images, in order to enhance its

detection accuracy for various targets, especially small targets and occlared targets, in complex aerial photography scenes, while maintaining the reasoning speed of the model as much as possible. This paper proposes an improved YOLOv7 small target detection method aiming at the problems existing in small target detection, such as insufficient feature extraction ability, poor multi-scale feature fusion effect, and limited improvement of the positioning accuracy of small targets by the loss function. Its main contributions are as follows:

1. This paper designs an improved CBFR module and introduces the Funnel activation function to enhance the feature extraction ability of small targets. The Funnel activation function, with its unique structure, can capture the characteristic information of small targets more accurately and effectively improve the detection accuracy.
2. This paper improves the original FPN+PAN structure and proposes a new multi-scale feature fusion method to construct a multi-scale object detection head. Through this method, the model can efficiently utilize feature information of different scales and significantly enhance the detection ability for small targets. Introduce the BiFormer bidirectional routing attention mechanism to improve the detection accuracy of the model for small targets. BiFormer significantly optimizes the model detection performance by adaptively focusing on the key feature regions of small targets.
3. Based on the fusion loss function of NWD and IOU, this paper proposes a loss function that fuses NWD and IOU to improve the positioning accuracy of small targets. This loss function combines the advantages of both to achieve a balance between positioning accuracy and classification accuracy and improve the overall performance of the model.

II. Related Work

In the research of computer vision, small object detection is an extremely difficult technical challenge. Such targets usually refer to objects in images with a pixel size of less than 32×32, vehicles in drone aerial photography, early-stage lesions in medical images, or minor defects in industrial quality inspection. Their core challenge lies in the fact that the limited pixels are difficult to carry sufficient discriminative information such as texture and shape, while traditional detection models are prone to losing key details during the feature extraction process. When Faster R-CNN processes 4K resolution remote sensing images, the missed detection rate of small buildings with a resolution of less than 10 pixels is as high as 40%, which directly affects the application effect in key scenarios such as disaster monitoring. In the public test in 2024, the EfficientDet-Lite model equipped with Bi-FPN achieved 52.2% in the small object detection task (AP_S) on the COCO dataset [1]. In response to the demands of mobile and edge computing, VoVNet[2] proposed the One-time Aggregation (OSA) module, which combines dense connections and efficient feature reuse to reduce the computational load while maintaining accuracy. It is suitable for real-time small target detection. YOLObile[3] achieves a detection speed 7 times faster than YOLOv3 on mobile devices through the collaborative design of model compression and compilation, while maintaining high precision. It is suitable for drone and security scenarios.

Aiming at the problem of large target scale variation in aerial images, Feng et al. introduced the RepConv module, Transformer Encoder and BiFPN structure in the YOLOv5 framework. By enhancing the multi-scale feature fusion ability, the detection accuracy of small targets was significantly improved. And achieved average accuracies of 90.29% and 90.06% respectively on the HRSC2016 and UCAS-AOD datasets [4]. Furthermore, the RRNet model proposed by Chen et al., through the design of a hybrid detector, effectively dealt with the problems of dense targets and complex backgrounds in aerial images, and was listed as a representative method in the VisDrone Challenge [5]; To solve the problems of uncertainty in the target direction and discontinuity of the boundaries of the rotating box, the RSDet algorithm adopts the eight-parameter regression of the four corner points of the rotating box, avoiding the discontinuity of losses caused by the periodicity of angles [6]. S2A-Net (2021) combines the feature alignment module and the active rotation filter to generate high-quality rotation anchor boxes, enhancing the representation ability of direction-sensitive features [7]. When Wu et al. improved YOLOv5, they introduced the Circular Smooth Label, which effectively alleviated the boundary discontinuity problem of angular regression. Sommer et al. optimized the Region generation network (RPN) for small targets by adjusting the anchor box size and output resolution of Faster R-CNN, and verified its effectiveness on the unmanned aerial vehicle dataset [8]; The PPYOLOE model proposed by the domestic team is based on the free design of anchor boxes and combined with task alignment learning. It simplifies the deployment process while maintaining high accuracy and is applicable to the embedded platform of unmanned aerial vehicles [9].

III. Math

YOLOv7 [10] is an upgraded version of the YOLO (You Only Look Once) series of object detection models, with high accuracy and real-time performance as its core advantages. Its network architecture is based on the CSPDarknet feature extraction backbone, integrating the ELAN (High Efficiency Layer Aggregation Network) module and the SPP (Spatial Pyramid Pooling) module, significantly enhancing the multi-scale feature fusion capability; The training efficiency and generalization performance are optimized through dynamic label allocation and model scaling techniques, supporting flexible adaptation from lightweight to high-precision models. The network structure of YOLOv7 mainly consists of four parts: Input (input layer), Backbone (feature extraction network), Neck (feature fusion network), and Head (output layer). The overall structure diagram of YOLOv7 is shown in Figure 2-6. This article is based on YOLOv7 and involves a series of improvements.

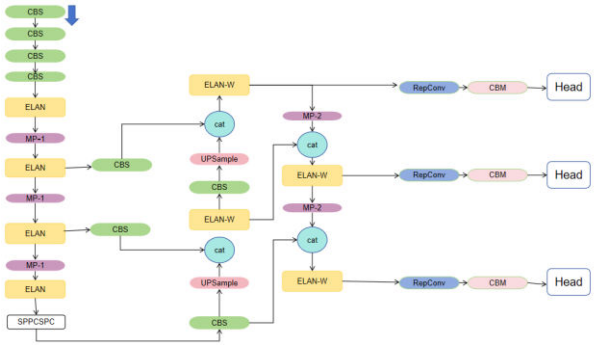

Fig 3-1 Overall structure diagram of YOLOv7

A. CBFR module

FReLU[11] is an improved activation function aimed at enhancing the flexibility and expressiveness of the ReLU activation function Q. Unlike the single form of the ReLU function, FReLU achieves richer nonlinear characteristics by introducing adjustable parameters or conditions. Specifically, FReLU can be defined and implemented in multiple ways:

1.Piecewise linear functions :FReLU may represent different input ranges through different linear segments, thereby increasing the complexity of the function. For example, the form of FReLU can be defined as formula (3-1):

$$f(x) = \begin{cases} \alpha_1 x + \beta_1 & \text{if } x \geq 0 \text{ and } x < t_1 \\ \alpha_2 x + \beta_2 & \text{if } x \geq t_1 \text{ and } x < t_2 \\ \vdots & \\ \alpha_n x + \beta_n & \text{if } x \geq t_{n-1} \end{cases} \quad (3\text{-}1)$$

Among them, is an adjustable parameter. This form allows FReLU to use different linear functions within different input ranges, thereby enhancing the flexibility of the model.

2.Adjustable parameters: Adjustable parameters can be introduced into the FReLU function, and these parameters can be adjusted through optimization algorithms during the training process. For example, FReLU can adjust the slope of the non-negative part through the parameter γ :

$$f(x) = \max(\gamma x, \beta x) \quad (3\text{-}2)$$

Among them, γ and β are hyperparameters. Usually, $\beta \leq 0$ to avoid an output that is completely zero.

3.Nonlinear extension: FReLU may adopt more complex nonlinear transformations, such as exponential functions or logarithmic functions, to enhance its expressive power. Formula 3-3:

$$\begin{cases} \alpha e^{\beta x} - 1 & \text{if } x \geq 0 \\ \alpha \log(1 + e^{\beta x}) & \text{if } x < 0 \end{cases} \quad (3\text{-}3)$$

α and β are adjustable parameters. The introduction of exponential or logarithmic functions enables FReLU to handle more complex input patterns.

In this paper, by replacing the SiLU activation function [45] used by the CBS module in the backbone network of the YOLOv7 algorithm with the FReLU activation function, the algorithm has the ability to obtain spatial context information and pixel-level modeling ability with almost no increase in computational overhead. The structure diagram of the improved CBFR module of the CBS module is shown in Figure 3-2. And apply this module to the ELAN module and EP module in the backbone part of YOLOv7 as shown in Figures 3-3 and 3-4. SiLU, as a smooth nonlinear activation function, although to a certain extent, it alleviates the vanishing gradient problem, its modeling ability for pixel-level spatial information is relatively weak. In contrast, FReLU, by introducing spatial conditions, extends the activation function from one dimension to two dimensions and is able to better capture pixel-level spatial information. This spatial modeling ability is particularly important for the target detection task because it can enhance the model's perception ability of the target's edges and details, thereby improving the detection accuracy. The form of FReLU is, where T(x) is a simple and efficient spatial context feature extractor. This design not only retains the efficiency of ReLU, but also enhances the flexibility and robustness of feature extraction through spatial conditions. In practical applications, FReLU can significantly enhance the model's modeling ability for complex visual layouts without significantly increasing the computational overhead.

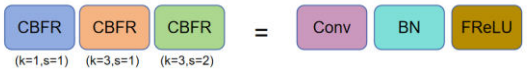

Fig 3-2 CBFR Module structure diagram

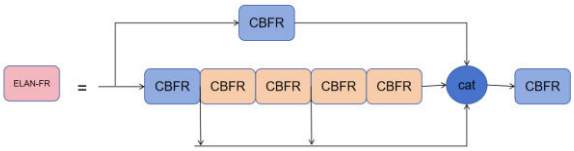

Fig 3-3 ELAN-FR Module structure diagram

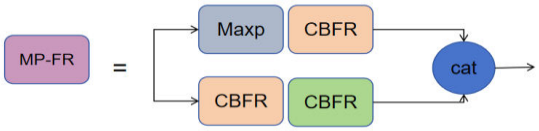

Fig 3-4 MP-FR Module structure diagram

B. Multi-scale target detection head

In the network architecture of the original YOLOv7 algorithm,the input image first undergoes a preprocessing stage to adjust the original image to a uniform pixel size of 640×640 to meet the input requirements of the network. Subsequently, the adjusted image enters the backbone network, and the feature information of the image is gradually extracted through a series of convolutional layers and downsampling operations. The design of the backbone network aims to efficiently capture multi-level features in the image and provide rich semantic information for the subsequent object detection tasks.

In the scenario of small object detection, the 80×80 feature map of the YOLOv7 algorithm corresponds to the receptive field range of 8×8 pixels in the input image. When the size of the target instance is less than 8 pixels, its contour features are prone to fuse with the background texture, resulting in the loss of key information during the feature extraction process, thereby significantly reducing the detection accuracy. In

order to better detect small targets, the 160×160 feature map generated by 4 times downsampling (corresponding to 4×4 pixel receptive fields) can retain more detailed spatial position information and detailed features, thereby performing better in the detection of small targets. To address this issue, the FPN[46]+ PAN[47] feature fusion architecture in the original YOLOv7 network was improved, and the feature pyramid structure as shown in Figure 3-5 was proposed. This structure further enhances the detection ability for small targets by expanding the number of layers of the feature pyramid and fusing the feature map with four times downsampling. The improved feature pyramid structure effectively enhances the feature expression ability of small targets through the fusion of multi-scale features, reduces the interference of background information at the same time, and significantly improves the detection accuracy.

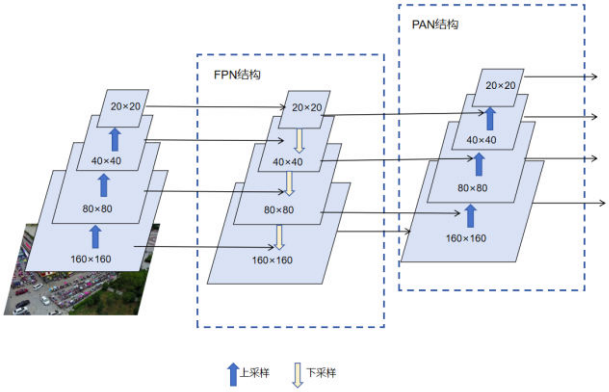

Fig 3-5 The FPN+PAN structure after adding the detection layer

C. *BiFormer:Bidirectional routing attention*

The structure of BiFormer[12] is shown in Figure 3-6 as follows:

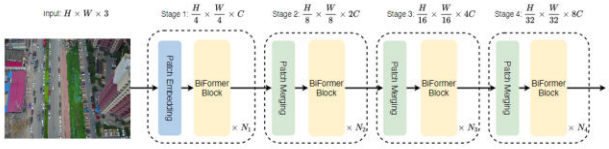

Fig 3-6 Model structure

In the i-th stage, when i=1, the overlapping patch embedding is adopted, while when i=2, 3, 4, the patch merging module is used to reduce the spatial resolution of the input and increase the number of channels at the same time. Subsequently, Transformer operations are performed on the input features using a connected BiFormer block. The insertion position of the BiFormer's attention is shown in Figure 3-7.

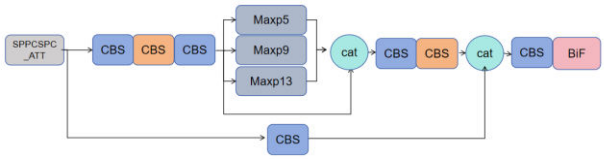

Fig 3-7 SPPCSPC improvement module

D. *Based on the fusion loss function of NWD and IOU*

NWD[13], as a brand-new measurement method, is used to measure the similarity between boxes. It constructs the box as a Gaussian distribution and uses the Wasserstein distance to measure the similarity of the two distributions, thereby replacing the IoU. The advantage of this distance lies in that even if the two boxes do not overlap at all or only overlap very little, the similarity between them can still be accurately measured.

Suppose a horizontal bounding box R=(cx,cy,w,h) is defined, where (cx,cy) represents the central coordinate; w and h are the widths and heights of the bounding box respectively. Thus, the inscribed ellipse of this bounding box can be expressed by formula (3-4) as:

$$\frac{(x-cx)^2}{(\frac{w}{2})^2}+\frac{(y-cy)^2}{(\frac{h}{2})^2}=1 \tag{3-4}$$

The probability density function formula of the two-dimensional Gaussian distribution (3-5) is as follows:

$$f(x|\mu,\Sigma)=\frac{exp(-\frac{1}{2}(x-\mu)^{T}\Sigma^{-1}(x-\mu))}{2\pi|\Sigma|^{\frac{1}{2}}} \tag{3-5}$$

In the formula: x represents the coordinates of the Gaussian distribution (x, y); It is the mean vector of the Gaussian distribution; Σ is the covariance matrix of the Gaussian distribution. When the conditions conform to the following formula (3-6),

$$(X-\mu)^{T}\Sigma^{-1}(X-\mu)=1 \tag{3-6}$$

The inscribed elliptic equation described by the above formula can be regarded as the probability density function of the two-dimensional Gaussian distribution. Therefore, the horizontal bounding box can be modeled as a two-dimensional Gaussian distribution N(). The degree of similarity between two bounding boxes can be quantified by comparing the distances between their corresponding Gaussian distributions. The values of sum Σ are given by the following formulas (3-7) and (3-8) :

$$\mu=\begin{pmatrix}cx\\cy\end{pmatrix}, \tag{3-7}$$

$$\Sigma=\begin{bmatrix}\frac{w^2}{4} & 0\\ 0 & \frac{h^2}{4}\end{bmatrix} \tag{3-8}$$

Suppose there are two Gaussian distributions 1 and 2. The distance between these two distributions can be calculated by means of the second-order Wasserstein distance. The second-order Wasserstein distance is shown by the following formula (3-9) :

$$W_2^2(\mu_1,\mu_2)=\|m_1-m_2\|_2^2+\mathrm{Tr}(\Sigma_1+\Sigma_2-2(\Sigma_2^{\frac{1}{2}}\Sigma_1\Sigma_2^{\frac{1}{2}})^{\frac{1}{2}})=\|m_1-m_2\|_2^2+\left\|\Sigma_1^{\frac{1}{2}}-\Sigma_2^{\frac{1}{2}}\right\|_F^2 \tag{3-9}$$

In the formula: is the Frobenius norm; According to the Gaussian distribution, it is simplified to the following equation (3-10):

$$W_2^2(\mu_1,\mu_2) = \left\| \left(\left[cx_1, cy_1, \frac{w_1}{2}, \frac{h_1}{2} \right]^T, \left[cx_2, cy_2, \frac{w_2}{2}, \frac{h_2}{2} \right]^T \right) \right\|_2^2 \quad (3\text{-}10)$$

The second-order Wasserstein distance provides a method to measure the difference between two distributions. However, it is essentially a distance measure and cannot be directly used as an indicator to measure the similarity between Gaussian distributions. Moreover, when the two borders change from overlapping to separating, the rapid growth of the Wasserstein distance also brings training difficulties. Therefore, The Wasserstein Distance (NWD) is Normalized through the exponential transformation function, and the final Normalized Wasserstein Distance is shown as the following formula (3-11).

$$\mathrm{NWD}(\mu_1,\mu_2) = \exp\left(-\frac{\sqrt{W_2^2(\mu_1,\mu_2)}}{C} \right) \quad (3\text{-}11)$$

In the formula, C is a constant related to the data set (C = 12.8). After fully considering objects of different scales in the aerial images, a mixed border loss based on IoU and Wasserstein distance was constructed, as shown in Equations (3-12) and (3-13) specifically.

$$L = \beta L_{NWD} + (1-\beta) L_{CIoU} \quad (3\text{-}12)$$

$$L_{NWD} = 1 - \mathrm{NWD}(\mu_1,\mu_2) \quad (3\text{-}13)$$

In the above formula, represents the original IOU-based border loss and is the adjustment coefficient.

IV. Experiments

A. Datasets

The VisDrone dataset [14] is an authoritative visual benchmark dataset from the perspective of unmanned aerial vehicles (UAVs) jointly launched by Tianjin University and other research teams in 2018. It consists of 288 video clips consisting of 261,908 frames and 10,209 static images. It contains 6,471 training sets, 548 validation sets and 3,190 test set images (the results need to be submitted to the official evaluation platform), covering complex low-altitude scenarios such as urban roads, transportation hubs and dense crowds. As shown in Figure 3-12, it is the data distribution of the VisDrone dataset. This dataset provides pedestrians, people, bicycles, cars, vans, trucks, tricycles, awning tricycles, buses, and motorcycles There are a total of 11 types of refined annotations for the vehicle (motor) and the invalid area (others). Among them, the others category is usually excluded in algorithm research because it marks non-solid targets such as blurred areas and reflection shadows to enhance the effectiveness of the model. The challenge of the data is reflected in the unique characteristics of unmanned aerial vehicles, such as dense small targets, blurred dynamic motion, drastic changes at multiple scales, and complex background interference. Its annotation also additionally includes attribute labels such as illumination conditions, occlusion rates, and target attitudes, providing the academic community with a core benchmark for measuring the robustness of algorithms. It is widely applied in unmanned aerial vehicle (UAV) visual tasks such as smart city management, traffic flow monitoring, and emergency rescue, and continuously promotes innovative breakthroughs in technologies such as multi-scale feature fusion and occlusion perception networks.

B. Implementation Details

All the experiments in this paper were run on the AutoDL cloud platform server. This platform has outstanding hardware performance. Its multi-core processor is highly efficient in operation, large-capacity memory ensures smooth operation, and high-performance graphics cards accelerate deep learning tasks. In terms of software, the Ubuntu20.04 operating system and the PyTorch framework were used to support the conduct of the experiment. The specific configuration of hardware and software resources is shown in Table 4.1 as follows:

Table 4-1 Experimental environment configuration

Environmental name	Name	Specific configuration
Hardware environment	CPU	Intel(R) Xeon(R) Platinum 8255C
	GPU	RTX 3090(24GB)
	Memory	43GB
Software configuration	Operating system	ubuntu20.04
	Python	Python 3.8
	Pytorch	PyTorch 1.10.0
	CUDA	Cuda 11.3

Set epochs to 300; Set the batch size to 16; The initial learning rate is set to 0.01, and the input image resolution is set to 640×640.

C. Ablation study

In order to verify the effectiveness of several improvement points proposed in this chapter for the YOLOv7 algorithm, through ablation experiments, the contributions made by each improvement point and the combined improvement points to the performance improvement of the YOLOv7 algorithm were analyzed based on the specific mAP, Params and FLOPs evaluation index values in the experimental results. As shown in Table 4-2 below, the specific experimental results of each improved algorithm on the Visdrone dataset are presented. In the table, FR represents the improvement of the CBFR module for the YOLOv7 algorithm, P2 represents adding a small object detection layer to the YOLOv7 algorithm, and B represents the Biformer

attention added at the connection between the backbone network and the head network of the YOLOv7 algorithm. P2B represents adding a small object detection layer to the YOLOv7 algorithm and adding Biformer attention at the connection between the backbone network and the head network, and so on. Yolov7-frpbn represents the final algorithm after all four improvements are added.

Table 4-2 Experimental results of YOLOv7-FRPBN algorithm in Visdrone ablation

算法	Params(M)	FLOPs(G)	mAP (%)
YOLOv7	37.6	106.7	48.5
YOLOv7-FR	37.8	108.4	49.1
YOLOv7-P2	38.2	122.3	49.7
YOLOv7-B	38.7	107.6	48.9
YOLOv7-FRP2	38.3	124.0	50.8
YOLOv7-FRB	38.9	109.4	50.5
YOLOv7-P2B	39.5	123.2	50.3
YOLOv7-FRPB	41.3	124.1	51.0
YOLOv7-FRPBN	41.3	125.0	51.6

D. Performance Comparison

Table 4-3 shows the comparison of the experimental results between YOLOv7-FRPBN and other algorithms. Experimental data show that YOLOv7-FRPBN achieves the optimal balance between model efficiency and detection accuracy. Compared with other models, it achieved the highest 0.516 mAP@0.5 with 125G of computing power (GFlops) ** and 41.3M of parameters, which is 4.2% higher than the suboptimal model ACAM-YOLO (0.495 mAP@0.5) and only 2.6 times the number of parameters. Compared with the models of the same order of magnitude, YOLOv7-FRPBN maintains a relatively low computational complexity (lower than 165.1G of YOLOv8 and 237.8G of YOLOv9-C) while having a significant advantage in accuracy, verifying the effectiveness of its improved strategy in feature extraction and multi-scale fusion.

Table 4-3 Compare with the experimental results of other algorithms

Model	GFlops	Params	mAP@ 0.5
TPH-YOLOv5[15]	129.8G	46.4M	0.429
YOLOXL[16]	155.6	54.1M	0.435
YOLOv7-CPS[17]	122.2G	38.8M	0.482
ACAM-YOLO[18]	130.8G	15.9M	0.495
YOLOv8	165.1G	43.5M	0.479
YOLOv9-C[19]	237.8G	50.9M	0.481
YOLOv7-FRPBN	**125G**	**41.3M**	**0.516**

E. Overall network structure

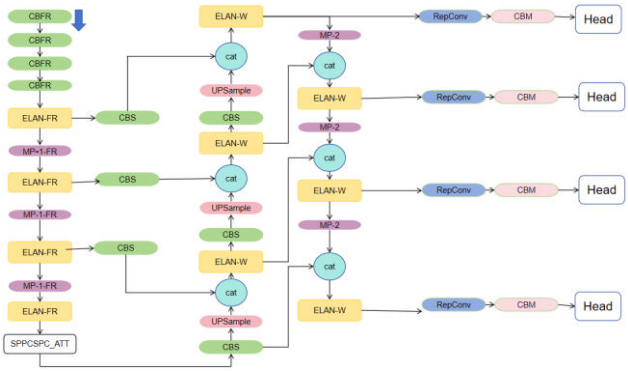

Fig 4-1 The structure diagram of the improved YOLOv7 model

V. CONCLUSION

In this paper, an in-depth study of the YOLOv7 algorithm is conducted, and a series of improvement measures are proposed, aiming to enhance the detection accuracy and robustness of the model for small targets. Aiming at the deficiencies of YOLOv7 in small object detection, this paper proposes a series of improvement measures. The improved CBFR module was introduced. Through the method of cross-block feature fusion, the feature expression ability of small targets was enhanced, effectively solving the problem of feature information loss of small targets. A multi-scale target detection head was designed. By fusing feature maps of different scales, the detection efficiency of the model for multi-scale small targets was further enhanced. The BiFormer bidirectional routing attention mechanism was introduced. This mechanism significantly improves the detection accuracy of the model for small targets by adaptively focusing on the key feature regions of small targets. Meanwhile, a loss function based on the fusion of NWD and IoU was proposed. By combining the advantages of both, the positioning accuracy and classification accuracy of small target detection were better balanced, thereby improving the overall performance of the model. The experimental results show that the improved YOLOv7 small object detection algorithm is superior to the original YOLOv7 algorithm and other similar algorithms in terms of detection accuracy, recall rate and average accuracy. Future research can further explore more efficient small object detection methods to meet the diverse needs in practical application scenarios.

REFERENCES

[1] M. T ,R. P ,Q.V. L .EfficientDet: Scalable and efficient object detection[J].Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2020,10778-10787.

[2] Zhang Haoyang, Wang Ying, Dayoub Feras, et al. VarifocalNet: An IoU-Aware Dense Object Detector[C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8514 - 8523.
[3]]Cai Yuxuan, Wang Yanzhi, Ren Bin, et al. YOLObile: Real-Time Object Detection on Mobile Devices via Compression-Compilation Co-Design[J]. ArXiv abs/2009.05697, 2020.
[4] Cao F ,Xing B ,Luo J , et al.An Efficient Object Detection Algorithm Based on Improved YOLOv5 for High-Spatial-Resolution Remote Sensing Images[J].Remote Sensing,2023,15(15):Zhang, L., Chen, S., Wang, T., & Liu, Z. (2012). Automatic facial expression recognition based on hybrid features. *Energy Procedia*, 17, 1817-1823.
[5] Chen H , Shrivastava A .Group Ensemble: Learning an Ensemble of ConvNets in a single ConvNet[J]. 2020.DOI:10.48550/arXiv.2007.00649.
[6] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. ArXiv abs/1404.5997, 2012.
[7] Han Jiaming, Ding Jian, Li Jie, et al. Align Deep Features for Oriented Object Detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-11.
[8] Sommer L W , Schuchert T , Beyerer J .Fast Deep Vehicle Detection in Aerial Images[C]//2017 IEEE Winter Conference on Applications of Computer Vision (WACV).IEEE, 2017.DOI:10.1109/WACV.2017.41.
[9] Xu S , Wang X , Lv W ,et al.PP-YOLOE: An evolved version of YOLO[J]. 2022.DOI:10.48550/arXiv.2203.16250.
[10] Wang Y, et al. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
[11] Ma N, Ningning et al. Funnel Activation for Visual Recognition[J]. ArXiv abs/2007.11824, 2020.
[12] Zhu Lei, Wang Xinjiang, Ke Zhanghan, et al. BiFormer: Vision Transformer with Bi-Level Routing Attention[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 10323-10333.
[13] Wang Jian, Xu Xingang, Zhang Hao, et al. A Normalized Gaussian Wasserstein Distance for Tiny Object Detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 10323-10333.
[14] Zhu P F, Wen L Y, Du D W, Bian X, Fan H, Hu Q H and Ling H B. 2022. Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11): 7380-7399.
[15] Zhu X, Lyu S, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios[C]. IEEE International Conference on Computer Vision， Montreal, Quebec, Canada, 2021: 2778-2788.
[16] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding Yolo Series in 2021[J]. arXiv Preprint arXiv:2107.08430, 2021.
[17] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information,"2024, arXiv:2402.13616.