# Small Object Image Detection Based on YOLOV5 Improved Algorithm

**Mawei Chen[a,*], Peng Yang[b], Zhongguo Liu[c]**

School of Computer Science, Nanjing Audit University, Nanjing, Jiangsu, China

[a]897133978@qq.com, [b]1356219836@qq.com, [c]11ylab@21cn.com

*Corresponding author

**Keywords:** Small Goals, YOLOV5, CA Attention Mechanism, BiFPN, Effective Improvement

**Abstract:** For long-distance aerial images with blurring, unclear features, and small volume. The existing algorithms are not ideal for detecting small targets.This article proposes an improved algorithm model for YOLOV5. Firstly, we improved the Backbone layer of YOLOV5.Adjust the original fast pyramid layer SPPF to an SPPFP layer. Then, we added a CA attention mechanism to improve the recognition of small targets.Secondly, in the Neck layer, adjust the road force aggregation network PANet to BiFPN. Finally, we adjust the loss function CIOU of the original network to EIOU, improve the robustness and generalization ability of the model, and accelerate the convergence speed of the network. The improved model mAP in this article has increased by 10%, indicating that the performance of our proposed model has been effectively improved and has certain application prospects.

## 1. Introduction of YOLOV5

YOLOv5 is divided into four parts: Input, Backbone, Neck and Output.[1]the improvement model proposed in this paper is based on YOLOv5s in YOLOv5 version 6.1. the structure of the YOLOv5 model in version 6.1 is shown in Figure 1. [2]

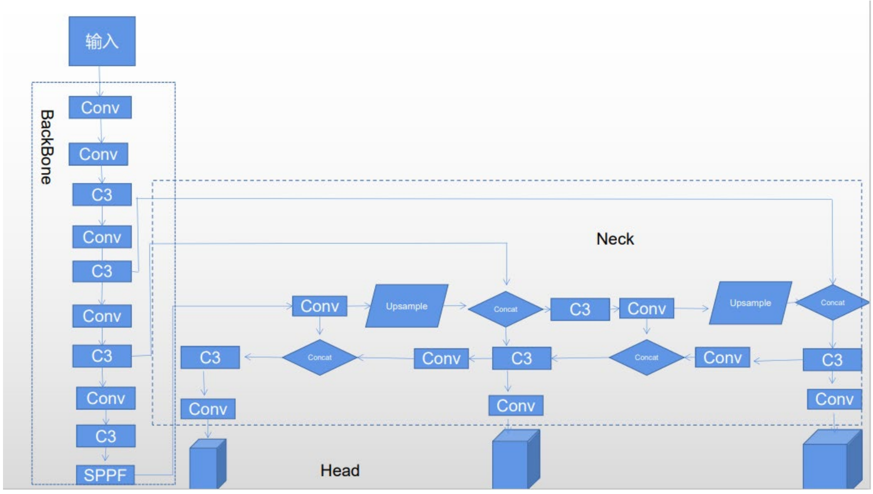
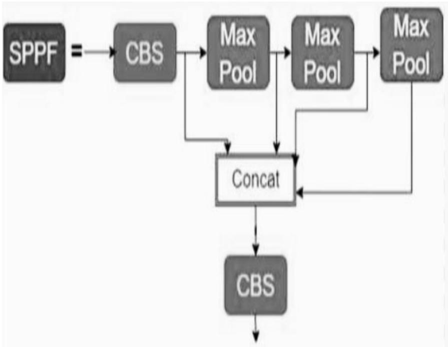


Figure 1 YOLOV5 network structure.

## 2. Model improvement of YOLOV5

### 2.1. SPPF improved to SPPFP

SPPF is available as the last module of Backbone in YOLOv5 version 6.1. [3]the SPPF module is a series of three 5 x 5 sized MaxPool layers through which the input is passed sequentially and the Concat operation is performed on the output of the three MaxPool layers before the CBS operation is performed. the structure of SPPF is shown in Figure 2. MaxPooling and jump connection at different scales enable the image to learn features at different scales, and thus fuse local and global features to enrich the image features.[4] Among them, maximum pooling divides the image into several rectangular regions and outputs the maximum value for each subregion. Although the maximum pooling operation can reduce redundant information, it also tends to cause the loss of feature information.



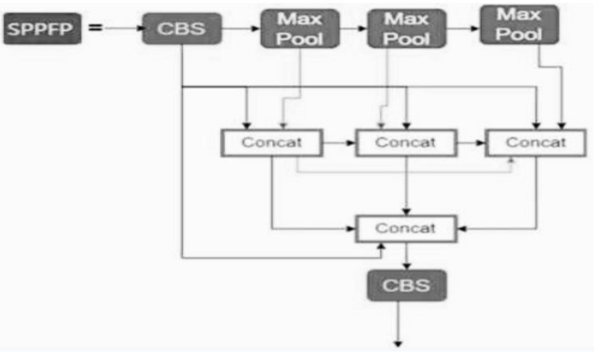


Figure 2 SPPF structure.

Figure 3 SPPFP structure.

In this paper, we improve SPPF by borrowing the idea of constructing dense links from BenseNet to enhance feature reuse. Then, we obtain the SPPF module, which reduces the feature information loss due to the maximum module pooling. the SPPF module can obtain better global information about the small target. Figure 3 shows the structure of SPPFP.[5]

### 2.2. Add CA attention mechanism

CA mechanism is a new mechanism to embed location information into channel attention. At the shallow level of the network (e.g. image-level), the extracted spatial feature map is too large and the number of channels is too small, resulting in the obtained channel weights not summarizing the specific features and the extracted spatial weights not being generalized enough due to the small number of channels.[6] In the later layers of the network, too many channels tend to lead to overfitting. More critically, the closer to the classification layer, the more sensitive the action of the attention mechanism is to the classification results, which affects the decision of the classification layer. Therefore, in this paper, the attention mechanism is added to the middle part of the network.The structure as figure 4.

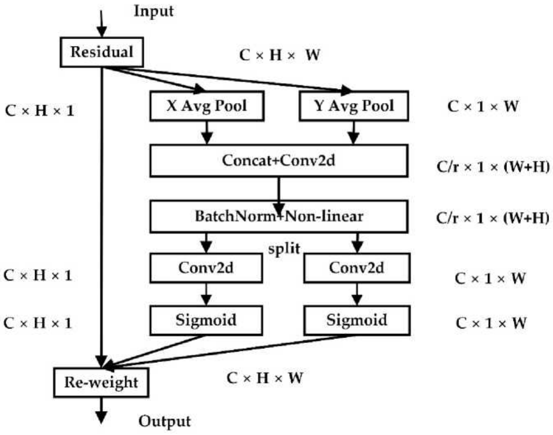


Figure 4 CA structure.

### 2.3. Adjusting the neck network PANet to BiFPN

Bidirectional Feature Pyramid Network (BiFPN) is a typical complex bidirectional feature fusion pyramid structure. [7]First, BiFPN simplifies the feature network by removing intermediate nodes with only one input edge on top of the standard feature pyramid. Second, additional edges are added at the same layer between the input and output nodes to fuse more features at low cost. [8]Finally, BiFPN treats each bi-directional (top-down and bottom-up) path as a single feature network layer and repeats the same layer multiple times to achieve a higher level of feature fusion. As shown in Figure 5, P3-P7 in the figure represent the different levels of fused features. One represents a top-down path and one represents a bottom-up path, and jump arrows indicate adding additional edges to the input and output nodes in the same layer. [9]Since different input features have different resolutions, different weights should be attributed to the final output at each node where feature fusion is performed. Therefore, BiFPN introduces training weights that add additional weights to each input to adjust the contribution of different inputs to the output feature map.BiFPN selects weights using a fast normalized fusion that divides the sum of all values directly by the weights and normalizes the normalized weights to a range from 0 to 1.[10]

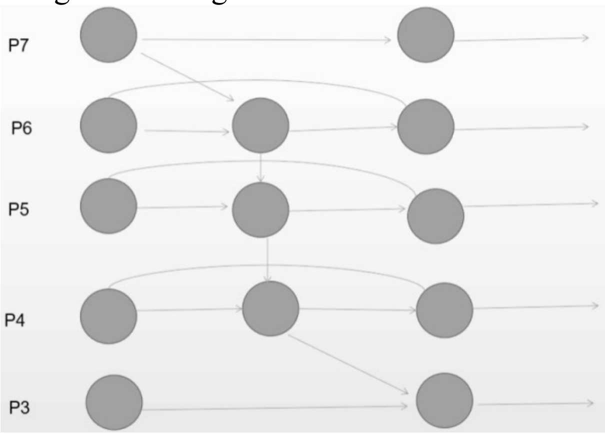


Figure 5 BiFPN structure.

## 2.4. Improving the loss function of YOLOV5

During model training, CIoU Loss only calculates the distance between the prediction frame and the center point of the real boundary, the overlap area and aspect ratio of the two frames to perform the regression of the bounding box, but does not take into account the direction of matching the prediction frame with the real frame, which leads to slow convergence and low efficiency of the network, and may produce worse models due to the arbitrary matching of the prediction frame during training. In order to solve the problems of the currently used bounding box loss function. A new function EIoU is introduced. as Eqs. (1) The loss function contains three components: overlap loss, center distance loss, and width-height loss. the first two components continue the approach in CIOU, but the width-height loss directly minimizes the difference between the width and height of the target box and the anchor box, making the convergence faster. Where Cw and Ch are the height and width of the minimum external box covering the two Boxes.

$$L_{EIOU} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{C_w^{\ 2}} + \frac{\rho^2(h, h^{gt})}{C_h^{\ 2}} \tag{1}$$

The improved overall structure of YOLOV5 is shown in Figure 6:

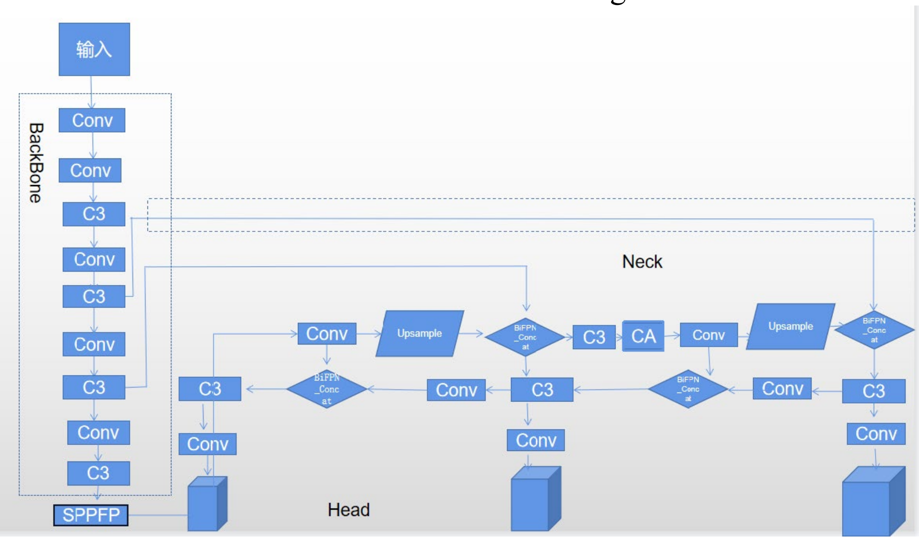


Figure 6 YOLOV5 improved structure.

## 3. Experimental results and analysis

### 3.1. Evaluation Criteria

The evaluation system of this test includes recall rate R (recall), precision rate P (precision), average precision AP (average precision) and mean average precision mAP (mean average precision), and the specific formulas as follows:

$$mAP = \frac{\sum_{i=1}^{k} AP_n}{k} \tag{2}$$

### 3.2. Comparison experiments

#### 3.2.1. Adding CA attention mechanism in different locations

The CA attention mechanism ablation experiments made in this paper were added to the backbone after the SPPF layer and in the middle of the neck network, respectively, to further evaluate the effect of CA addition location on the algorithm performance. The experimental results are shown in Table 1. We compare the accuracy values of CA added at different positions and conclude that the best results are obtained when CA is added at the Neck position.

Table1 Different CA attention experiments.

| Add Location | mAP |
|---|---|
| Not added | 0.801 |
| CA Added in Backbone | 0.843 |
| CA added in Neck | 0.865 |

#### 3.2.2. Comparative analysis of different loss functions

We verify experimentally to illustrate that using EIoU loss function can improve the performance of the model, speed up the detection and make the network more stable. the YOLOV5x model uses CIoU, GIoU, DIoU and EIoU respectively to compare the experiments. Through Table 2, we can see that EIoU has better performance.

Table 2 Comparison of different loss functions.

| Model | mAP |
|---|---|
| YOLOV5X-CIoU | 0.810 |
| YOLOV5X-GIoU | 0.809 |
| YOLOV5X-DIoU | 0.798 |
| YOLOV5X-EIoU | 0.830 |

#### 3.2.3. Ablation verification experiments and analysis

In order to verify the effectiveness of several improvement methods proposed in this paper. We did ablation experiments for each models.The results of the five models are shown in Table 3:

Table 3 YOLOV5 modified model ablation experiments.

| Model | mAP | Precision | Recall |
|---|---|---|---|
| YOLOV5 | 0.808 | 0.817 | 0.716 |
| YOLOV5_SPPFP | 0.853 | 0.834 | 0.803 |
| YOLOV5_SPPFP_CA | 0.873 | 0.857 | 0.814 |
| YOLOV5_SPPFP_CA_BiFPN | 0.896 | 0.863 | 0.826 |
| YOLOV5_SPPFP_CA_BiFPN_EIoU | 0.902 | 0.895 | 0.835 |

## 4. Conclusion and outlook

In this paper, we propose an improved YOLOV5 model. Firstly, the SPPF is enhanced to SPPFP in the last layer of the backbone network. secondly, the CA attention mechanism is introduced in the neck network, which allows our model to extract the image Rui features quickly. Then the PANet is adjusted to BiFPN. lastly, the loss function is modified and the original function is replaced with EIoU. the final improved model has a 10.2% improvement in mAP over the original YOLOV5 compared to that on VisDrone2019. The detection time and detection accuracy have been significantly improved.

Target detection has a wide range of application scenarios, such as aerial fear image military applications, smoke detection, and traffic recognition. The current detection network still does not have high enough detection accuracy for small target images, and the future direction of improvement could be to replace backBone's C3 module with MobileNetV3 to create a lightweight network with higher performance.

## References

[1] Wang M, Li Q, Gu Y, et al. SCAF-Net: Scene Context Attention-Based Fusion Network for Vehicle Detection in Aerial Imagery[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19(2): 1–5.

[2] Chen P, Huang L, Xia Y, et al. Detection and recognition of road traffic signs based on Mask R-CNN for UAV images[J]. Remote Sensing of Land Resources, 2020, 32(4): 61-67.

[3] Wu Xiaohui, Tian Qichuan. A review of traffic sign recognition methods[J]. Computer Engineering and Applications, 2020, 56(10):20-26.

[4] Jia Kexin, Ma Zhenghua, Zhu Rong, et al. Attention mechanism to improve lightweight SSD model for sea surface small target detection [J]. Chinese Journal of Graphical Graphics, 2022, 27(4): 1161-1175.

[5] Yuan S, Ma X, Liu S. Improved YOLOv3 algorithm for pedestrian-vehicle target detection[J]. Science, Technology and Engineering, 2021, 21(8): 3192-3198.

[6] Gu YL, Zong X. A review of deep learning-based target detection research [J]. Modern Information Technology, 2022, 6(11): 76-81.

[7] Zhao Xingke, Li Minglei, Zhang Gong, et al. Unmanned airborne thermal infrared image target detection method based on significant map fusion[J]. Journal of Automation, 2021, 47(9): 2120-2131.

[8] A.Almahairi, N.Ballas, T.Cooijmans, et al. Dynamic capacity networks[C]. International Conference on Machine Learning, 2016, 2549-2558

[9] XIAO Jie, LIU Gang, GUO Guofa. Weed detection method based on multiscale fusion module and feature enhancement[J]. Journal of Agricultural Machinery, 2022, 53(04): 254-260.

[10] WANG Ruiqing, WANG Huiqin, WANG Ke. Fire smoke detection by fusing detailed features with hybrid attention mechanism[J]. Liquid Crystal and Display, 2022, 37(07): 900-912.