Smoking Detection Algorithm Based on Improved YOLOv5

Yingying Cao[1,a], Mingkun Xu[1,b,*]

[1]School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing, China
[a]3378912452@qq.com, [b]165605847@qq.com
[*]Corresponding author

Abstract: *Aiming at the current problems in smoking detection in public places, such as many small and medium-sized cigarette targets, low pixels occupied by cigarette targets, indistinct characteristics of cigarette targets, and difficult detection, an improved smoking detection algorithm YOLOv5_EC is proposed. Based on the YOLOv5, this paper redesigns the neck structure and proposes the EFFN (Enhanced Feature Fusion Neck) structure, which fuses three adjacent feature maps of different sizes to better retain the positioning information of the target, further improves the feature expression ability of small objects, and then introduce the CBAM attention module in the network, so that the model can focus on the important areas of the image, suppress the interference of irrelevant information, enhances the model's ability to learn features, and improves the accuracy of model detection. Experiments show that the mAP@0.5 of the model proposed in this paper is improved by 1.5% compared with the original YOLOv5 model, and the improved model can effectively identify smoking behavior in actual scenes.*

Keywords: *smoking detection, YOLOv5 model, C3_CBAM, EFFN*

1. Introduction

Smoking is not only harmful to health, but also has potential safety hazards such as fire. There are currently over 350 million smokers in my country. This represents 26.6% of the country's total population. Some figures who is smoking can often be found in public places. In this regard, my country has adopted a series of tobacco control measures. The traditional manual supervision method consumes manpower and material resources and cannot achieve real-time detection. Thus, developing an efficient and reliable smoking detection algorithm for public venues has become a major concern.

At present, image recognition technology is becoming more and more mature, smoking detection methods based on computer vision are divided into two branches: traditional target detection methods [1] and deep learning-based target detection methods [2]. Traditional target detection methods are mostly based on smoke features to identify smoking behavior, Shuai L et al. [3] proposed a smoke detection algorithm based on fast self-correcting background difference segmentation and smoke analysis and judgment, which judges the static and dynamic smoke of the candidate area and identifies the smoke area. There are problems such as low cigarette smoke concentration, easy diffusion, and inconspicuous edges, making detection more difficult. Target detection methods based on deep learning are mainly divided into two categories, namely regression-based one-stage [4,5] algorithm and candidate region-based two-stage [6,7] algorithm. Poonam G et al. [8] first applied the Faster RCNN target detection algorithm to the detection of cigarette targets, which effectively improved the detection accuracy, but the detection speed was relatively slow. Z. Rentao et al. [9] proposed a smoking detection algorithm by using the k-means clustering algorithm to select the appropriate anchor frame and adding a small target detection layer to YOLOv3-tiny, which effectively improves the detection accuracy, but is not relatively robust. Since real-time processing of video data is required in the actual scene of smoking detection, the algorithm needs to have a faster detection speed. Therefore, this paper uses the one-stage algorithm YOLOv5 as the target detection model, and improves its network structure to realize the detection of smoking behavior.

To sum up, in view of the situation that the cigarette target scale changes greatly, there are many small and medium-sized targets, the pixels occupied by the cigarette target are low, and the features are not obvious, and this paper proposes the YOLOv5_EC algorithm. This paper mainly does the following two tasks. Firstly, the neck network structure is redesigned, and the EFFN structure is proposed to replace the original Neck of YOLOv5, and the feature maps of three adjacent sizes are fused together, so that it can better retain the positioning information of the target and further enhance the accuracy of small target detection. Then embed the fusion CBAM attention in the YOLOv5 backbone network to better extract important information of the target.

2. Related Work

2.1. YOLOv5 algorithm model

The YOLOv5 algorithm is mainly composed of four parts: Input, Backbone, Neck and Head. As shown in Figure 1.

The Input part is mainly to preprocess the input image data using methods such as Mosaic data enhancement. This method randomly selects four images to expand the scale and richness of the data, reduce the size of some objects, and improve the accuracy of cigarette detection by zooming in, zooming out, changing the color position, cutting and combining, etc.

The Backbone part is mainly responsible for the feature extraction of the input, and generates the corresponding feature map for the subsequent use of Neck. This module is mainly composed of the C3, the SPPF, and the Conv. Among them, YOLOv5 draws on the core idea of the cross-stage local network, designs the C3 module composed of 3 standard convolutional layers and multiple Bottlenecks, and uses it for Backbone and Neck, among them, the C3_1 module of Neck is a C3 module without residual connection, and the C3 module of Backbone mainly learns residual features, and can reduce the amount of network parameters on the basis of unchanged or improved accuracy. SPPF ensures that the final input features are consistent through maximum pooling, and further increases the receptive field of the prediction frame.

The Neck part is mainly responsible for fusing the feature maps of multiple sizes to generate a feature pyramid. YOLOv5 uses the structure of Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) for feature fusion, among them, FPN transmits information of deep-level features from top to bottom, better integrates deep feature information and shallow feature information, and obtains feature maps of different scales for prediction. PAN transfers the shallow spatial feature information from the bottom up, splicing it with the deep feature information, so that the shallow high-resolution feature information is transmitted to the deep layer, so that the deep network can obtain rich location information.

Finally, in the Head layer, predictions are made for the three sizes of small, medium, and large.

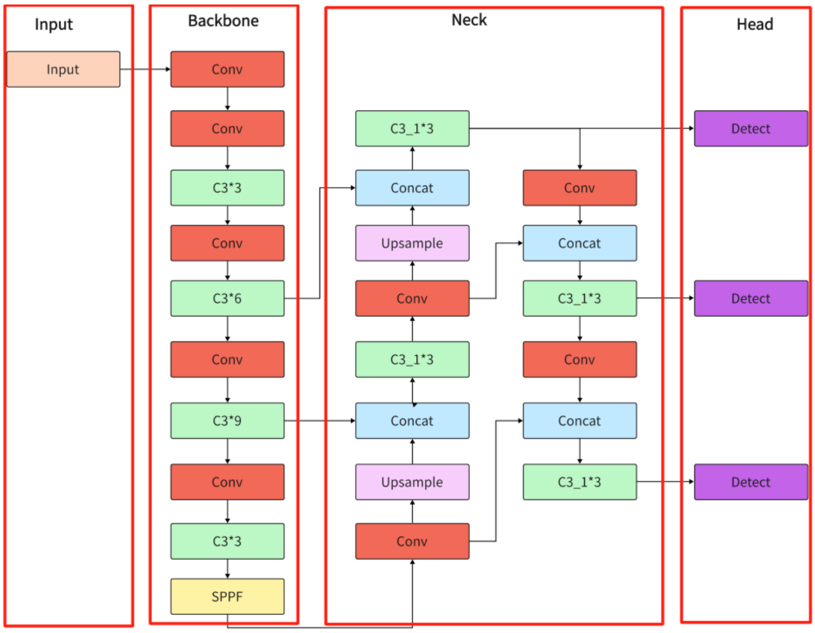

Figure 1: YOLOv5 Algorithm Module.

2.2. CBAM

Convolution Block Attention Module (CBAM) is mainly composed of two core modules, one of which is Spatial Attention Module (SAM) focusing on spatial features, and the other is Channel Attention Module (CAM) focusing on channel features. The structure is shown in Figure 2.

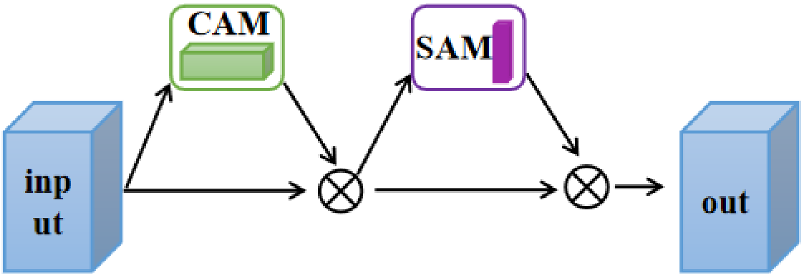

Figure 2: Convolution Block Attention Module (CBAM)

In CBAM attention, the $F \in \mathbb{R}^{C\times H\times W}$ initial input feature map first passes through the CAM module, compressing its spatial dimension, preserving the semantic information of the object as much as possible, and obtaining a 1-D channel attention feature map. An $C \times H \times W$ intermediate feature map is then obtained by multiplying the input feature map with a channel-by-channel weighting. Then its channel dimension is compressed by the SAM module to obtain a 2-D spatial attention feature map that has a focus on location features. Finally, it is weighted with the intermediate feature map to obtain the final feature-enhanced feature map.

In CAM, the Avg Pool and Max Pool operations are first performed on the input feature map, and it is compressed into two feature maps of $C \times 1 \times 1$ size. Then send it to the Share MLP module, in which the number of channels is compressed to the original 1/r, and then it is promoted to the original dimension, and two feature maps are generated after passing the ReLU activation function. Then, the two feature maps are added and the sigmoid activation function is passed to obtain the channel attention feature map. Its structure is shown in Figure 3.

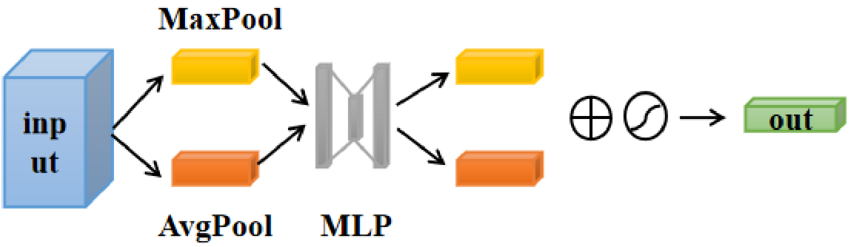

Figure 3: Channel Attention Module (CAM).

In SAM, the intermediate feature map is firstly subjected to Max Pool and Avg Pool operations to compress the channel to 1, and two $1 \times H \times W$ feature maps are obtained. Then perform the Concat operation based on the channel, this is compressed into a 1-channel feature map by Conv and sent to the sigmoid activation function to ultimately generate a spatial attention feature map. Its structure is shown in Figure 4.

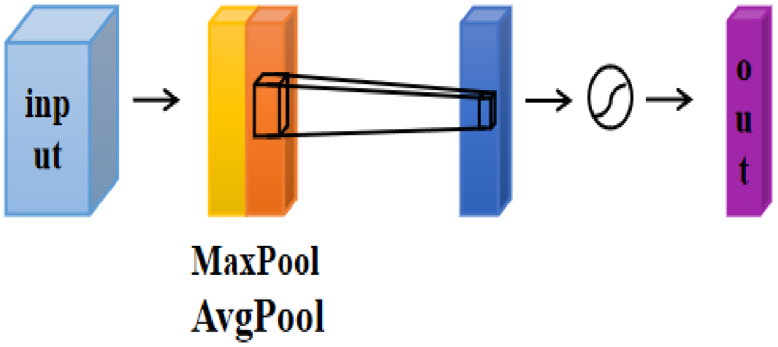

Figure 4: Spatial Attention Module (SAM).

2.3. CARAFE

CARAFE is a lightweight and efficient upsampling operator, which can provide a large receptive field, perceivable content and a relatively lightweight model. Therefore, using CARAFE as an upsampling module in the network can not only reduce the loss of features, but also not bring too much calculation to ensure calculation efficiency.

As shown in Figure 5. CARAFE consists of two main parts: the upsampling convolution kernel prediction module and the feature reconstruction module. Suppose the upsampling ratio is σ, given an input feature map of $H \times W \times C$, we first perform channel compression to C_m to reduce the amount of subsequent calculations. Then use a $K_{up} \times K_{up}$ convolution kernel to predic, expand the re-encoded feature map by channel and are normalized by using Softmax. Finally, the dot product operation is performed between the predicted upsampling kernel of each pixel and the area of $K_{up} \times K_{up}$ in the center of the pixel of the input feature map, so that the same upsampling kernel can participate in the calculation of different channels at the same position, and finally obtain an output feature map of $\sigma H \times \sigma W \times C$.

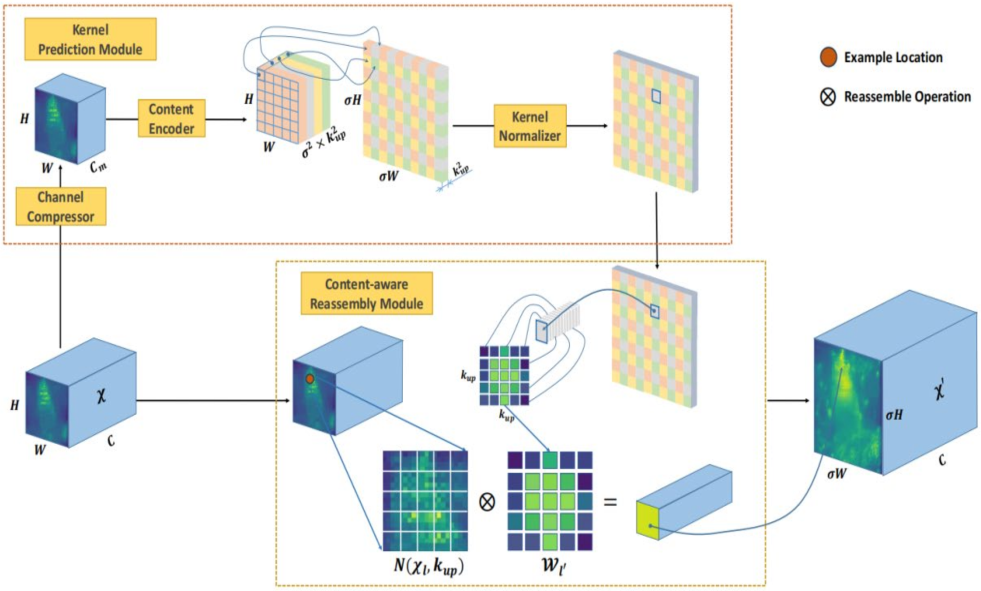

Figure 5: CARAFE Structure.

The formula for the calculation of the operator is the following, where k_{up} is the upsampling convolution kernel of K size, and ⊗ represents the convolution operation:

$$CARAFE = \left[Softmax\left(\sigma H \times \sigma W \times {k_{up}}^2\right)\right] \otimes N\left(X_i, k_{up}\right) \quad (1)$$

3. Method of This Paper

3.1. Main Network

Based on the YOLOv5 algorithm, this paper designs and proposes the YOLOv5_EC algorithm. The structure is shown in Figure 6. The structure mainly including Backbone, Neck and Head. Among them, the CBAM attention is embedded in the Backbone network, allowing the important information in the picture is learned more fully during feature extraction. Then the Enhanced Feature Fusion Neck (EFFN) is redesigned as Neck, the EFFN can better integrate the features of the three size feature maps adjacent to the Backbone, retain more accurate positioning information of the cigarette target, and the effect of small cigarette detection is enhanced, further avoiding the loss of small targets.

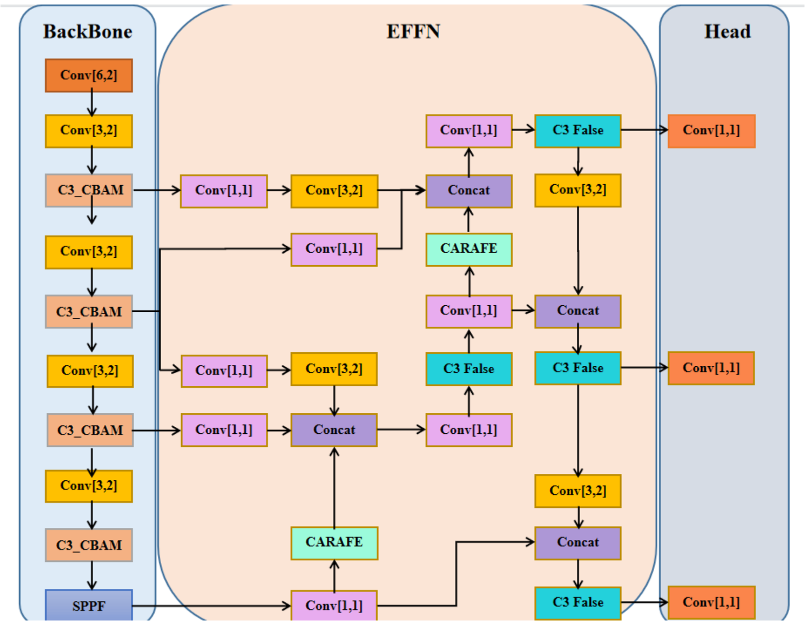

Figure 6: YOLOv5_EC Algorithm Module.

3.2. EFFN

In order to better retain accurate location information, enrich feature semantic information, and enhance the effect and efficiency of small cigarette detection without increasing the amount of calculation, this paper redesigned the neck and proposed an EFFN structure. As shown in Figure 7.

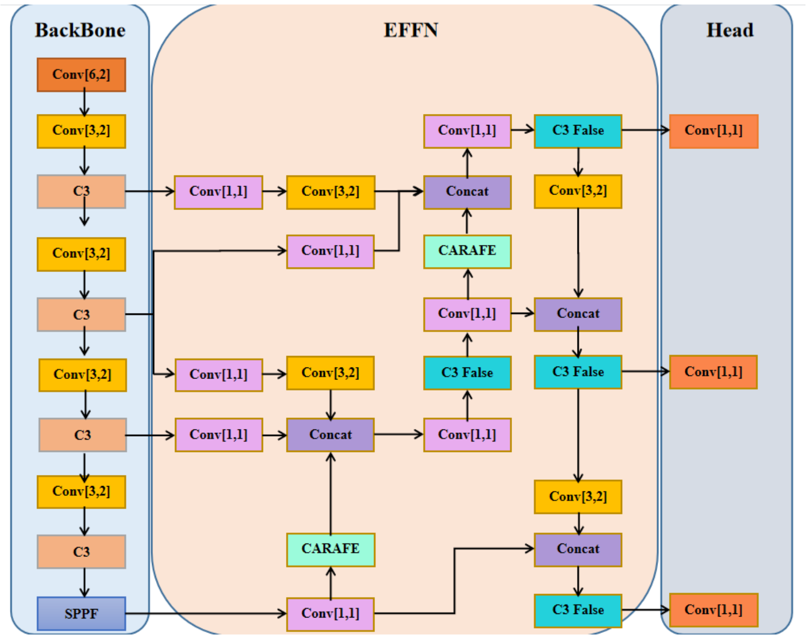

Figure 7: EFFN Network Structure

The core idea of the EFFN structure is to perform two-way feature fusion, that is, to fuse feature maps of the same scale, large scale, and small scale, so as to better retain the location feature information of the target and generate richer semantic features. In order to better integrate features, in the EFFN structure, for feature maps of the same scale, its dimensionality is reduced using the 1×1 Conv operation. For large-scale feature maps, first use the 1×1 Conv to decrease the dimension of the feature map, then use a 3×3 Conv with a step size of 2 to downsample the feature map. For the feature map which size is smaller, use the CARAFE operator to upsample. Finally, the processed three-part feature map is Concat, and then the 1×1 Conv is used to reduce the dimension again to generate a new feature map with more accurate position information and richer semantic information. Among them, the structure uses the

CARAFE operator as the upsampling module, which has a large receptive field and have a good grasp of the content, while ensuring a certain calculation speed.

3.3. C3_CBAM

By assigning different weights, the attention can enable the network to better recognize important features in the image, suppress irrelevant information, and improve detection accuracy. Due to the small size of the cigarette target in the smoking detection, low pixels in the image, few image features, and complex image background, it has caused great interference to the detection of the cigarette target. Therefore, to better improve the feature extraction ability of the model, this paper chooses to integrate CBAM into the C3 module of the backbone network of the YOLOv5, and designs the C3_CBAM, whose structure is shown in Figure 8. First, add CBAM attention after the two convolution operations in the bottleneck module to build the CBAMBottleneck module, to enable the model to better focus on important feature information during feature extraction, then replace the original bottleneck module in the C3 module with CBAMBottleneck to build the C3_CBAM module.

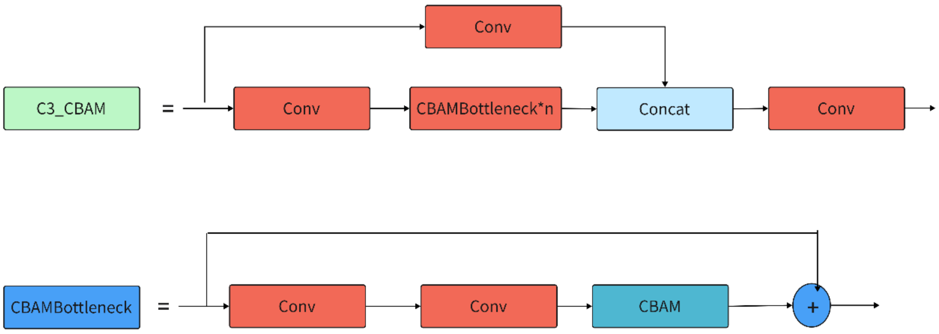

Figure 8: C3_CBAM Module.

This paper designs the positions of two C3_CBAM modules based on YOLOv5, and its structure is shown in Figure 9. The first way is to replace each C3 module of Backbone with a C3_CBAM module to obtain the YOLOv5_C1, It is designed to force the algorithm to focus on important features at each stage of extracting features. The second way is to replace the last two C3 modules of Backbone with C3_CBAM modules to obtain the YOLOv5_C2 model.

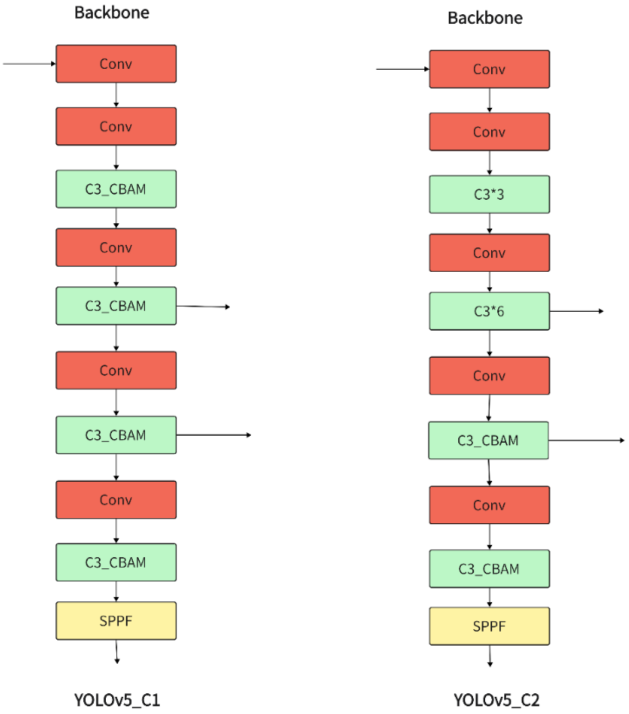

Figure 9: YOLOv5_C1. YOLOv5_C2.

4. Experimental Results and Analysis

4.1. Experimental Environment and Data

The experimental environment of this article runs on a 15-core CPU, 80GB memory, and NVIDIA GeForce RTX 3090 GPU computing platform, the development environment is PyTorch1.11.0, Python 3.8.10, cuda11.7.

Since there is currently no open source smoking detection dataset, this experiment collects smoking pictures on the Internet, intercepts smoking pictures in surveillance videos, etc., and sorts out 4848 pictures containing different scenes, scales and camera angles, as the data set used in this article is labeled with the Label Img tool.

4.2. Evaluation Index

In order to better evaluate the network performance of the algorithm in this paper, this paper uses three indicators of Precision (P), Recall (R) and mAP as evaluation criteria. Among them, the precision rate mainly reflects the proportion of data correctly detected as smoking behavior out of all data predicted to be smoking behavior, that is, a measure of whether there are false detection results. The recall rate mainly reflects the proportion of data correctly identified as smoking behavior out of all smoking behavior data. In other words, it measures whether there are missing detection results. Its calculation formula is as formula (2)～(4):

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+\mathrm{FN}} \quad (3)$$

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \quad (4)$$

Among them, TP refers to the number of smoking behavior is correctly predicted as smoking behavior, TN refers to the number of non-smoking behavior is correctly predicted as non-smoking behavior, FP refers to the number of non-smoking behavior is wrongly predicted as smoking behavior, FN refers to the number of smoking behavior is wrongly predicted as non-smoking behavior, AP is the average precision of object detection. This article uses the mAP value when the IOU threshold is 0.5 for evaluation, that is, mAP@0.5.

4.3. Analysis of Results

4.3.1. Comparative Experimental Results Embedded with CBAM Attention

To further verify the effectiveness of the embedding of the CBAM module in YOLOv5 and to explore the best position for the embedding of the attention module, this paper conducts comparative experiments on the two embedding methods based on YOLOv5.Among them, YOLOv5_C1 embeds each module in Backbone into CBAM, and YOLOv5_C2 embeds CBAM in the third and fourth C3 modules in Backbone. The results of the comparative experiment are shown in Table 1.

Table 1: Comparative experimental results embedded with CBAM Attention.

Model	Precision	Recall	mAP50
YOLOv5	0.804	0.789	0.800
YOLOv5_C1	0.821	0.793	0.812
YOLOv5_C2	0.827	0.786	0.807

By observing the results of three sets of comparison experiments, it can be seen that after the introduction of the C3_CBAM structure proposed in this paper at different positions of the network, the evaluation indicators have been improved to different degrees in comparison with YOLOv5. Amongst these, the introduction of CBAM in each C3 module of the backbone network has a better performance in tobacco detection. Compared with YOLOv5, the performance of the YOLOv5_C1 and YOLOv5_C2 models has improved to a certain extent, which indicates that this paper introduces the CBAM module into the C3 module of the YOLOv5 backbone network, which helps to improve the detection accuracy. From the comparison of the two models, YOLOv5_C1 and YOLOv5_C2, it can be seen that YOLOv5_C1 has a greater improvement in network performance and is more conducive to the detection of small cigarette targets.

4.3.2. Ablation Experiment

To further analyse the reason why the proposed structure improves the performance of the algorithm, and to verify the enhancement effect of each proposed structure on the detection effect of YOLOv5, four groups of ablation comparison experiments were conducted to prove that each enhancement scheme improves the accuracy of the network. The ablation experiments in this paper mainly include the original YOLOv5, embed CBAM attention in the C3 module of the backbone network on the basis of the original YOLOv5, use the proposed EFFN as the neck of YOLOv5, and simultaneously embed CBAM and use EFFN as the neck. The ablation experiment uses precision rate, recall rate, and mAP50 as evaluation indicators. The results of the ablation experiment are shown in Table 2.

Table 2: Ablation experiment results.

Model	YOLOv5	CBAM	EFFN	Precision	Recall	mAP50
YOLOv5	√			0.804	0.789	0.800
YOLOv5_C1	√	√		0.821	0.793	0.812
YOLOv5_EFFN	√		√	0.832	0.782	0.813
YOLOv5_EC	√	√	√	0.830	0.796	0.815

By observing the results of four sets of ablation experiments, it can be seen that after the introduction of various structures proposed in this paper, the evaluation indicators have been improved to varying degrees in comparison with YOLOv5. Of these, YOLOv5 has the best performance for cigarette detection by incorporating CBAM and using the EFFN structure as a neck. The comparison between YOLOv5 and YOLOv5_C1 shows that the performance of the YOLOv5_C1 model is better, indicating that the YOLOv5_C1 model introduces the CBAM module into the backbone network, which helps to improve the detection accuracy. The comparison of YOLOv5 and YOLOv5_EFFN shows that YOLOv5_EFFN is better, indicating that the proposed EFFN structure can more accurately obtain the position information of the cigarette objects in the feature fusion stage. The comparison of the results of YOLOv5_EC and other models shows that YOLOv5_EC introduces the CBAM module and adopts EFFN as the neck structure to improve the accuracy of tobacco detection and also improve the accuracy of cigarette position.

To visually represent the detection effect of the improved model, two images were selected for detection comparison. As shown in Figure 10, it can be clearly seen that the improved YOLOv5_EC algorithm has a good detection result, which further verifies the detection capability of the YOLOv5_EC algorithm.

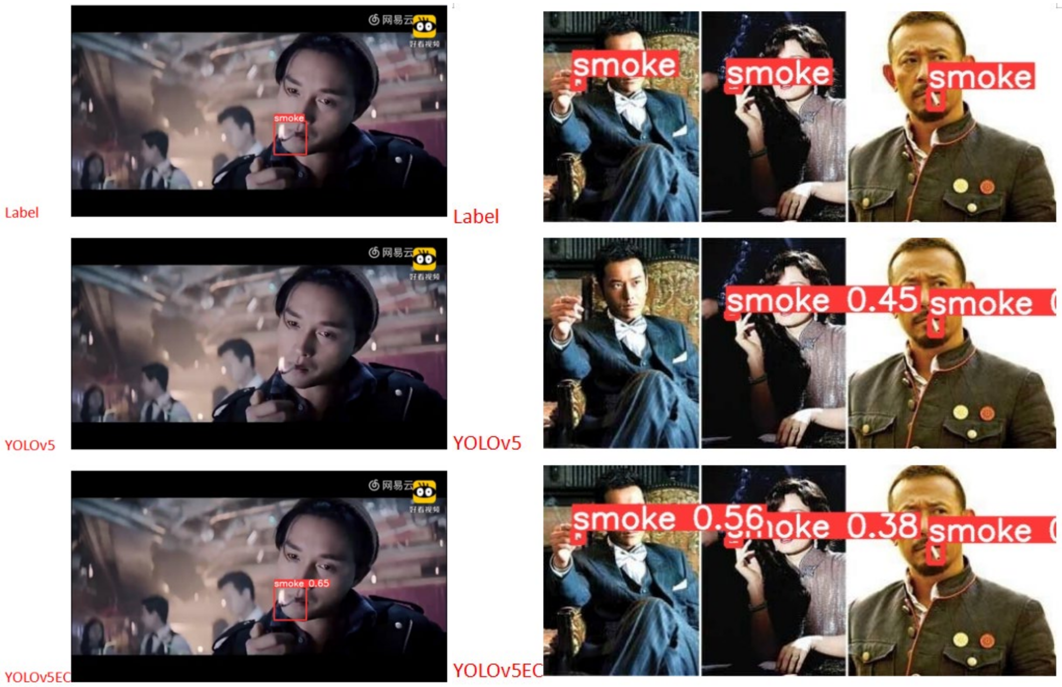

Figure 10: Comparison chart of experimental results.

5. Conclusion

This paper analyzes the network model structure of YOLOv5, and based on the original network, the YOLOv5_EC network model based on CBAM and EFFN is designed and proposed, which improves the accuracy of cigarette target detection, especially the detection accuracy of small cigarette targets. The network model structure embeds CBAM into Backbone's C3 module, which improves the network's attention to important features, reduces the interference of irrelevant information, and improves the performance of feature extraction. In addition, the neck network is redesigned, and the EFFN structure is proposed, which fuses three feature maps of different scales together, retains the positioning information more accurately, while greatly improving the detection accuracy of small cigarette and further avoiding the loss of small targets.

References

[1] H. Tian, W. Li, L. Wang and P. Ogunbona, A Novel Video-Based Smoke Detection Method Using Image Separation[C]. 2012 IEEE International Conference on Multimedia and Expo, Melbourne, VIC, Australia, 2012, 532-537.

[2] Lisu Han, Leilei Rong, Yongquan Li, Zhikang Qin, and Yan Xu. CA-SSD-Based Real-time Smoking Target Detection Algorithm[C]. 2021 5th International Conference on Digital Signal Processing. Association for Computing Machinery, New York, NY, USA, 2021, 283–288.

[3] Shuai L, Bo W, Ranran D, et al. A novel smoke detection algorithm based on Fast Self-tuning background subtraction[C] IEEE 2016 Chinese Control and Decision Conference (CCDC) - Yinchuan, China, 2016, 3539-3543.

[4] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection[C] 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). LasVegas: IEEE, 2016, 779-788.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD:Single Shot MultiBox Detector[C].Computer Vision–ECCV 2016. Amsterdam:Springer, Cham, 2016,21-37.

[6] GIRSHICK R, DONAHUE J, DARRELL T, et al.Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C].2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus:IEEE,2014:580-587.

[7] RENS Q, HEK 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.

[8] Poonam G, Shashank B N, Rao A G. Development of framework for detecting smoking scene in video clips[J]. Indonesian Journal of Electrical Engineering and Computer Science, 2019, 13(1):22-26.

[9] Z. Rentao, W. Mengyi, Z. Zilong, L. Ping and Z. Qingyu, Indoor Smoking Behavior Detection Based on YOLOv3-tiny[C]. 2019 Chinese Automation Congress (CAC), Hangzhou, China, 2019, 3477-3481.