Journal of Computers Vol. 36 No. 1, February 2025, pp. 127-142
https://doi.org/10.63367/19911599202502360100 9

A Lightweight Model of Multi-UAVs Object Detection with Dynamic Layer Aggregation

Ruyun Chen, Yong Ding*, and Xuepeng Lu

The College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China
{chenry, dingyong, luxuepeng2000}@nuaa.edu.cn

Received 27 May 2024; Revised 29 August 2024; Accepted 1 October 2024

Abstract. Aiming at the problems of missing detection, false detection and poor real-time performance in the UAV small object detection of aerial images under the complex background of dim, dense trees and buildings, we propose a lightweight YOLO small object detection algorithm with dynamic layer aggregation. Firstly, we propose a dynamic efficient layer aggregation lightweight Backbone, and gradient path planning is adopted to improve the feature extraction ability for small targets. Secondly, we propose a dynamic Neck with Omni-Dimensional dynamic convolution. It helps obtain the variability of each dimension and richer context information, and reduce the missing rate of the model for small targets. Thirdly, SIoU loss function is designed to improve the detection accuracy of small targets and accelerate the convergence of the model. Finally, the channel pruning operation is carried out to compress the size of the model without affecting the performance of the model. Multi-UAVs detection dataset was constructed to evaluate the performance of the proposed model. Compared with the baseline YOLOv5, mAP@0.5 of the proposed model improved by 49.2%, which means its good performance in detecting small targets under complex background environments. And FPS reaches 78.236, meeting the requirements of real-time detection.

Keywords: UAV, multi small objects detection, dynamic ELAN, channel pruning

1 Introduction

Target detection based on UAV (unmanned aerial vehicles) aerial images refers to the autonomous identification of vehicles, drones, obstacles and other targets in the video images captured by a visual camera on the UAV platform, and then subsequent tasks such as flight path planning and obstacle avoidance decisions can be carried out successively. It has been widely used in many fields such as UAV exploration [1], visual obstacle avoidance [2], and traffic supervision [3].

As a basic problem of computer vision, object detection has been extensively studied. Due to the rapid development of deep learning, object detection methods based on deep learning have gradually replaced the traditional object detection methods and become the mainstream methods at present. According to whether there is a target candidate box generation stage, the algorithm is mainly divided into two categories: the two-stage network represented by R-CNN [4] and the single-stage network represented by SSD [5] and YOLO [6]. Compared with the two-stage detection algorithm [7], the single-stage detection algorithm tends to have lower accuracy but faster detection speed [8]. Among the single-stage algorithms, SSD method is a multi-frame prediction method [9], which uses the method of multi-scale feature graph to detect objects and has good detection performance. However, the calculation cost of this method is high. YOLO uses the global information of images to make predictions [10].

Based on this, many scholars have proposed an optimization algorithm for object detection from the UAV perspective. To solve the problems of objects scale change and background interference, Wang et al. [11] extracted scale specific information by using parallel deconvolution spatial pyramid pool, improved attention mechanism by using multi-path residual module and CBAM module. To improve the problem of information loss during down-sampling and effectively improve the average detection accuracy, Zhang et al. [12] introduced spatial

* Corresponding Author

depth convolution blocks and added small object detection heads. To solve the occlusion problem, Liu et al. [13] introduced mosaic data enhancement, cross-small batch normalization and self-adversarial training to YOLO network, improved the performance of object detection and tracking. To solve the problem of mesoscale transformation of marine vessel surveillance of UAVS, Cheng et al. [14] used the dynamic convolution method and introduced ConvNeXt to improve the network, which improved both detection speed and detection accuracy. To solve the occlusion problem in UAV detection and tracking, Yang et al. [15] introduced DiOUS-NMS and added CBAM to increase network feature extraction, reducing the problem of missed detection and false detection caused by occlusion. To solve the problem of low detection accuracy of multi-scale targets, Lu et al. [16] added a small target detection layer and used K-Means++ clustering algorithm and optimized the size of prior frames. To solve occlusion problem, Li et al. [17] build an occlusion guided multi-task network and use an occlusion decoupling head to replace the conventional detection head, thus improved the detection ability of occluded target. To solve the difficulty of detecting small targets of UAVs, Chen et al. [18] introduced an adaptive fusion mechanism to improve the fusion mode of deep and shallow features. To improve the detection accuracy of UAV images, Sahin et al. [19] improved the YOLO network structure by increasing the number of prediction layers of the YOLO model and integrating Transformer method. To solve the problem of dense small and medium targets, Cao et al. [20] integrated lightweight GhostConv method into YOLO, and deleted the large object detection head to obtain a high detection accuracy of small targets. To improve the dense small targets detection performance in complex backgrounds, Jia et al. [21] added CA mechanism to the Backbone YOLO model, extracted important features by embedding location information, enhanced the regression and positioning capabilities of the model, and improved the detection accuracy and robustness of the model.

It can be seen that although there have been some research results on the detection of multi-small targets from the perspective of UAVs, most of the current UAV target detection algorithms can hardly meet both real-time and accuracy due to the scale change, light transformation, small target size, interference from background and other problems in the images captured by UAVs when flying in the air. Moreover, the hardware condition of the drone platform is limited by various factors, thus the computing power that the UAV platform can carry is limited, which makes the object detection under the perspective of UAV more difficult. Therefore, aiming at the problem that the detection effect of multi-small targets under the UAV aerial photography perspective is easily affected by dim and complex interference background, we propose a lightweight multi-UAV small target detection model with dynamic layer aggregation. We focus on the small target feature extraction problem of the lightweight target detection model, and verify the feasibility of the proposed algorithm through experiments. The experimental results demonstrate that the proposed algorithm has a good effect on the real-time detection of small targets under the UAV view.

The technical contributions we proposed are summarized below

(1) We propose a Neck network structure of omni-dimensional dynamic convolution, leveraging the advantages of dynamic convolution in solving the feature loss problem. The adaptive adjustment of the weight of the convolutional kernel is based on input, which enhances the model's ability to extract features for small targets in dim and complex interference environments.

(2) We propose a lightweight Backbone network structure with dynamic efficient layer aggregation, and designed a neural network based on gradient path planning to improve the model's ability to detect multi-small targets, which further solve the problem of multi-UAV small-target detection under complex background interference environment.

(3) The Angle cost is introduced, and the SIoU loss function was designed to further improve the detection performance of the model for small targets of multiple UAVs, which can accelerate the convergence of the model and improve the training speed.

(4) Through channel pruning of the model, the detection performance is less affected, model parameters are reduced, computation is reduced, less computing resources are taken up, and model detection efficiency is improved.

The paper is structured as follows: firstly, we introduce the structure of the YOLOv5s model and analyze its advantages and disadvantages when used in the UAV platform. Secondly, to solve the problem of poor detection effect of multi-UAV small targets in complex interference environments, the backbone network and neck network of the model are improved, and the whole model is pruned. Finally, the simulation results show that the proposed model has been greatly improved in robustness, convergence speed, detection accuracy and detection speed.

2 Related Work

YOLOv5s network is mainly composed of Backbone, Neck and Head. Its structure is shown in Fig. 1. Firstly, the input image is extracted in the backbone network, and the multi-layer feature map of the image is obtained through a series of convolution and pooling operations. Secondly, in the neck network, the features of different levels are fused through up-sampling and down-sampling operations. Finally, these features are passed to the prediction head for regression prediction, prediction frame generation and classification. YOLOv5s has the fastest model detection and performs best on devices with limited computing resources, such as mobile or edge devices.

Although YOLOv5s model is commonly used in target detection tasks performed on UAV platforms, this method has low accuracy for small object detection due to the fixed size of the detection frame. At the same time, the images taken by drones when flying at low altitudes contain dim, complex scenes with dense trees or buildings. Besides, small objects such as lights and birds in the background are easily mistakenly detected as targets, and targets are misclassified as background because of their small size and similar colors to the background. Not only that, but the drone's motion during the shooting also causes the light transformation in the image, and the size and shape of the target can change significantly. All these factors will make it more challenging to extract object features, resulting in worse performance of UAV object detection under these complex interference backgrounds.

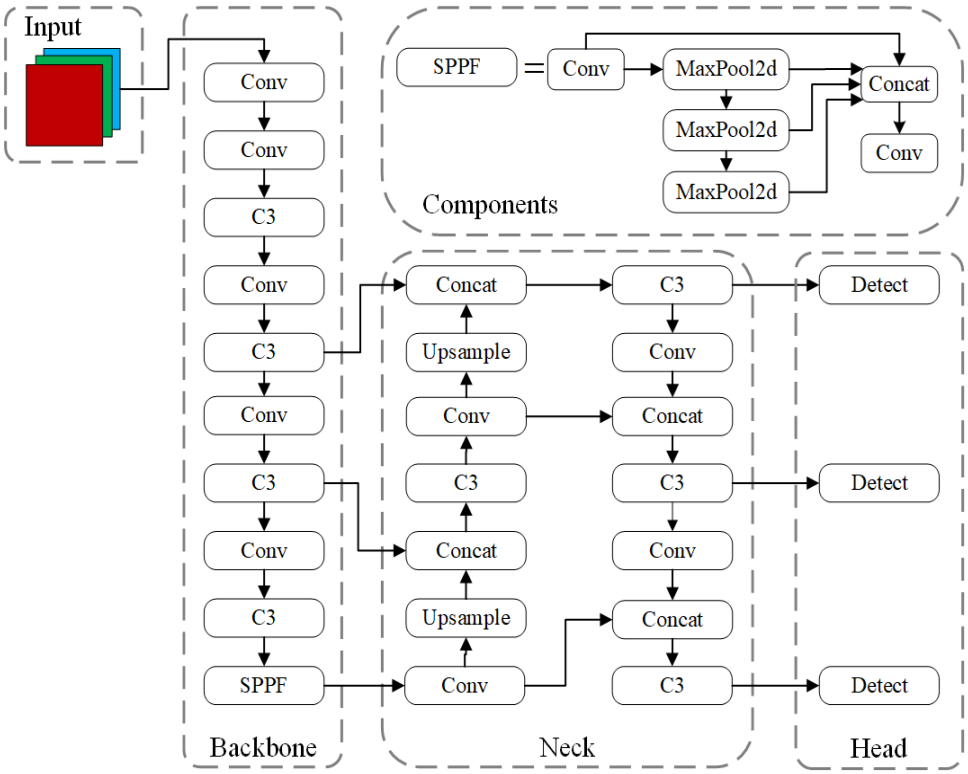


Fig. 1. Overall network structure of YOLOv5s

To solve these problems, we propose a lightweight multi-small target detection model of improved YOLOv5s based on fusion dynamic layer aggregation, focusing on the feature extraction problem of multi-small targets under complex background interference to reduce the error detection and missing detection rate of small targets under complex background, meet the requirements of real-time target detection of UAV aerial images and obtain high detection accuracy.

3 Proposed Methods

In this section, the YOLOv5 structure of the object detection network is improved to meet the requirements of UAV aerial photography object detection tasks under complex backgrounds, and a lightweight YOLOv5 multi-small target detection network with dynamic layer aggregation is proposed. The overall block diagram is shown in Fig. 2. Firstly, the traditional volume layer of the YOLOv5s neck network is replaced by an omni-dimensional dynamic convolution layer. Secondly, the backbone network is improved to a dynamic efficient layer aggregation network structure, and the pool layer module structure is modified to further improve the feature extraction capability of small targets under complex backgrounds. In addition, to solve the problem of the model's difficulty in converging, the loss function is optimized, and the Angle cost is introduced. The Angle relationship between the prediction box and the ground-truth box is considered, and the robustness and training speed of the small target detection model are improved. Finally, channel pruning is carried out on the whole model so that the model can maintain good detection performance on the UAV platform with limited computing power.

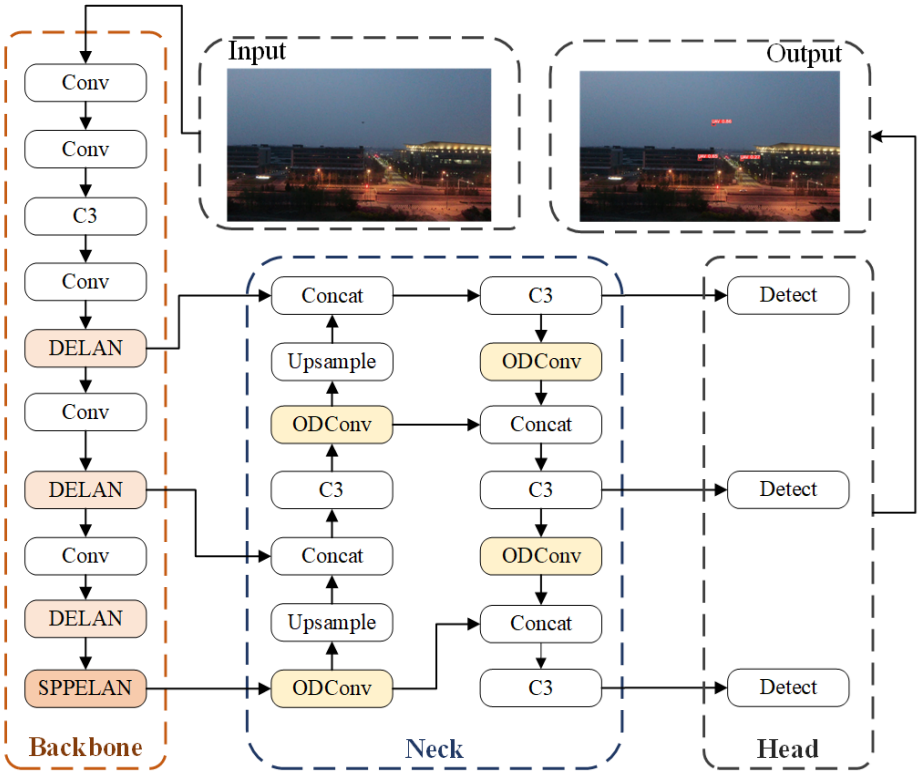


Fig. 2. Structure of lightweight multi-small target detection model we proposed

3.1 The Neck with Omni-Dimensional Dynamic Convolution

Omni-dimensional dynamic Convolution (ODConv) is used to replace the convolution layer in Neck for the multi-small target detection lightweight model from the UAV perspective. By multiplying the dimensions of position, channel, filter, and kernel with different attention, the variability of each dimension can be obtained, which can improve the performance of the model and capture rich context information with better performance. It greatly improves the feature extraction ability of convolutional and enhances the detection ability of multi-small targets, which is helpful in solving the problem of UAV small target detection in complex environments.

The principle of ODConv is to dynamically adjust the shape and size of the convolution kernel according to the characteristics of the input data in the convolution process to adapt to different input data. The ODConv structure is shown in Fig. 3.

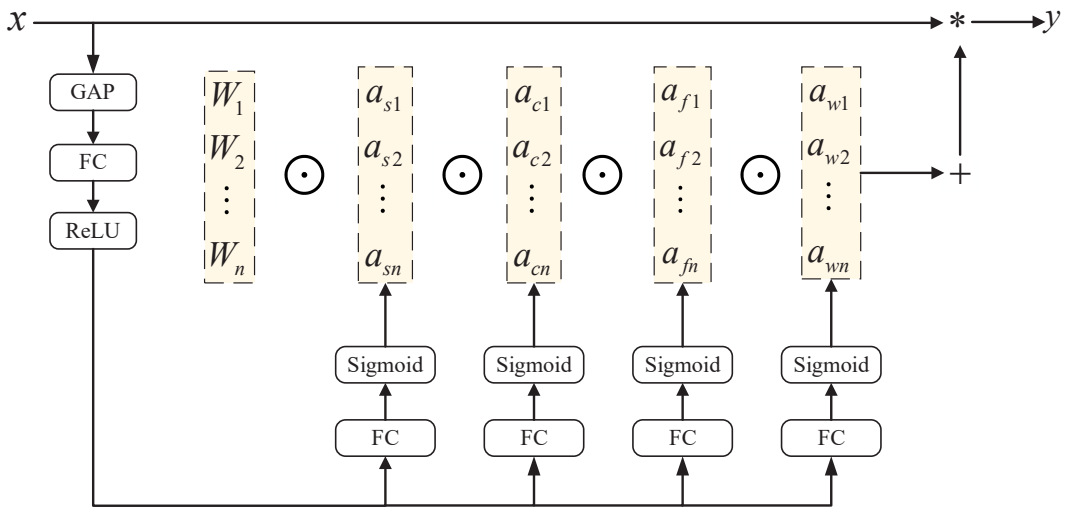


Fig. 3. Structure of ODConv

It is defined as

$$y = \left(\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + \cdots + \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n\right) * x \tag{1}$$

Where $\alpha_{wi} \in \mathbb{R}$ represents the attention scalar of convolution kernel W_i . $\alpha_{si} \in \mathbb{R}^{k \times k}$, $\alpha_{ci} \in \mathbb{R}^{c_{in}}$ and $\alpha_{fi} \in \mathbb{R}^{c_{out}}$ represent the attention introduced by the spatial dimension, the input channel dimension and the output channel dimension of the kernel space of convolution kernel W_i respectively. $\odot$ indicates multiplication is performed on all dimensions of the kernel space.

ODConv considers the dynamic characteristics of spatial space, input channel, output channel, and other dimensions at the same time. By introducing a multi-dimensional attention mechanism that uses parallel strategies, the multi-dimensional attention mechanism learns along the four complementary dimensions of the kernel space. By gradually multiplying the convolution W_i along the position, channel, filter, kernel and other dimensions with different attention, the convolution operation will have different dimensions for the input, providing better performance to obtain rich context information.

3.2 Dynamic Efficient Layer Aggregation Lightweight Backbone Network

Efficient Layer Aggregation Network (ELAN) is a lightweight network architecture that maximizes accuracy within a minimal parameter budget. In terms of target detection, ELAN has shown good performance in terms of accuracy and detection speed at different computing modules and depth Settings. ELAN is a neural network designed with gradient path planning.

In order to further improve the detection performance of multi small objects under complex background interference, a lightweight dynamic efficient layer aggregation Backbone network is proposed. The structure of the main part is shown in Fig. 4. Compared with ELAN method, dynamic ELAN we proposed uses dynamic convolution method instead of traditional convolution method in ELAN structure, which has stronger feature extraction ability. The feature extraction ability of the model is greatly improved in a variety of complex background environments while only a small amount of computation is increased.

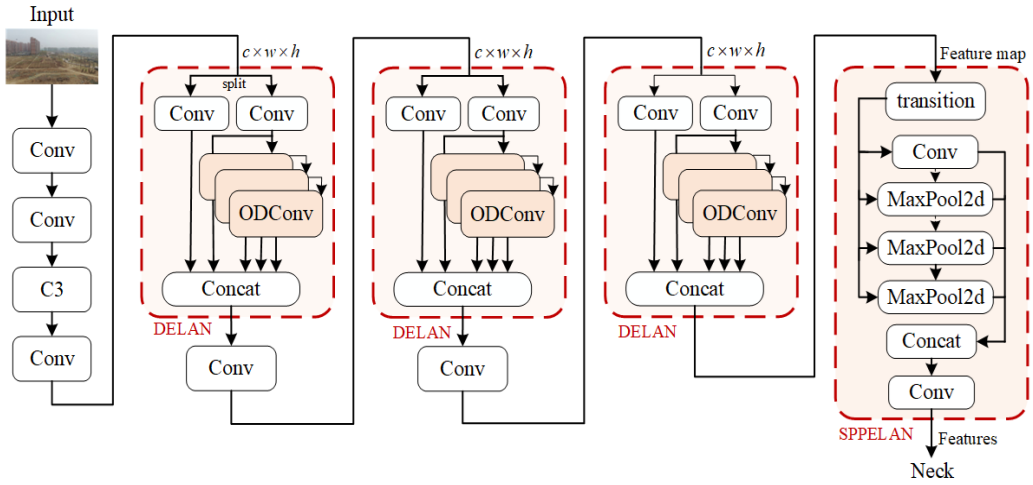


Fig. 4. Structure of Dynamic efficient layer aggregation lightweight Backbone network

The dynamic efficient layer aggregation backbone network we proposed adopts a full-dimensional dynamic convolution method, where multiple parallel convolution cores are dynamically aggregated, and the weight of each convolution kernel is dynamically adjusted according to the input, thus generating an adaptive dynamic convolution. As is shown in Fig. 4, the input size of the Dynamic ELAN module is $c \times w \times h$, c is the number of channels, w is the height of the number of frames, and h is the width of the number of frames. After splitting, the input is divided into two parts, one of which is integrated with the input of the other part after a series of convolution, and the number of output channels is c. After a series of convolution and dynamic ELAN convolution operations, the resulting feature maps are pooled.

In the backbone network, we not only replace some C3 modules with dynamic ELAN modules, but also the original SPPF (Spatial Pyramid Pooling) module is replaced with the SPPELAN (Spatial Pyramid Pooling Enhanced with ELAN) module. As is shown in Fig. 4, the SPPELAN module combines the pooling function of a spatial pyramid with an efficient feature aggregation network structure to capture spatial information at different scales. Compared with the SPPF module, the SPPELAN method makes the model more lightweight while maintaining the robustness of the model, which helps the model reduce the calculation and improve the reasoning speed.

3.3 Design of the Loss Function

In order to solve the problem of slow convergence in the training of the improved model, the loss function of the model is modified. Loss function is needed to measure the degree of prediction error of the model [22]. Compared with the original CIoU (Complete Intersection over Union) loss function used in YOLOv5, which has the problem of slow convergence during training, SIoU (SCYLLA- Intersection over Union) loss function enables the model faster convergence because it adds directionality to the cost function and further considers the angle relationship between prediction box and ground truth box. It can improve the model training efficiency, and it is more suitable for small object detection model. Thus we use SIoU to calculate the loss function.

The SIoU loss function consists of four cost functions: Angle cost, Distance cost, Shape cost and IoU cost. By calculating the width difference, height difference, distance, angle α and angle β of the center point between prediction box and ground truth box, we can get Angle cost L_A , Distance cost L_D , Shape cost L_S and IoU cost $L_{IoU\,Cost}$. The details are as follows.

Angle cost L_A is defined as:

$$L_A = 1 - 2 * \sin^2\left(\arcsin(x) - \frac{\pi}{4}\right) \tag{2}$$

Where $x=\frac{c_h}{\sigma}=\sin(\alpha)$, $\sigma=\sqrt{\left(b_{c_x}^{gt}-b_{c_x}\right)^2+\left(b_{c_y}^{gt}-b_{c_y}\right)^2}$, $c_h=\max\left(b_{c_y}^{gt},b_{c_y}\right)-\min\left(b_{c_y}^{gt},b_{c_y}\right)$. And α is the angle between the center point of the prediction box and the ground truth box. σ is the distance between the prediction box and the ground truth box. c_h is the height difference between the prediction box and the ground truth box. b_{c_x}, b_{c_y} is the center coordinate of the prediction box. $b_{c_x}^{gt}$, $b_{c_y}^{gt}$ is the center coordinate of the ground truth box.

Distance cost L_D is defined as:

$$L_D=\sum_{t=x,y}\left(1-e^{-\gamma\rho t}\right) \tag{3}$$

Where $\rho_x=\left(\frac{b_{c_x}^{gt}-b_{c_x}}{c_w}\right)^2$, $\rho_y=\left(\frac{b_{c_y}^{gt}-b_{c_y}}{c_h}\right)^2$, $\gamma=2-L_A$ and c_h, c_w is the height and width of the minimum external rectangle for the prediction box and the ground truth box.

Shape cost L_S is defined as:

$$L_S=\sum_{t=w,h}\left(1-e^{-\omega_t}\right)^{\theta} \tag{4}$$

Where $\omega_w=\frac{\left|w-w^{gt}\right|}{\max(w,w^{gt})}$, $\omega_h=\frac{\left|h-h^{gt}\right|}{\max(h,h^{gt})}$. θ is the degree of concern for shape loss. h, w is the height and width of the predicted box. h^{gt}, w^{gt} is the height and width of the ground truth box.

IoU cost $L_{IoU\,Cost}$ is defined as:

$$L_{IoU\,Cost}=1-IoU \tag{5}$$

Where $IoU=\frac{\left|B\cap B^{GT}\right|}{\left|B\cup B^{GT}\right|}$, B is a prediction box, B^{GT} is the ground truth box.

Regression loss function L_{box} is defined as:

$$L_{box}=1-IoU+\frac{\Delta+\Omega}{2} \tag{6}$$

The loss function L consists of two parts: classification loss and box loss:

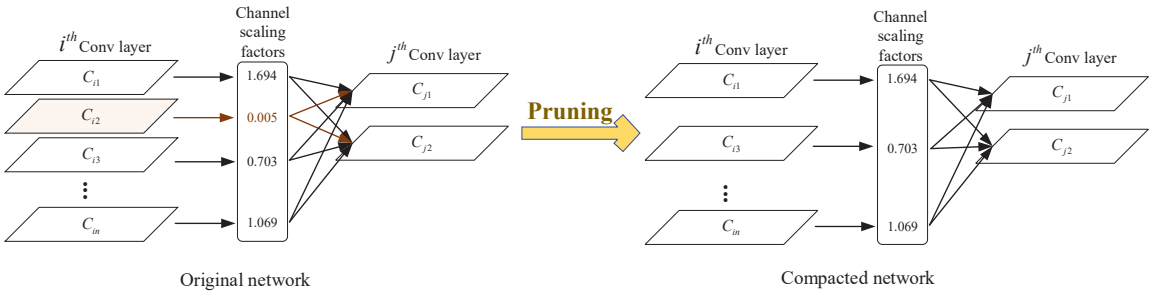
$$L=W_{box}L_{box}+W_{cls}L_{cls} \tag{7}$$

Where W_{box} is box loss weight, W_{cls} is classification loss weight, and L_{cls} is focal loss.

3.4 Model Compression

YOLO consists of tens of millions of parameters and requires nearly 100 billion floating-point operations, which makes it difficult to detect the target in real-time on the UAV platform, so we introduce model compression technology. Commonly used model compression methods include pruning [23], knowledge distillation [24] and quantization [25]. Although the quantization method can significantly reduce the model size, the performance of the model will be greatly reduced due to the uneven data distribution. Knowledge distillation can accelerate network convergence, but the model trial and error cost is high, and the interpretability is poor. In contrast, the pruning method has less impact on the model performance while reducing the amount of computation, so it is more used in the model compression of deep neural networks.

Aiming at the lightweight requirement of the YOLOv5s high-performance model proposed in this paper, Channel Pruning [26] is adopted to compress the improved YOLOv5s model after training. Channel pruning is a method to reduce the number of parameters and computation of deep neural networks, which can effectively reduce the model size and improve the efficiency of target detection and is more suitable for optimizing the network structure of YOLO. The channel pruning method selects a certain number of channels to delete according to the characteristics of the model after the basic model training and further fine-tune. The pruning diagram is shown in Fig. 5. In the figure, (a) is the initial network before pruning, and (b) is the network after pruning. The channel removed by pruning operation is the part marked orange in the figure (a), which has a small weight and negligible effect on the overall model.



(a) Before pruning (b) After pruning

Fig. 5. Pruning process

For the improved YOLOv5 model we proposed, the pruning rate was set to 80% according to the weight of each layer, and the convolution layers Conv, C3 and SPPELAN modules in Backbone with low weight were pruned. After pruning is completed, the pruning model is trained again for fine-tuning to further adjust it to reduce the impact of compression operation on model performance.

4 Experiment

The experimental running environment we used is set up in the deep learning framework PyTorch, and the GPU parallel accelerated computation is carried out in the CUDA environment of the GPU server NVIDIA GeForce RTX3090. The experimental environment is shown in Table 1.

Table. 1. Experimental environment

Device	Configuration
Operating System	Windows 10.0
CPU	Intel Xeon Gold 6133
GPU	NVIDIA GeForce RTX3090
Training Environment	CUDA 11.1 cuDNN 8.2.1
Developing Environment	Python 3.8.10 Pytorch 1.10.2

Since there are few multi-UAV detection datasets from the UAV perspective, we use images collected from the MOT-FLY dataset and modify them into object detection datasets. The multi-UAV image data used in this paper includes different background scenes, viewing angles, UAV sizes, flight heights, and lighting conditions. After selected and annotated images by LabelImg software, 1290 images were used as the training set and 270 images were used as the validation set. In order to verify the performance of the model under dim and complex interference backgrounds, the validation set mainly consists of pictures of night and complex backgrounds. Some of the annotated images are shown in Fig. 6.

(a) Part of the annotated training set

(b) Part of the annotated validation set

Fig. 6. Part of the annotated dataset

4.1 Model Performance Evaluation Metric

This experiment uses the commonly used indicators of target detection performance evaluation. FPS (Frames Per Second) represents the number of images that can be processed per second and is used to evaluate the model's processing speed on the hardware. *R* (Recall) represents the proportion of correctly identified positive samples to all predicted samples, which is defined as

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

Where *TP* (True Positives) represents instances where the model correctly predicted the positive sample, and *FN* (False Negatives) represents instances where the positive sample was incorrectly predicted to be negative.

P (Precision) represents the proportion of correctly predicted positive samples to all positive samples, which is defined as

$$P = \frac{TP}{TP + FP} \tag{9}$$

Where *FP* (False Positives) represents instances where the negative sample was incorrectly predicted to be positive.

mAP (mean Average Precision) represents the mean of the average precision of all individual classes, which is defined as

$$mAP = \frac{1}{n}\sum_{i=1}^{n}\left(\int_0^1 PdR\right)_i \tag{10}$$

Where *n* represents the number of all individual classes.

4.2 Ablation Experiments

To verify the effectiveness of the lightweight YOLOv5 multi-small object detection method we proposed, the ablation experiment was designed. That is, dynamic ELAN backbone network structure, omni-dimensional dynamic convolution ODConv neck network and SIoU loss function were added to the baseline YOLOv5s model. At the same time, the various kinds of combinations of above improvement methods are added, and the comparison before and after pruning was made. The results are shown in Table 2.

The results show that after replacing the backbone network with a dynamic ELAN structure, mAP@0.5 increases by 33.8%. After introducing ODConv into the neck network, mAP@0.5 increased by 18.0%. Therefore, the detection performance of the model can be greatly improved by using the Backbone with dynamic efficient layer aggregation we proposed, and the detection accuracy and recall rate of the model can be further enhanced by introducing ODConv convolutional module. Besides, after modifying the loss function to the SIoU method, the detection performance does not get much improvement. However, the training time of the model can be shortened, which also proves that the SIoU loss function is more conducive to model convergence and can improve the robustness of the model. Moreover, it has also been proved that the model we proposed has good detection performance before and after pruning, and mAP@0.5 can reach more than 95%.

Table. 2. Ablation experiment results

YOLOv5	GELAN	ODConv	SIoU	Pruning	mAP@0.5	Precision	Recall
√	-	-	-	-	0.458	0.810	0.411
√	√	-	-	-	0.796	0.919	0.738
√	-	√	-	-	0.638	0.859	0.635
√	-	-	√	-	0.551	0.866	0517
√	√	√	-	-	0.963	0.983	0.934
√	√	√	√	-	0.966	0.980	0.938
√	√	√	√	√	0.950	0.967	0.899

4.3 Comparative Experiments

Comparison with other detection methods in the YOLOv5 series. To verify that the method used in this paper can more effectively improve the detection performance of small targets under complex background, a comparative experiment is conducted. The experiment compares the method we used with other detection methods based on the YOLOv5s model. The comparison includes the baseline YOLOv5s model, the model with introduction of Self Attention mechanism [27], the introduction of SPATIAL-SHIFT(S2) attention mechanism [28], the convolutional layer modified as dynamic convolution CondConv [29], the convolutional layer modified as full-dimensional dynamic convolution ODConv, and the loss function modified as CIoU [30], EIoU [31], SIoU. The comparison experiment designed in this section is also based on the multi-UAV small target detection dataset with complex interference background, mainly comparing the model size, mAP@0.5, mAP@.5:.95, Accuracy, Recall and other performance indicators, as shown in Table 3.

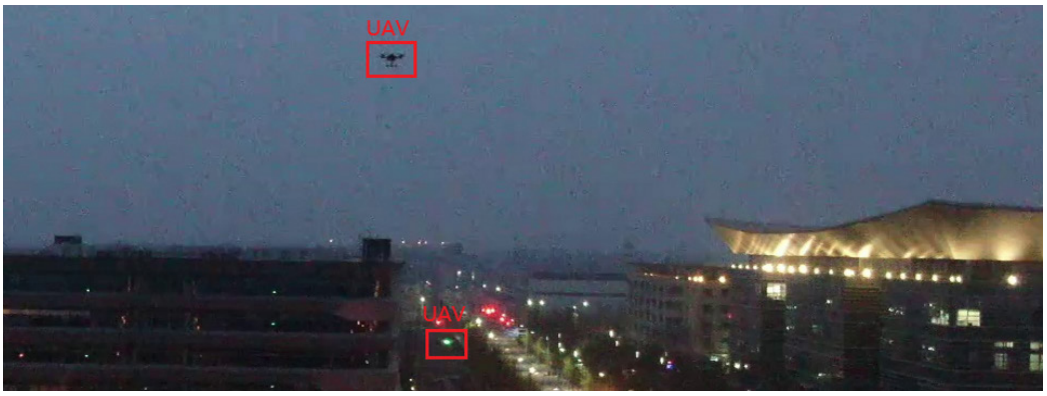
As is shown in Table 3, firstly the Self Attention mechanism is integrated into the baseline YOLOv5 model. Although the model size can be reduced, the detection performance is worse. Subsequently, the self-attention mechanism is substituted with the S2 attention mechanism, resulting in an increase in the size of the trained model without any improvement in detection performance. Furthermore, the convolution layers within the baseline model are replaced with dynamic convolution. In comparison to the CondConv dynamic convolution method, it

is found that utilizing the full-dimensional dynamic convolution ODConv we used has greater improvements in detection accuracy, mAP@0.5 incresed by 25.6%, and Recall increased by 46.0%. Additionally, modifications are made to the loss function of the bsaeline model. It is observed that using SIoU loss function outperforms CIoU and EIoU in terms of Accuracy and Recall. Compared to IoU used in the baseline model, employing SIoU results in an increase of mAP@0.5 by 20.3% and Recall by 25.8%. Moreover, due to implementing SIoU loss function, there is also a reduction in training time for model convergence which demonstrates its ability to accelerate convergence process.

Table. 3. Comparison of small target detection based on different YOLOv5s models under complex background

Models	Size	mAP@0.5	mAP@.5:.95	Precision	Recall
Baseline YOLOv5s	13.7M	0.458	0.299	0.810	0.411
+Self Attention [27]	11.6M	0.409	0.285	0.737	0.425
+S2-MLPv2 [28]	17.6M	0.449	0.274	0.751	0.415
+CondConv [29]	14.6M	0.508	0.316	0.819	0.435
+ODConv (We used)	14.7M	*0.638*	*0.425*	*0.859*	*0.635*
+CIoU [30]	13.7M	0.510	0.379	0.829	0.423
+EIoU [31]	13.7M	0.524	0.473	0.860	0.502
+SIoU (We used)	13.7M	0.551	0.479	0.866	0.517

In order to further analyze and explain the superiority of the proposed method in complex interference scenes, we have selected the detection effects of UAVs under two complex interference backgrounds for demonstration, as marked in Fig. 7. The red box indicates the location of UAVs. In Fig. 7(a), Scene 1 is a dim environment with various light disturbances, including two drones. In Fig. 7(b), Scene 2 depicts a scene with small objects such as cars and trees interfering in the background, including three drones.



Scene 1

Scene 2

Fig. 7. Two annotated scenarios with complex interference

The test results are presented in Fig. 8. In Fig. 8(a), the detection results of three models in Scene 1 are shown, including the addition of a self-attention mechanism, the addition of an S2 attention mechanism, and the im-

provement of the convolutional layer to ODConv layer from top to bottom. Fig. 8(b) displays the detection results of these three models in Scene 2.

In Scene1, all three methods miss detectig a drone in the background with light interference. The model incorporating the Self Attention method and the model using full-dimensional dynamic convolution did not exhibit false detections. However, the model introducing S2 attention mechanism mistakenly detected building and background light as a drone. In Scene 2, both methods introducing attention mechanisms resulted in some false detections by mistaking trees and vehicles in the background as drones. Although our proposed full-dimensional dynamic convolutional layer model also failed to detect all UAVs, it significantly reduced false detection rates.

From both training data and model detection images obtained from experiments, it is evident that our proposed method can more effectively address multi-UAV small target detection issues, especially those related to false detections under complex interference backgrounds by using full-dimensional dynamic convolution methods.

Fig. 8. The detection results of different improved models under complex background interference

Comparison of the proposed model before and after pruning with the baseline model. To verify the performance superiority of the proposed model, we design a comparison experiment, which includes the comparison of the performance of the baseline YOLOv5s with the model before and after pruning of the model we proposed. The multi-UAV small target detection dataset for verification contains the following five complex backgrounds: Scene 1 is the overall dim light with interference background such as lights and buildings, Scene 2 is the dense forest background interference background, Scene 3 is the complex background including woods, buildings, land and other disturbances, Scene 4 is the dim light with dense building interference background, Scene 5 is the background with dense building interference. The detection results of the three models under five different complex backgrounds are shown in Fig. 9, where column (a) is the results of the baseline YOLOv5s model, column (b) is the results of the improved YOLOv5s model of fusion dynamic layer aggregation we proposed, and column (c) is the test results of the model used in column (b) after pruning.

According to the results displayed in Fig. 9, the detection performance of the baseline YOLOv5s model is poor. There are missing detection problems in the first, second, fourth and fifth scenes, and the problem of false detection is prone to occur in the fourth scene with dim light and dense building interference. Then, we use the model we proposed, which contains the dynamic layer aggregation Backbone, Neck with ODConv and SIoU loss function. After the proposed improvement, the missed detection and false detection are effectively solved, and it had a good small-target detection performance. After the improved model is pruned, the model is trained again. According to the detection results, most of the UAVs can still detect, and only in the second scene with dense

forest interference, there is a case of missing detection, which has little impact on the performance of the pruned model.

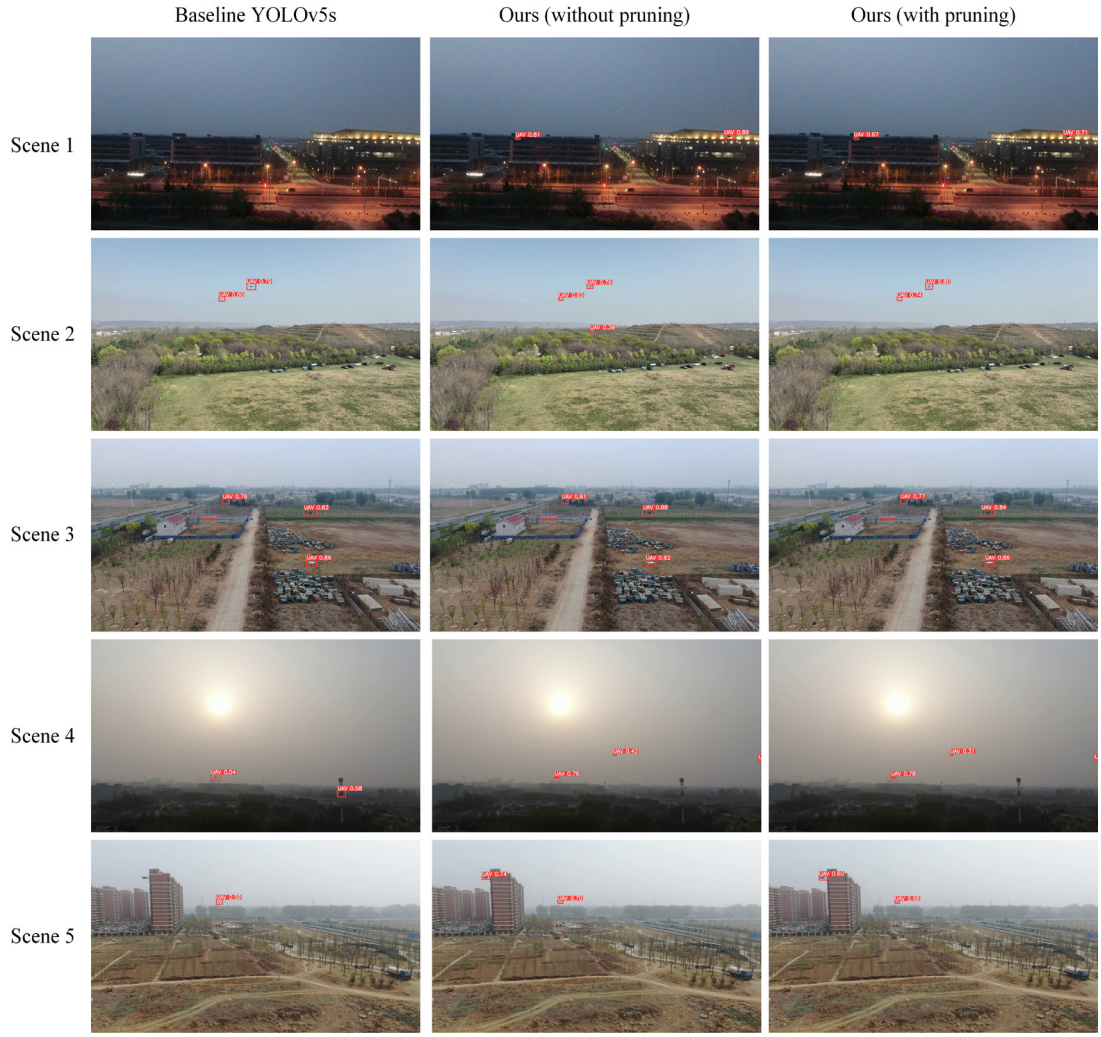


Fig. 9. Comparison of detection results of different models on multiple small targets in complex scenes

The data of the performance indicators of the baseline model and our proposed model are shown in Table 4.

Table. 4. Comparison of multi small object detection performance of different models in complex scenarios

Model	Size	mAP@0.5	mAP@.5:.95	Precision	Recall	FPS
Baseline YOLOv5s	13.7M	0.458	0.299	0.810	0.411	75.996
Ours (without pruning)	23.3M	*0.966*	*0.655*	*0.980*	*0.938*	77.353
Ours (with pruning)	16.4M	0.950	0.593	0.967	0.899	*78.236*

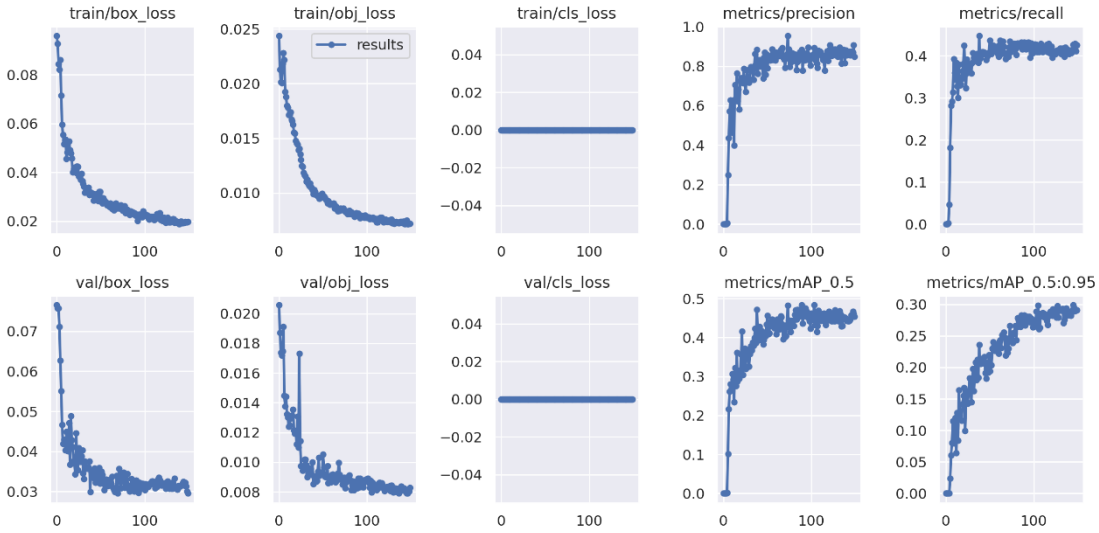
According to the comparative experimental results in Table 4, it is evident that the introduction of dynamic layer aggregation Backbone, Neck with ODConv, and SIoU loss function has significantly improved the performance of the model compared with the baseline YOLOv5s model. Specifically, there is a 50.8% increase in mAP@0.5, a 35.6% increase in mAP@.5:.95, a 17.0% improvement in accuracy, and 52.7% enhancement in Recall. The Accuracy under dim background was notably enhanced while reducing false detection rate and missing detection rate under complex background interference. After conducting channel pruning operation on the proposed model, there is a reduction of 29.6% in Size. However, this reduction only resulted in little decreases

in mAP@0.5 (1.6%), mAP@.5:.95 (6.2%), Precision (1.3%), and Recall (3.9%) compared to the model before pruning.

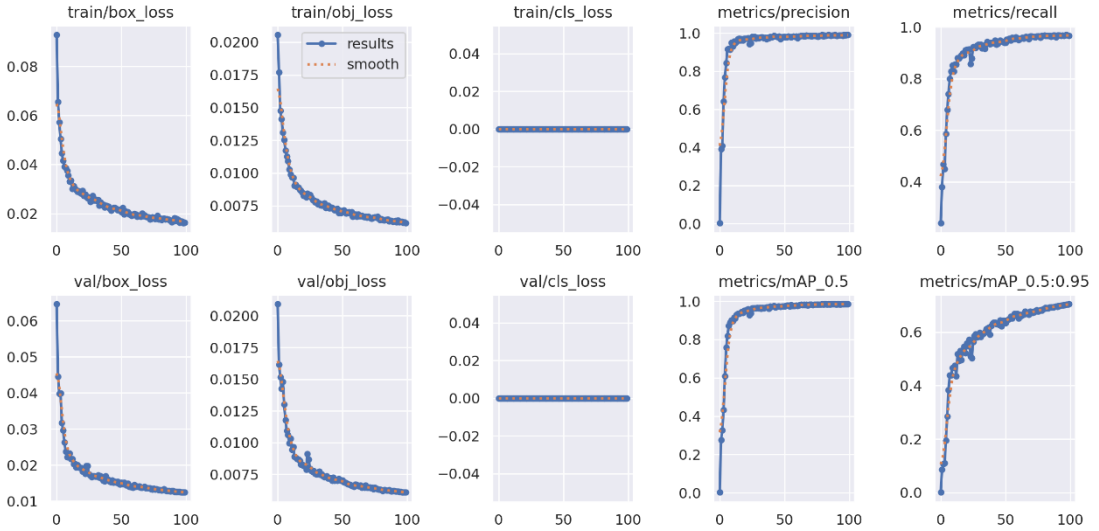
Compared with the baseline model, mAP@0.5 of the proposed model with pruning improves by 49.2%, mAP@.5:.95 by 29.4%, Precision by 15.7%, Recall by 48.8%, and FPS by 2.9%. The comparison experiment shows that the proposed model has a great improvement in Precision, Recall and Accuracy compared with the baseline model. And after pruning operation, the performance of the model does not drop much, whose FPS reaches 78.236, meeting the real-time requirements.

In order to further demonstrate the superiority of ours proposed model, the training effect is compared with the baseline model. The training curve of the baseline YOLOv5 and the dynamic layered lightweight model we proposed in this paper are shown in Fig. 10(a) and Fig. 10(b) respectively.

As shown in Fig. 10, the loss box_loss and obj_loss curves of the model we proposed smoothly reach a small value, reaching the performance level after only 50 rounds of training compared to 150 rounds required by the baseline model. Compared from the aspects of robustness, convergence speed and detection effect, The performance of the model proposed in this paper is far superior to the baseline model in all aspects.



(a) Training curves of the baseline YOLOv5s model



(b) Training curves of lightweight model with dynamic layer aggregation

Fig. 10. Comparison of training curves of the baseline model and ours

5 Conclusion

Aiming at the problem that multi-UAV objects are small and difficult to detect in aerial images when UAV faces complex interference background at low altitudes, we focus on the feature extraction problem of multi-small targets under complex interference backgrounds and propose a lightweight YOLO of multi-UAV small target detection algorithm with dynamic layer aggregation is proposed. First, the original convolutional layer of the neck network is improved to an omni-dimensional dynamic convolutional layer, which improves the feature extraction ability of the model for small targets. Secondly, a lightweight backbone network with dynamic efficient layer aggregation is proposed, and the convolutional layer and pooling layer are modified to further improve the feature extraction capability of small targets under complex backgrounds. Thirdly, the loss function of the model is modified, the Angle cost function is introduced, and the SIoU loss function is used to accelerate the convergence of the model and improve the robustness of the model. Finally, the channel pruning of the model is carried out to reduce the number of model parameters and the amount of calculation while reducing the model detection performance. Through experiments, the lightweight YOLO multi-UAV small target detection algorithm proposed in this paper is compared with the benchmark YOLOv5 model that mAP@0.5 improved by 49.2%, mAP@.5:.95 improved by 29.4%, accuracy improved by 15.7%, recall rate improved by 48.8%, FPS improved by 2.9%. The proposed model effectively improves the detection ability of multi-UAV small objects in complex backgrounds, and verifies that the model has good real-time performance.

Although the proposed algorithm performs well in the task of detecting small targets of UAVs, its performance in detecting multi-scale targets in this environment needs to be improved. The multi-scale target detection task will be further studied in the future.

6 Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 62173180).

References

[1] M. Aibin, Y.-X. Li, R. Sharma, J.-Y. Ling, J.-N. Ye, J.-S. Zhang, L. Coria, Advancing Forest Fire Risk Evaluation: An Integrated Framework for Visualizing Area-Specific Forest Fire Risks Using UAV Imagery, Object Detection and Color Mapping Techniques, Drones 8(2)(2024) 39-57.

[2] N. Zhang, F. Nex, G. Vosselman, N. Kerle, End-to-End Nano-Drone Obstacle Avoidance for Indoor Exploration, Drones 8(2)(2024) 33.

[3] S.-C. Li, X.-D. Yang, X. Lin, Y.-Z Zhang, J.-H Wu, Real-Time Vehicle Detection from UAV Aerial Images Based on Improved YOLOv5, Sensors 23(12)(2023) 5634.

[4] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.-C. Berg, SSD: Single Shot MultiBox Detector, in: Proc. 14th European Conference on Computer Vision, 2016.

[6] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[7] Z. Jian, K. Feng, W.-X. Li, J. Han, F. Pan, TS4Net: Two-stage sample selective strategy for rotating object detection, Neurocomputing 501(2022) 753-764.

[8] H.-Y. Zhang, W.-R. Li, Y.-Y. Qi, H.-N Liu, Z.-B. Li, Dynamic fry counting based on multi-object tracking and one-stage detection, Computers and Electronics in Agriculture 209(2023) 107871.

[9] S.-C. Liu, H. Shi, Z. Guo, Remote sensing image object detection based on improved SSD, in: Proc. 3rd CVIDL & ICCEA, 2022.

[10] M. Hussain, YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection, Machines 11(2023) 677.

[11] J. Wang, F. Zhang, Y. Zhang, Y.-H. Liu, T. Cheng, Lightweight Object Detection Algorithm for UAV Aerial Imagery, Sensors 23(13)(2023) 5786.

[12] J. Zhang, G.-Y. Wan, M. Jiang, G.-F. Lu, X.-W. Tao, Z.-Y. Huang, Small object detection in UAV image based on improved YOLOv5, Systems Science and Control Engineering 11(1)(2023) 1-12.

[13] X. Liu, Z. Zhang, A Vision-Based Target Detection, Tracking, and Positioning Algorithm for Unmanned Aerial Vehicle, Wireless Communications and Mobile Computing 2021(2021) 5565589.

[14] S.-X. Cheng, Y.-S. Zhu, S.-H. Wu, Deep learning based efficient ship detection from drone-captured images for maritime surveillance, Ocean Engineering 285(2)(2023) 115440.
[15] H. Yang, Y. Ge, Research on Detecting and Tracking Algorithm of UAV Intrusion Based on YOLOv5+DeepSort, in: Proc. 3rdCVIDL & ICCEA, 2022.
[16] Q. Lu, Y.-Q. Yu, D.-M. Xu, Q. Zhang, Improved YOLOv5 Small Drones Target Detection Algorithm, Computer Science 50(S2)(2023) 212-219.
[17] X.-X. Li, W.-H. Diao, Y.-Q. Mao, P. Gao, X.-H. Mao, X.-M. Li, X. Sun, OGMN: Occlusion-guided multi-task network for object detection in UAV images, ISPRS Journal of Photogrammetry and Remote Sensing 19(2023) 242-257.
[18] R. Chen, H.-F. Zhen, H.-Y Jiang, Y.-W. Guo, Combination of simulation-based transfer learning and adaptive fusion for UAV small object detection, Journal of Chinese Computer Systems 44(8)(2023) 1743-1749.
[19] O. Sahin, S. Ozer, YOLODrone+: Improved YOLO Architecture for Object Detection in UAV Images, in: Proc. 30th Signal Processing and Communications Applications Conference, 2022.
[20] J. Cao, W.-S. Bao, H.-X. Shang, M. Yuan, Q. Cheng, GCL-YOLO: A GhostConv-Based Lightweight YOLO Network for UAV Small Object Detection, Remote Sensing 15(20)(2023) 4932.
[21] Y.-J. Jia, K. Fu, H. Lan, X. Wang, Z.-B. Su, Maize tassel detection with CA-YOLO for UAV images in complex field environments, Computers and Electronics in Agriculture 217(2024) 108562.
[22] P.-K. Sekharamantry, F. Melgani, J. Malacarne, Deep Learning-Based Apple Detection with Attention Module and Improved Loss Function in YOLO, Remote Sensing 15(2023) 1516.
[23] G.-Y. Liu, Y.-Z. Li, Y.-C. Song, Y.-M. Liu, X.-F. Xu, Z. Zhao, R.-H. Zhang, A lightweight convolutional network based on pruning algorithm for YOLO, in: Proc. International Conference on Graphic and Image Processing, 2023.
[24] X.-Y. Liu, T. Wang, J.-M. Yang, C.-W. Tang, J.-C. Lv, MPQ-YOLO: Ultra low mixed-precision quantization of YOLO for edge devices deployment, Neurocomputing 574(2024) 127210.
[25] L.-P Gan, R.-Z Cao, N. Li, M. Yang, X.-C. Li, Focal Channel Knowledge Distillation for Multi-Modality Action Recognition, IEEE Access 11(2023) 78285-78298.
[26] Y. He, X. Zhang, J. Sun, Channel Pruning for Accelerating Very Deep Neural Networks, in: Proc. 2017 IEEE International Conference on Computer Vision, 2017.
[27] Y.-M. Xie, L.-W. Zhang, X.-Y. Xu, W. Xie, YOLO-MS: Multispectral Object Detection via Feature Interaction and Self-Attention Guided Fusion, IEEE Transactions on Cognitive and Developmental Systems 15(4)(2023) 2132-2143.
[28] T. Yu, X. Li, Y.-F. Cai, M.-M. Sun, P. Li, S2-MLP: Spatial-Shift MLP Architecture for Vision, in: Proc. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, 2022.
[29] G.-K. Ji, R. Wang, S.-F. Peng, Person re-identification method based on attention mechanism and CondConv, Journal of Beijing University of Aeronautics and Astronautics 50(2)(2024) 655-662.
[30] H. Zhong, S.-L. Hu, Target Detection Method of Apple Harvesting Robot Based on Improved YOLO v5, in: Proc. 2023 35th Chinese Control and Decision Conference, 2023.
[31] W.-M. Qi, H.-G. Chen, Y.-T. Ye, G.-S. Yu, Indoor object recognition based on YOLOv5 with EIOU loss function, in: Proc. 3rd International Conference on AASIP, 2023.