ISSN 2616-5775 Vol. 6, Issue 12: 12-19, DOI: 10.25236/AJCIS.2023.061202

Small Object Detection Algorithm in Drone Aerial Images Based on Improved YOLOv5

Feng Fei[1,a,*], **Hu Yu**[1,b]

[1]College of Big Data Engineering, Kaili University, Kaili, China
[a]991656217@qq.com, [b]383849383@qq.com
[*]Corresponding author

Abstract: *In response to the low accuracy of conventional object detection algorithms in small object detection during drone aerial missions, this paper proposes an improved YOLOv5 algorithm for small object detection. Firstly, we introduced a 160x160 small object detection head and removed the 20x20 large object detection head, adopting a strategy of fusing feature maps of P2, P3, and P4 scales to enhance small object detection performance. Secondly, by eliminating the P5 scale network layers in the backbone network, we strengthened the high-order semantic information of the P4 layer. Finally, we utilized adaptive spatial feature fusion technology to replace the direct concatenation method in the Neck network for merging different scale feature maps, optimizing the fusion effect of feature maps of different scales. The experimental results on the VisDrone2019 small object dataset show that, compared to the original YOLOv5, the proposed algorithm in this paper achieved significant improvements of 7.3%, 7.1%, and 9.2% in Precision, Recall, and mAP@0.5, respectively. Furthermore, the model's parameter quantity was reduced by 40%, indicating that the improved model presented in this paper has superior performance in drone small object detection tasks and is highly practical.*

Keywords: *YOLOv5, Aerial Images, Small Object Detection, Adaptive Spatial Feature Fusion*

1. Introduction

As drone technology continues to advance, its applications have gradually integrated into various aspects of our daily lives. Drone aerial photography technology is extensively used in areas such as transportation, agriculture, public security, and geological exploration. However, due to the altitude at which drones operate, the captured images often contain numerous small targets against complex and variable backgrounds, posing significant challenges for target detection in drone images.

Currently, aerial image target detection is mainly categorized into two methods: traditional target detection algorithms and deep learning-based methods. Traditional algorithms primarily rely on manually extracted features, such as HOG[1] and Haar features[2]. However, these methods have low generalization capabilities in small target detection and are challenging to apply widely in aerial images.

With the innovative development of convolutional neural networks, deep learning-based aerial image object detection algorithms have shown superior performance and have become the mainstream technology for aerial image object detection[3]. For instance, Kisantal et al.[4] improved the network's contribution to small targets by increasing their proportion in the dataset through copying and pasting. However, this method did not fully utilize the feature information of small targets. Liu et al.[5] introduced the PAN network architecture, which enhances the model's multi-scale feature extraction capabilities by merging different network layers in a top-down and bottom-up approach. However, during the fusion process, the method of direct concatenation was used, which did not fully integrate feature information from different scales. TPH-YOLOv5[6] used the CBAM attention mechanism and a Transformer[7] structure in the detection head, improving the feature utilization and prediction accuracy of the network. Yang[8] proposed QueryDet, using a novel query mechanism to accelerate the inference speed of target detection based on feature pyramids. Chen Jiahui et al.[9] designed a multi-scale feature extraction module with a residual structure, C3Res2Block, and a Decoupled Head, effectively improving the model's localization and detection accuracy. In summary, although existing small target detection algorithms have improved detection performance to some extent, there is still considerable room for improvement. This paper's work on small target detection is as follows:

(1) Optimizing the network structure by adding a 160x160 small target detection head to enhance the model's detection capability for small targets; removing the 20x20 large target detection head, which has limited contribution to small target detection; merging feature maps of P2, P3, and P4 scales for target detection; deleting the network layers of the P5 scale in the backbone network to enhance the high-level semantic information of P4.

(2) Optimizing the fusion of features at different scales by using adaptive spatial feature fusion technology to replace the direct cascading method in the neck network for merging feature maps of different scales, effectively resolving the information conflict between different layers and improving the effect of feature fusion at different scales.

2. YOLOv5 Network Structure

In the field of object detection, YOLOv5 is one of the widely used algorithms. This algorithm comes in five versions based on the main differences in network depth and width: YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. These versions range from the smallest YOLOv5n to the largest YOLOv5x, with the primary distinctions being in the network's depth, width, and the number of parameters, while maintaining a consistent overall structure. Among them, YOLOv5s is widely used for real-time object detection tasks due to its balance between detection accuracy and speed. In this paper, YOLOv5s is selected as the base network for the study of aerial image object detection.

The YOLOv5 network architecture is divided into four main parts: Input, Backbone, Neck, and Head. The Input part is responsible for image preprocessing, including data augmentation, anchor box calculations, and image scaling. The Backbone, which is the core of the model, focuses on extracting basic features of the image and includes CBS, SPPF, and C3 modules. CBS consists of Convolution, Batch Normalization, and SiLU activation functions. The SPPF layer speeds up the process by fusing feature maps with different receptive fields and using smaller pooling kernels instead of larger ones. The C3 module combines CBS units and residual modules, enhancing feature extraction and learning capabilities. The lower layers of the Backbone extract basic features like edges, colors, and textures, while the higher layers extract more complex and abstract features, providing rich multi-scale feature information for feature fusion and classification. The Neck part uses the FPN+PAN structure, merging multi-scale features in both top-down[10] and bottom-up[11] approaches to enhance detection capabilities for targets of varying sizes. Finally, the Head part classifies the multi-scale features merged in the Neck. The structural diagram of YOLOv5s is shown in Figure 1.

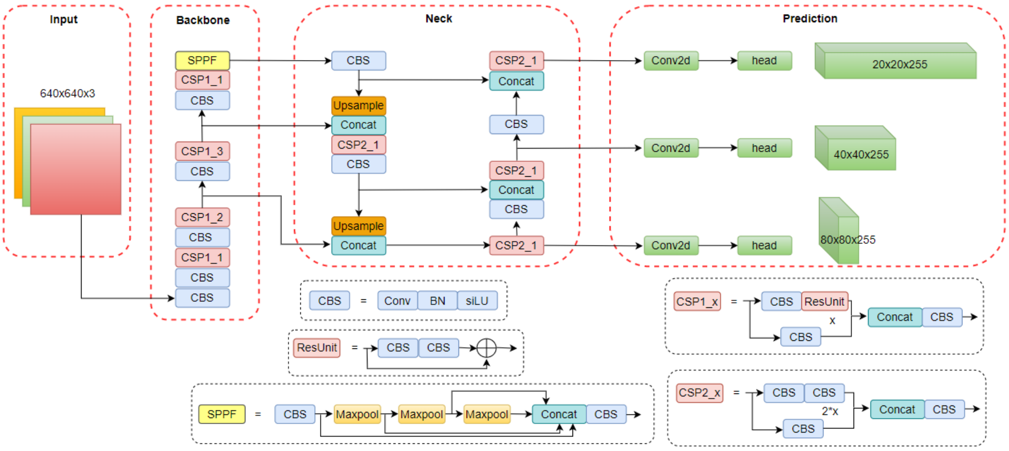

Figure 1: YOLOv5 Network Structure.

3. YOLOv5s Model Improvements

3.1 Optimizing Network Structure

In the Backbone of YOLOv5s, as the network deepens, five sets of feature maps P1, P2, P3, P4, and P5 are generated. When the input image size is 640x640, these feature maps are produced through 2x, 4x, 8x, 16x, and 32x downsampling, resulting in feature maps of sizes 320x320, 160x160, 80x80, 40x40, and 20x20, respectively. Shallow, larger feature maps are used to detect smaller targets, while deeper, smaller feature maps are for detecting larger targets. YOLOv5s primarily fuses feature maps P3, P4, and P5, enabling the model to detect regional targets of sizes 8x8, 16x16, and 32x32 in the original image. However, the detection of tiny targets smaller than 8x8 is not ideal.

Based on the distribution graph of the target size relative to the entire image in the training dataset, as shown in Figure 2, it is evident that there are a large number of tiny targets. To address this issue, some researchers have added a 160x160 small target detection head to the existing network[12][13], which to some extent improved the detection of small targets, but also increased the model's parameter count. Building on this, this paper makes further improvements by adding a 160x160 small target detection head while removing the 20x20 large target detection head, as the latter contributes limitedly to small target detection and adds extra parameters. The model now fuses feature maps from P2, P3, and P4 layers for target detection. Since the last layer in the Backbone contains the richest semantic information, removing the P5 layer makes the P4 layer the highest, enhancing its semantic information and thereby improving the overall performance of the network. The improved network structure is shown in Figure 3.

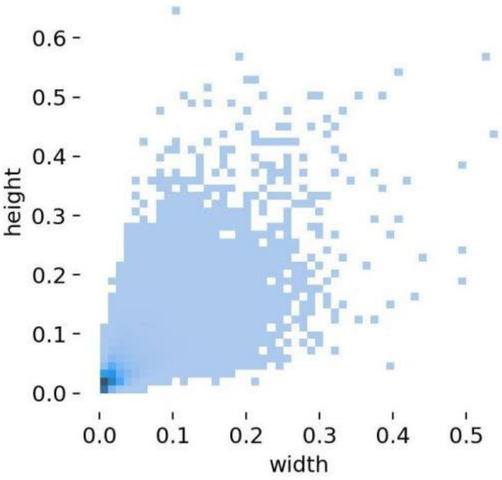

Figure 2: Distribution Graph of Target Aspect Ratio.

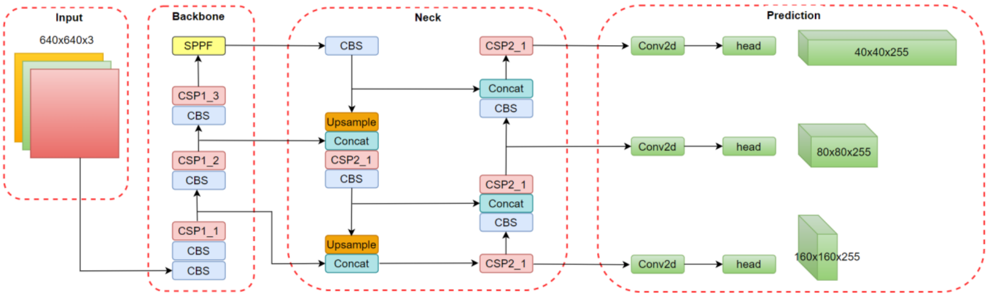

Figure 3: Optimized Network Structure.

3.2 Adaptive Spatial Feature Fusion

Multi-scale fusion is a key concept in the field of computer vision, enriching feature information by fusing feature maps of different scales, thereby enhancing the model's performance. In deep neural networks, features extracted by different network layers vary: lower layers capture basic and general features (such as edges, colors, textures), while higher layers extract more semantically rich information. YOLOv5s employs an FPN+PAN structure for feature fusion, combining both top-down and bottom-up fusion methods. In YOLOv5s, feature fusion typically uses a direct concatenation approach, but this method does not effectively resolve the conflict of feature information between different layers.

To improve this, this paper proposes the use of Adaptive Spatial Feature Fusion (ASFF) technology [14] to fuse feature maps of different scales for more effective feature fusion, as shown in the network structure in Figure 4. ASFF technology adaptively learns the spatial weights of feature maps at different scales between layers, effectively resolving the information conflict between layers. Taking ASFF_3 as an example, P2, P3, and P4 represent feature maps of different scales extracted through the Backbone network. After Neck feature fusion, the feature maps at different scales obtained are L1, L2, and L3, respectively. These feature maps are first compressed through 1×1 convolution to the same number of channels as L3. Then, they are adjusted to the same dimensions as L3 through 4x and 2x upsampling, resulting in RL1 and RL2. Next, RL1, RL2, and L3 generate weight parameters α, β, and γ through 1×1 convolution. Finally, RL1, RL2, and L3 are each multiplied by these weights and then added together to produce a new fused feature map. The formula for adaptive spatial feature fusion can be specifically expressed as:

$$y^l_{\mathrm{ij}} = \alpha^l_{ij} \times x^{1\to l}_{ij} + \beta^l_{ij} \times x^{2\to l}_{ij} + \gamma^l_{ij} \times x^{3\to l}_{ij} \tag{1}$$

In the formula: y^l_{ij} represents the new feature map obtained through ASFF; $x^{n\to l}_{ij}$ represents the feature vector at position (i, j) from layer n to layer l; α^l_{ij}、β^l_{ij}、γ^l_{ij} represents the weight values of three different levels of feature maps, which are normalized to meet the condition $\alpha^l_{ij} + \beta^l_{ij} + \gamma^l_{ij} = 1$, α^l_{ij}、β^l_{ij}、$\gamma^l_{ij} \in [0,1]$ using the Softmax function.

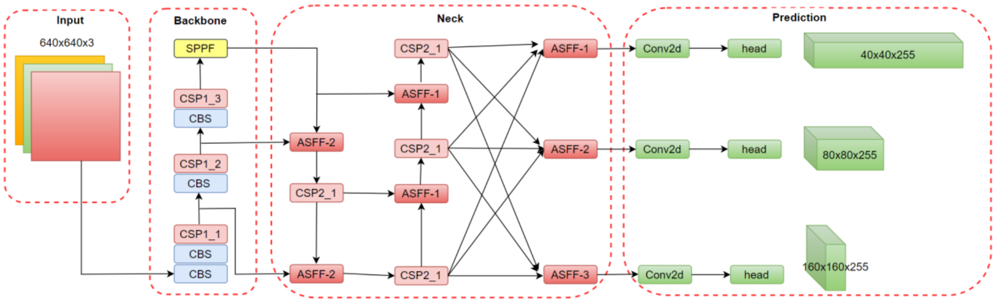

Figure 4: Network Structure with Adaptive Spatial Feature Fusion Technology.

4. Experimental Results and Analysis

4.1 Experimental Dataset

The experimental dataset used in this paper is the VisDrone2019 dataset[15], collected and organized by the AiskYeye team from the Machine Learning and Data Mining Lab at Tianjin University. The VisDrone2019 dataset comprises 288 video clips 261908 frames of video, and 10209 static images. Of these 6471 images are used for training the model 548 for validation, and 3190 for testing. The dataset's images cover various scenes from daily life and include ten different categories: pedestrian, people, bicycle, motor, car, bus, truck, van, awning-tricycle, and tricycle, with a total of 2.6 million annotations. The targets in the images are relatively small, making the dataset suitable for researchers studying small object detection and recognition.

4.2 Experimental Environment and Evaluation Metrics

The experimental environment configuration for this paper is outlined in Table 1.

The experiment utilizes the Stochastic Gradient Descent (SGD) method to optimize and adjust the network, with some of the hyperparameters settings shown in Table 2.

The evaluation metrics used in this paper are as follows: Precision, which indicates the proportion of correctly predicted targets among all predictions made by the model; Recall, which refers to the proportion of correctly predicted targets among all actual targets; AP (Average Precision), measuring the average accuracy of the model for each specific target; and mAP@0.5, assessing the detection capability of the trained model across all detection targets. The specific formulas for these calculations are as follows:

$$\Pr ecision = \frac{TP}{TP+FP} \tag{2}$$

$$\operatorname{Re} call = \frac{TP}{TP+FN} \tag{3}$$

$$AP = \frac{1}{N}\sum_{1}^{N} precision \tag{4}$$

$$mAP = \frac{1}{N}\sum_{1}^{N} AP \tag{5}$$

In the evaluation metrics for categorical models, if the focus is on the correctness of a specific category's label, it equates to a binary classification task, with only two types of classification targets: positive and negative. Binary classification tasks use a confusion matrix for definition, as shown in Table 3.

Table 1: Configuration of the Experimental Environment.

Environmental configuration	Version
Operating system	Windows
CPU	Intel i5-13490,2.5GHz
GPU	GeForce RTX 3090,24GB Graphics Memory
CUDA	11.8
Pytorch	2.0.1
python	3.11.4

Table 2: Settings of selected hyperparameters.

Training Parameter	Parameter Value
Initial Learning Rate(lr0)	0.01
Cyclic Learning Rate (lrf)	0.01
Momentum	0.937
Weight Decay	0.0005
Batch Size	16
Training Epochs	200

Table 3: Confusion matrix.

Confusion matrix		True Value	
		Positive	Negative
Predicted Value	Positive	TP	FP
	Negative	FN	TN

4.3 Experiment

4.3.1 Ablation Study

This paper designed four sets of ablation experiments to compare and analyze the effects of the improvements. The results of these experiments are summarized in Table 4.

The analysis of the experimental data indicates that the improved Model 1 showed increases of 3.3%, 4.7%, and 5.5% in Precision, Recall, and mAP@0.5, respectively, compared to the original YOLOv5s. This result confirms that adding a 160x160 small object detection head effectively enhances the model's ability to detect small objects, but also leads to a 2.9% increase in the model's parameters. Meanwhile, the improved Model 2 showed additional improvements of 0.6%, 0.6%, and 0.7% in these three metrics over Model 1, along with a 25% reduction in model parameters. This demonstrates that the large object detection head contributes limitedly to small object detection and instead increases the model's parameter count.

On the other hand, although the improved Model 3 saw a decrease of 0.9% in Precision compared to Model 2, its Recall and mAP@0.5 increased by 0.5% each. More importantly, the model's parameters decreased by 63%, indicating that removing the P5 layer from the Backbone to enhance the semantic information of the P4 layer can improve model performance. The final improved Model 4 showed increases of 4.3%, 1.3%, and 2.5% in Precision, Recall, and mAP@0.5, respectively, compared to Model 3. This improvement validates the superiority of adaptive spatial feature fusion technology over direct concatenation fusion.

As the models in this paper are improvements based on YOLOv5s, a comparison between the final improved Model 4 and the original YOLOv5s network was conducted. The results showed that the improved Model 4 achieved significant improvements of 7.3%, 7.1%, and 9.2% in Precision, Recall, and mAP@0.5, respectively, while reducing the model's parameters by 40%. This outcome demonstrates the superiority of the model proposed in this paper to some extent.

Table 4: Ablation Experiment Results.

Methods	Method One	Method Two	Method Three	Method Four	Precision (%)	Recall (%)	mAP@ 0.5(%)	Params (10^6)
Yolov5s	×	×	×	×	45.3	33.0	32.6	7.0
Improved Model 1	√	×	×	×	48.6	37.7	38.1	7.2
Improved Model 2	√	√	×	×	49.2	38.3	38.8	5.4
Improved Model 3	√	√	√	×	48.3	38.8	39.3	2.0
Improved Model 4	√	√	√	√	52.6	40.1	41.8	4.2

Note: Method One(Adding a 160x160 small object detection head), Method Two(Removing the 20x20 large object detection head),Method Three(Removing the P5 layer of the Backbone),Method Four(Using ASFF fusion technology),√ indicates the use of this method for improvement, × indicates not using this method for improvement.

4.3.2 Comparison with Different Models

To validate the advancement of the improved model proposed in this paper, a comparison was made with other current mainstream models on the VisDrone2019 dataset. The comparison of mAP@0.5 for different models is shown in Table 5. The table reveals that compared to the models Faster RCNN[16], QueryDet, and YOLOv4[17], the mAP@0.5 of the improved model increased by 13%, 11.6%, and 11.3%, respectively. This demonstrates the superior performance of the model improved from YOLOv5s in small object detection.

Table 5: Comparison Results of Different Models.

Model	mAP@0.5(%)
Faster RCNN	28.8
QueryDet	30.2
YOLOv4	30.5
The improved Model 4 of this paper	41.8

4.3.3 Visualization of Detection Results

The detection results of the YOLOv5 model and the improved Model 4 under different environmental conditions were visualized, as shown in Figures 5, 6, and 7. The improved Model 4 outperformed the YOLOv5 model in terms of detection quality and precision for small objects, making the proposed algorithm more suitable for detecting small objects in aerial images.

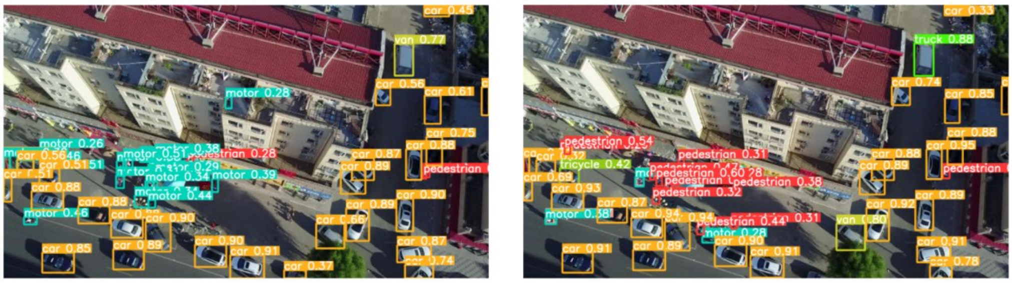

Figure 5: Comparison of detection effects in high-altitude shooting scenes between YOLOv5 (left) and the improved Model 4 of this paper (right).

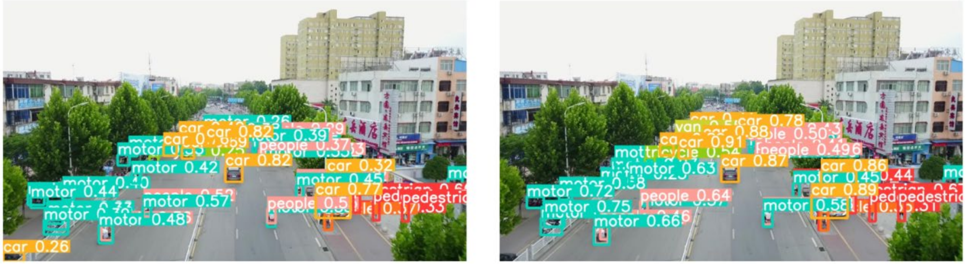

Figure 6: Comparison of detection effects on daytime road scenes between YOLOv5 (left) and the improved Model 4 of this paper (right).

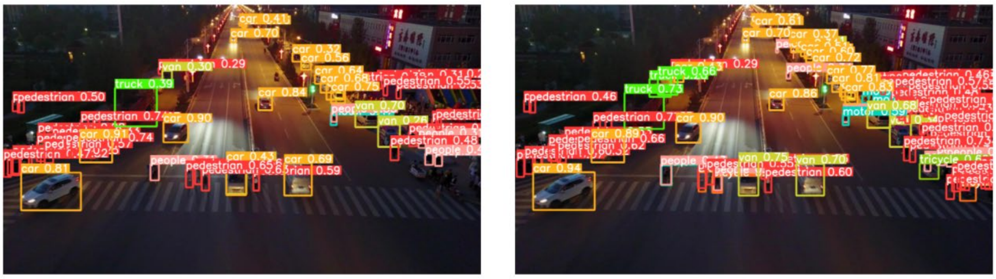

Figure 7: Comparison of detection effects in night-time dark scenes between YOLOv5 (left) and the improved Model 4 of this paper (right).

5. Conclusion

Faced with the challenge of numerous small targets in drone aerial images, conventional object detection algorithms often perform poorly in such scenarios. This study proposed an improved algorithm for small object detection, achieving significant effectiveness in drone aerial image target detection. Firstly, to enhance the algorithm's capability to extract features of small objects, we designed a 160x160 small object detection head. Secondly, we removed the 20x20 large object detection head, which contributed minimally, a strategy that not only improved detection performance but also reduced the model's parameter count. Additionally, we optimized the model's backbone network by removing the P5 layer and enhancing the high-level semantic information of the P4 layer, further improving the overall feature extraction capability of the model. Finally, this paper adopted adaptive spatial feature fusion technology to optimize the integration of feature maps of different scales, replacing the traditional direct concatenation method in the neck network. This allowed the model to adaptively fuse multi-scale feature information for more precise feature representation. Compared to the original YOLOv5 and other mainstream object detection algorithms, the algorithm proposed in this study demonstrated excellent performance in small object detection. In the future, we will continue to research how to further improve the detection accuracy of small objects in aerial images without increasing network complexity.

Acknowledgement

Funding Project: Social Service Robot for Basic Human Motion Recognition, Guizhou Education and Technology(2022)No. 366

References

[1] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, 2005, 1: 886-893.

[2] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]. Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. IEEE, 2001, 1: I-I.

[3] Li Xiaolin, Liu Dadong, Liu Xinman, et al. Improved YOLOv5 Algorithm for Drone Aerial Image Object Detection [J/OL]. Computer Engineering and Applications.2023:1-13. http://kns.cnki.net/kcms/detail/11.2127.TP.20231013.0942.002.html.

[4] Kisantal M, Wojna Z, Murawski J, et al. Augmentation for small object detection[J]. arXiv preprint arXiv:1902.07296, 2019.

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

[6] 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]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2778-2788

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. arXiv preprint arXiv: 1706.03762, 2017.

[8] Yang C, Huang Z, Wang N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection[C]//Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2022: 13668-13677.

[9] Chen Jiahui, Wang Xiaohong. Improved YOLOv5 Algorithm for Dense Small Object Detection in Drone Aerial Images [J/OL]. Computer Engineering and Applications.2023:1-11.http://kns.cnki.net /kcms/ detail/11.2127.TP.20230901.1731.002.html.

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

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

[12] Wang Haoxue, Cao Jie, Qiu Cheng, Liu Yaohui. Aerial Image Multi-Object Detection Method Based on Improved YOLOv4 [J]. Electronics Optics & Control, 2022, 29(05): 23-27.

[13] Chen Weibiao, Jia Xiaojun, Zhu Xiangbin, Ran Erfei, Xie Hao. Drone Aerial Image Object Detection Based on DSM-YOLO v5 [J]. Computer Engineering and Applications, 2023, 59(18): 226-233.

[14] Liu S, Huang D, Wang Y. Learning spatial fusion for single-shot object detection[J]. arXiv preprint arXiv:1911.09516, 2019.

[15] ZHU P, WEN L, DU D, et al. Detection and Tracking Meet Drones Challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 7380-7399.

[16] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[17] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXi preprint arXiv:2004.10934, 2020.