UAV Target Detection Algorithm based on Improved YOLOv5

Zijun Guan
School of Electric
Information and
Electrical Engineering
Yangtze University
Jingzhou, China

Abstract: Aiming at the problems of complex background, dense small targets, and large scale changes in UAV aerial images, this paper proposes a target detection algorithm based on improved YOLOv5s. First, ShuffleNetV2 is used to reconfigure the backbone network, which improves the detection speed while reducing the computational complexity and realizes the model lightweight. Second, the ASPP module is introduced to enhance the multi-scale feature extraction capability, reduce the small target feature loss, and improve the detection accuracy. Further, the CA attention mechanism is introduced into ShuffleNetV2 and YOLOv5s feature fusion network to strengthen the extraction and expression of key features. Finally, the SIoU loss function is substituted for the original CIoU to better adapt to multi-scale targets and improve the bounding box regression accuracy.Experiments conducted on the VisDrone2019 dataset demonstrate that the mAP0.5 and mAP0.5:0.95 of the improved model reach 39.5% and 21.5%, respectively, which are 5.0% and 3.1% higher than the original model, and the amount of parameters is reduced to 54%. The results show that the algorithm effectively improves the performance of target detection in complex scenes under the premise of ensuring the lightweight of the model, and has high application value.

Keywords: unmanned aircraft; target detection; small target detection; attention mechanism; lightweighting

1. INTRODUCTION

In recent years, unmanned aerial vehicle (UAV) technology has been widely applied in fields such as agricultural monitoring, urban security, intelligent logistics, surveying and mapping, and military reconnaissance [1][2]. With the development of computer vision and deep learning, unmanned aerial vehicles (UAVs) have gradually acquired autonomous perception and intelligent decision-making capabilities, especially demonstrating great potential in target detection and recognition tasks [3]. However, due to factors such as complex image backgrounds and large variations in target scales, how to achieve efficient and accurate small target detection remains an important research topic.

Among the current object detection algorithms, single-stage detection is more suitable for unmanned aerial vehicle (UAV) platforms with high real-time requirements due to its simple structure and high speed[4]. The YOLO series, as a representative among them, is widely used in embedded vision tasks due to its efficient and accurate detection performance[5]. The lightweight YOLOv5s has advantages in model size and inference speed, but it still has the problem of insufficient detection accuracy in complex scenarios such as dense small targets, large scale variations, and severe occlusion[6]. To this end, this paper proposes an improvement method based on YOLOv5s: using ShuffleNetV2 to reconstruct the backbone network to enhance the inference speed; Introduce the ASPP module to enhance multi-scale feature extraction; Incorporate the CA attention mechanism into the backbone and fusion structures to enhance key features; And SIoU is used to replace the CIoU loss function to improve the regression accuracy. The improved model has effectively enhanced the performance of small target detection for unmanned aerial vehicles (UAVs) in complex scenarios while maintaining lightweight.

2. IMPROVED YOLOv5 ALGORITHM

YOLOv5s is a typical lightweight object detection model in the YOLO series. It adopts a single-stage structure, converting the detection task into a single regression. It has excellent real-time performance and is suitable for deployment on embedded devices. Its overall network structure includes an input module, backbone network, feature fusion structure and prediction head. The network structure is shown in Figure 1.

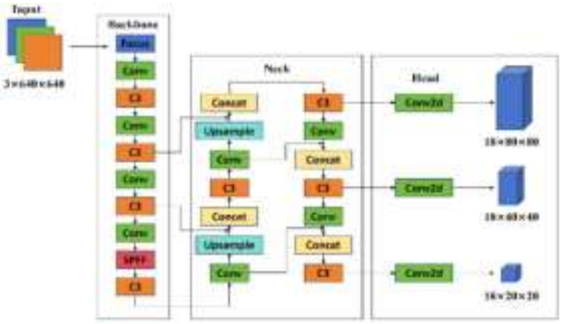


Figure. 1 YOLOv5s network architecture diagram

2.1 Lightweight backbone network

Although YOLOv5s, as a lightweight object detection algorithm, has certain advantages in terms of inference speed and model size, its backbone network, CSPDarkNet, still contains a large number of convolutional and residual structures, with high computational clutter, which is not conducive to deployment on resource-constrained embedded platforms such as unmanned aerial vehicles. To further reduce the number of parameters and enhance operational efficiency, this paper replaces the CSPDarkNet backbone network of YOLOv5s with the more efficient ShuffleNetV2 structure.

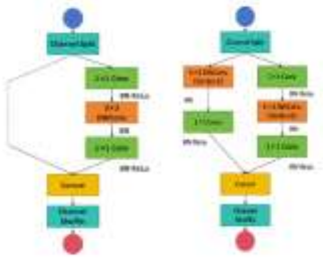


Figure. 2 ShuffleNetV2 unit structure diagram

ShuffleNetV2 achieves lightweight design by adopting a mixed strategy of group convolution and channel washing. The input features are divided into two branches. One is directly passed on, while the other undergoes 1×1 convolution for channel compression, 3×3 depth-separable convolution for feature extraction, and then 1×1 convolution to restore the channel dimension. Finally, channel washing and fusion are carried out to enhance the feature interaction and expression ability. This structure significantly reduces the computational load and parameter overhead while maintaining a strong feature extraction capability, thereby enhancing the real-time detection performance and deployment flexibility of the model on embedded platforms.

2.2 Improved Feature Pyramid Pooling

To enhance the detection capability of YOLOv5s for small targets and multi-scale targets, this paper replaces the original SPPF module with the ASPP module. By extracting multi-scale features in parallel through convolution with different void rates, this module effectively compensates for the information loss caused by Max pooling and enhances the global perception ability of the model. ASPP receives the high-level semantic feature map output by the backbone network (ShuffleNetV2) and outputs the enhanced features that integrate multi-scale context information, which are then transmitted to the Neck for further fusion and prediction.

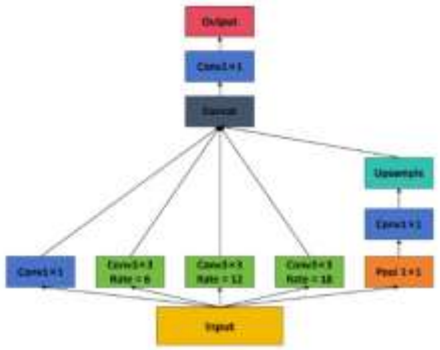


Figure. 3 ASPP network structure diagram

2.3 Attention mechanism module

Traditional convolution operations treat all channels and spatial positions equally, making it difficult to distinguish key areas from redundant backgrounds. Especially in the detection of small or dense targets, it is prone to causing information flooding and reducing detection accuracy. To enhance the model's perception ability of key information, this paper introduces the CA (Coordinate Attention) attention mechanism to highlight the responses of important regions in the feature map.

The CA module extracts channel-level global features through global average pooling and generates the attention weights of each channel through multi-layer perceptrons, thereby enhancing the feature expression of key channels and suppressing redundant information at the same time. Ultimately, the weighted feature map is input into the subsequent network to achieve more accurate representation and detection of key regions.

To enhance the model's detection ability for small and low-contrast targets, this paper introduces the CA attention mechanism into the backbone and Neck structure of the YOLOv5s. In the Backbone network, YOLOv5s originally adopted the CSP structure to build the Backbone, but in this article, it has been replaced by ShuffleNetV2.Therefore, we embed the CA module after each basic convolutional unit of ShuffleNetV2, with a focus on enhancing the expressive power of local region features. In the PAN structure of the Neck part, we insert the CA attention mechanism before and after the C3 module, enabling the model to refocus on important information during the multi-scale feature fusion stage.

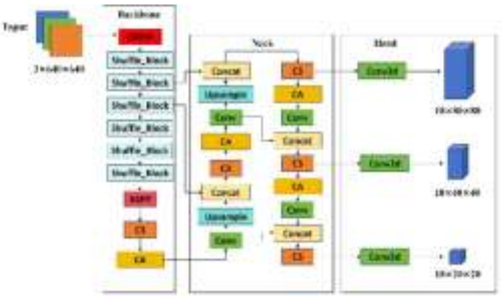


Figure. 4 Structure of YOLOv5s with the introduction of CA attention mechanism

2.4 Improvement of loss function

In the task of object detection, the regression loss function has a crucial impact on the positioning accuracy of the predicted bounding box. To enhance the model's adaptability to multi-scale targets, especially small ones, this paper introduces SIoU as the bounding box regression loss function to replace the original CIoU. SIoU not only takes into account the overlapping area between the predicted box and the real box, but also integrates multiple spatial factors such as Angle, distance and shape, optimizing the regression direction and scale matching from multiple dimensions, and has stronger geometric structure adaptability.In the object detection task, the regression loss function is used for prediction.

3. EXPERIMENT AND EVALUATION

3.1 Dataset selection

This paper uses the VisDrone2019 dataset to verify the effectiveness of the proposed algorithm[7]. This dataset was collected by the team from the Machine Learning and Data Mining Laboratory of Tianjin University. It was captured using various types of drones in different background scenes of 14 cities and rural areas in China, covering a wide range of weather and lighting conditions. The dataset contains a total of 8,599 images, including 6,471 in the training set, 548 in the validation set, and 1,580 in the test set.

3.2 Experimental evaluation indicators

In this study, to comprehensively evaluate the performance of the YOLOv5s improvement method, commonly used object detection evaluation metrics were selected. These metrics can measure the improvement effect of the model from multiple perspectives such as detection accuracy, model generalization ability, and operational efficiency. Among them, the accuracy rate represents the proportion of the number of correctly predicted positive samples to the total number of predicted positive samples, while the recall rate represents the proportion of the number of correctly predicted positive samples to the total number of positive samples. Average accuracy (AP) is the core indicator for measuring the performance of object detection. It is calculated through the precision and recall curves at different thresholds. Mean Average Precision (mAP) is the average of the AP values of all target categories, which is used to comprehensively evaluate the overall detection performance of the model. In addition, the detection speed (FPS) indicates the number of image frames that a model can process within a unit of time[8]. The higher the value, the faster the detection. FPS is mainly affected by factors such as model complexity, input image resolution, and the computing power of inference devices, and is usually compared based on the average detection speed under a unified hardware platform..

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Ablation experiment

To verify the effectiveness of each component in the model, this paper conducts ablation experiments to evaluate the impact of different modules on the functionality of the object detection algorithm under the same experimental conditions. The ablation experiment selected YOLOv5s as the benchmark model. To ensure the accuracy of the experiment, the same parameter configuration was adopted throughout the training process. The experimental results are shown in Table 1.

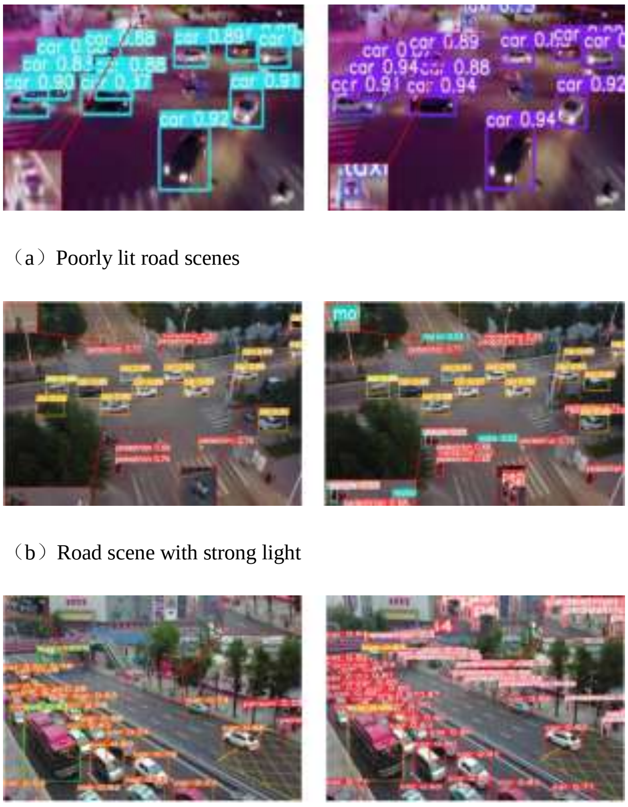
Table 1. Results of ablation experiments

Module	mAP_{50}	FLOPs	Params
Baseline	34.5	15.8	7.0
Baseline+ShuffleNet V2	32.7	8.12	3.8
Baseline+ShuffleNet V2+ASPP	33.8	8.13	3.8
Baseline+ShuffleNet V2+CA	35.0	8.13	3.8
Baseline+ShuffleNet V2+ASPP+CA	36.7	8.13	3.8
Baseline+ShuffleNet V2+ASPP+CA+SIoU	39.5	8.13	3.9

As can be seen from the experimental results in Table 1, after introducing the ShuffleNetV2 module alone, although the detection accuracy was slightly reduced, the number of model parameters and computational complexity were compressed to about half of the original model, significantly improving the real-time performance. To make up for the decline in accuracy, an ASPP module was added after it and a CA attention mechanism was introduced in the Neck part respectively. The detection accuracy was increased by 1.1% and 2.3% respectively, verifying its effectiveness. By further integrating the two, the accuracy was improved by 2.2% and 4.0% respectively compared to the original model and the ShuffleNetV2 model only introduced. After the SIoU loss function was finally introduced, the overall accuracy was improved by 5.0%, significantly enhancing the detection performance and inference speed on the basis of lightweight, meeting the real-time detection requirements.

4.2 Visual analysis

To evaluate the performance differences of the improved algorithm in this paper in unmanned aerial vehicle (UAV) target detection, typical scenarios in the VisDrone2019 test set were selected for comparative experiments, and the detection results were visually analyzed, respectively corresponding to road scenarios with insufficient light, strong light, and dense targets. The results show that the original YOLOv5s model has obvious missed detections, especially performing poorly in the detection of small targets. The improved model significantly reduced the missed detection rate in various complex scenarios, demonstrating stronger adaptability and robustness.



（a）Poorly lit road scenes

（b）Road scene with strong light

（c）Target-intensive road scenes

Figure. 5 Visualization comparison of detection effect before and after model improvement

5. CONCLUSION

Aiming at the problems of low detection accuracy of small targets and limited model deployment resources in unmanned aerial vehicle (UAV) aerial images, this paper proposes an improved method based on YOLOv5s. Firstly, the ShuffleNetV2 module is introduced in the backbone network section, effectively enhancing the feature extraction capability and significantly reducing the number of model parameters. In addition, the ASPP hollow space pyramid pooling module is adopted to replace the traditional pooling operation, which enhances the network's ability to extract multi-scale features and improves the detection accuracy of small targets. The CA attention mechanism was introduced to weight the channel dimension features, thereby enhancing the model's attention to the target area. In terms of the loss function, SIoU is adopted as the regression loss, which optimizes the bounding box positioning accuracy and accelerates the model convergence. The experimental results show that the method proposed in this paper achieves superior detection performance while maintaining the lightweight of the model. Although the model in this paper achieves a good balance between accuracy and lightweight, there are still some limitations. On the one hand, in extreme environments such as low light, strong interference, and complex scenes with blurred images, there is still room for improvement in the robustness of the model. On the other hand, the improvements in this paper mainly focus on network structure design and loss function optimization, and have not yet integrated emerging strategies such as self-supervised learning and multimodal fusion. The above research not only verified the effectiveness of the improved strategy, but also provided a research foundation and development direction for achieving higher-precision and more robust object detection in resource-constrained environments in the future..

6. REFERENCES

[1] SHAHI T B, XU Chengyuan, NEUPANE A, et al. Recent advances in crop disease detection using UAV and deep learning techniques [J]. Remote Sensing, 2023, 15(9): 2450.

[2] RAMACHANDRAN A, SANGAIAH A K. A review on object detection in unmanned aerial vehicle surveillance [J]. International Journal of Cognitive Computing in Engineering, 2021, 2: 215-228.

[3] YUAN Yubin, WU Yiquan, ZHAO Langyue, et al. Research progress of UAV aerial video multi-object detection and tracking based on deep learning [J]. Acta Aeronautica et Astronautica Sinica, 2023, 44(18): 028334.

[4] Zhang X, Li Y. An improved YOLOv5 algorithm for small object detection [C]// Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Hotel Royal, Singapore: IEEE, 2023: 1842-1847.

[5] ZHOU Huaping, GUO Wei. Improved YOLOv5 network in application of remote sensing image object detection [J]. Remote Sensing Information, 2022, 37(5): 23-30.

[6] WEI Yangyang. Research on Underwater Biological Detection Technology Based on Improved YOLOv5s Algorithm [D]. Wuhu: Anhui University of Engineering, 2023.

[7] ZHU Pengfei, WEN Longyin, BIAN Xiao, et al. Vision challenge meets drones: Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 777-793.

[8] Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the clear motmetrics[J]. EURASIP Journal on Image and Video Processing, 2008, 2008:1-10.