Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/41/2024CH0126

Research on Improved Crowd Detection Based on YOLOv5

Qi Wen[1], Kecheng Li[2,4], Yue Wang[3]

[1]School of Computer Science, University of Xi'an for Polytechnic, Xian, China
[2]College of Computer and Cyber Security, Chengdu University of Technology, Chengdu, China
[3]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China

[4]li.kecheng@student.zy.cdut.edu.cn

Abstract. With the acceleration of the process of modern urbanization and the improvement of residents' material living standards, the flow of people in the public space is gradually becoming saturated. The monitoring equipment in public places records a huge amount of people flow information all the time, but due to the crowds tend to be dense and crowded. Traditional machine learning cannot make accurate and efficient identification of a large number of dense crowds, if the deep learning technology can be used to process the crowded crowd captured by the surveillance camera and accurately identify the number of people in public places, it provides an important guarantee for the flow of people in public areas and safety construction. However, for crowded targets with occlusions, the traditional target detection algorithm sometimes performs poorly. Based on the above background, this paper introduces an enhanced deep learning framework utilizing the YOLOv5 neural network for crowd detection research. aiming at the characteristics of dense and crowded crowds in public areas. By improving convolutional layer C3 in the backbone structure of YOLOv5 neural network and adding CBAM attention mechanism. Compared with the original YOLOv5, the improved model has increased the maximum F1 value of crowd recognition at near, middle and far distances. To sum up, the deep learning framework improved by YOLOv5 neural network proposed in this paper has significantly improved the recognition accuracy of crowded people in public areas.

Keywords: YOLOv5, Crowds, Image recognition, CBAM attention mechanism.

1. Introduction

With computer vision technology, big data tracking and the development of convolutional neural networks, object detection plays a vital role in various fields. In practice, this method can monitor the flow of people in public places and tourist attractions. At present, YOLOv5, faster, R-CNN and other deep learning-based object detection methods have been applied to human flow detection. YOLO (You Only Look Once), as a classic real-time target detection algorithm, its fast speed and good effect make it widely used. It plays a crucial role in detecting pedestrian flow in public areas [1]. But when it comes to crowd counting, now it's difficult to count pedestrians based on dense and crowded crowd identification, suitable for large density, small target population, there's still a lot of room to explore, and current crowd-based target detection, deep learning algorithms are still in their infancy, difficult to apply in real life scenarios, it has such problems as poor robustness, low precision and large calculation amount [2]. To solve the above problems, there have been many studies that have made remarkable progress. For example, by adding the Focus layer [3], the amount of model computation can be reduced and the operation speed can be accelerated, thus improving the detection efficiency. In addition, by introducing attention mechanism and adding SE module in the stage of network feature fusion, the localization accuracy of information is improved. At the same time, using Soft-NMS to replace the original NMS, the mean detection accuracy mAP@0.5 increased by 1.5%, and the recall rate increased by 0.5% [4]. This paper designs an improved deep learning framework based on YOLOv5 neural network. By improving the convolutional layer C3 in the backbone structure of YOLOv5 neural network, add CBAM attention mechanism [5], to realize the improvement of crowd identification accuracy at near, middle and far distances, allowing it to more accurately identify obscured targets, thus improve the detection performance and practicability. The primary contents of this paper include: Identify targets for crowds, this paper completes the crowd count by detecting the torso of the person. First collect and create a data set for people detection and identification, the dataset consists of 15,000 images. Then they trained on the data they had collected, according to the training data, the characteristics of occlusion and congestion of the target are identified. Improvements to YOLOv5, by adding SENet attention mechanism to YOLOv5, that is, each Channel is pooled, through two fully connected layers, get the output vector, then, the nodes and channels of the second fully connected layer are aligned [6]. The final output, the F1 value after training increased by 0.03 compared with the original YOLOv5. The recall rate went up from 0.71 to 0.73, solve the problem of target recognition accuracy of crowded crowd at middle distance and far distance. However, the experiment found that its accuracy in short-range target recognition decreased compared with the original YOLOv5.To solve this problem, this paper introduces a new attention mechanism CBAM to improve YOLOv5. By adding two new modules before the data entry of the original C3 module of YOLOv5, channel attention mechanism and spatial attention mechanism, the two modules are multiplied after each calculation is completed, by suppressing information that is not important in terms of channel and space, respectively, the F1 value after training is 0.72, the confidence for accuracy of 1 is 0.968, when the confidence is 0.5, the accuracy is 0.768, The recall rate was 0.76, the F1 value of the original YOLOv5 is 0.68. The confidence for accuracy of 1 is 0.985, when the confidence is 0.5, the accuracy is 0.723. The recalls rate was 0.71. Experiments show that the target recognition accuracy of middle and far crowded crowd is improved. At the same time, it also maintains the recognition accuracy of the original YOLOv5 in the close-range target. Solved the model in the complex public scene, especially for a large number of people with small targets, the identification accuracy of the problem.

2. Research methods

2.1. Introduction to YOLOv5 architecture

YOLOv5 is an efficient target detection model, which is characterized by a simple model architecture and high efficiency of target recognition, especially suitable for real-time multi-target recognition scenarios. The YOLOv5s version is used in this paper, this version also boasts the smallest depth and feature map width among the YOLOv5 series. The basic components of YOLOv5s include Focus, Conv, C3, SPP. The function of Focus is to decompose the high-resolution feature map into many low-resolution feature maps, that is, by reducing the larger input image to a smaller input image to improve the speed of calculation and the accuracy of feature extraction. Conv is a conventional convolution layer in YOLOv5. The main goal is to convolve the input image through the convolution kernel operation to achieve the purpose of feature extraction and processing. C3 is the key complex convolutional layer module of YOLOv5s. The main idea revolves around dividing the input feature map into two parts and processing them separately, and finally merge them to reduce the amount of calculation as much as possible. SPP is a pooling module with a pyramid shape. It uses a maximum pooling method to extract features from different spatial scales and perform multi-scale fusion, so that the model has excellent recognition ability for various targets of different sizes. In general, the main idea of YOLOv5s is to extract features through complex convolutional layers and strengthen feature fusion through multi-scale pooling layers, to achieve the effect of fast and accurate recognition of different targets.

2.2. Introduction of complex convolution layer C3

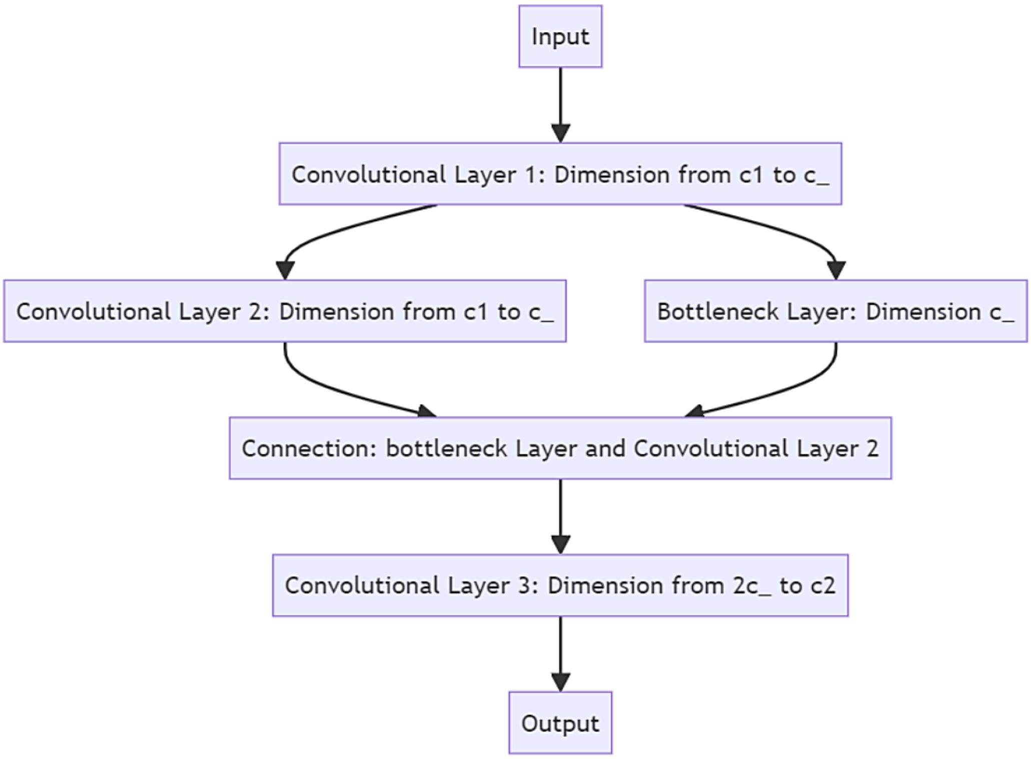
Input
Convolutional Layer 1: Dimension from c1 to c_
Convolutional Layer 2: Dimension from c1 to c_
Bottleneck Layer: Dimension c_
Connection: bottleneck Layer and Convolutional Layer 2
Convolutional Layer 3: Dimension from 2c_ to c2
Output

Figure 1. C3 fundamentals.

Figure 1 illustrates the fundamental principle of the C3 layer, with c1 denoting the input channels, c2 denoting the output channels, and c_ representing the channels generated during the intermediate convolution process. In general, there are four convolutional layers: convolutional layer 1 and convolutional layer 2 are exactly the same, and their role is to adjust the number of channels of the feature map for subsequent processing; The bottleneck layer is also a convolutional layer, similar to convolutional layer 1 and convolutional layer 2, which mainly provides a bypass for the convolution operation and directly connects to the subsequent connection layers; After the connection in the channel dimension, the convolution layer C3 can perform the convolution operation on the feature map as a whole, and the number of channels is converted from 2c_ to c2, so as to generate the final feature map. In general, the main feature of C3 layer is that it can fuse features of various scales more quickly through the combination of parallel convolution and concatenation operations to enhance the model's feature extraction capabilities.

2.3. Improvement scheme

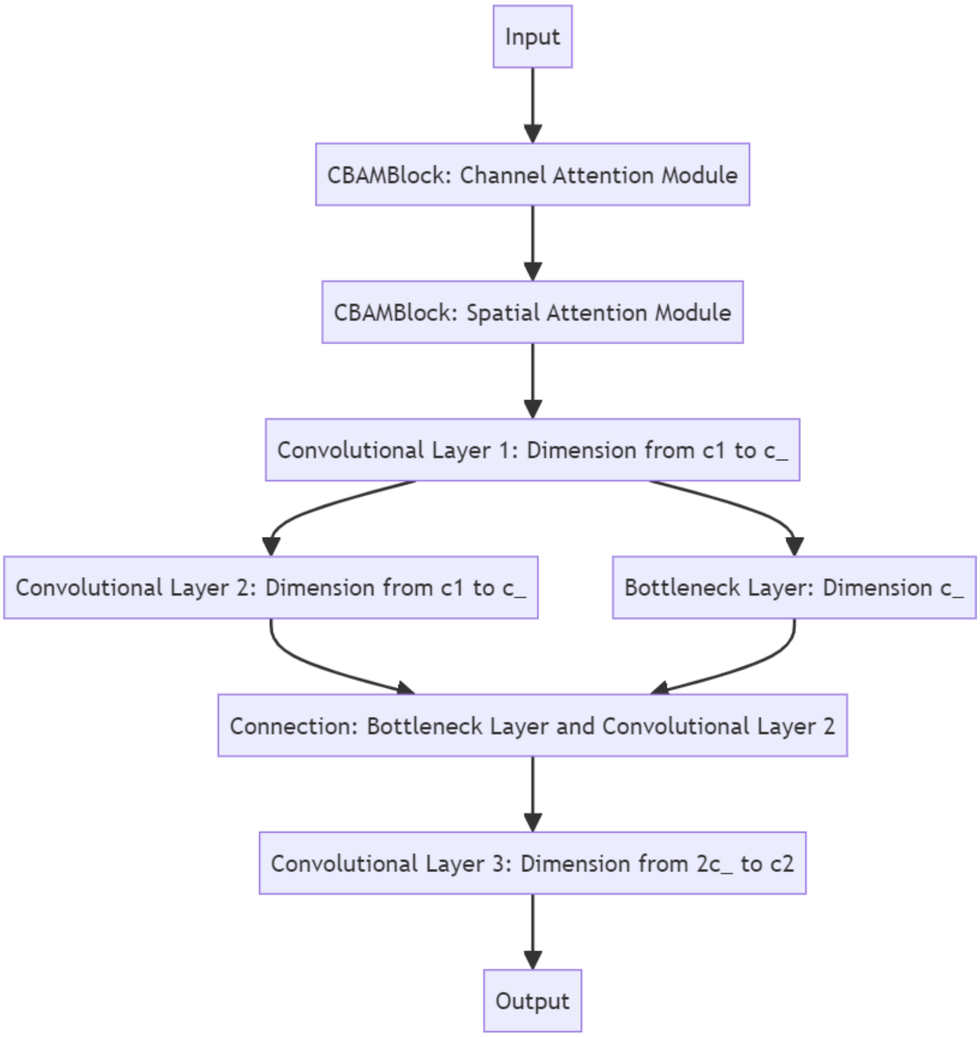
Input
CBAMBlock: Channel Attention Module
CBAMBlock: Spatial Attention Module
Convolutional Layer 1: Dimension from c1 to c_
Convolutional Layer 2: Dimension from c1 to c_
Bottleneck Layer: Dimension c_
Connection: Bottleneck Layer and Convolutional Layer 2
Convolutional Layer 3: Dimension from 2c_ to c2
Output

Figure 2. C3 fundamentals.

Based on the architecture of YOLOv5s, this paper makes code improvements. The main improvement scheme is to replace the original complex convolutional layer C3 of YOLOv5s with the C3CBAM module to increase its recognition ability for dense crowds. C3CBAM module actually adds two new modules: adding spatial attention and channel attention modules before the input data of the original C3 module. As shown in Figure 2, the two modules multiply after the calculation is completed respectively, so as to suppress the unimportant information in terms of channel and space respectively. The channel attention module can dynamically learn the significance of individual channels over time, reduce the effect of irrelevant feature channels to reduce redundant information, and enhance the model's robustness. The spatial attention module helps the model to understand the relationship between pixels more accurately through continuous learning of spatial position weights, so as to enhance the function of feature extraction [7]. By adding the C3CBAM module to the C3 layer of YOLOv5s, the scheme presented in this paper not only boosts the model's recognition accuracy for individuals but also enhances its effectiveness in complex scenes, particularly for numerous individuals with small targets.

2.4. Introduction of channel attention module

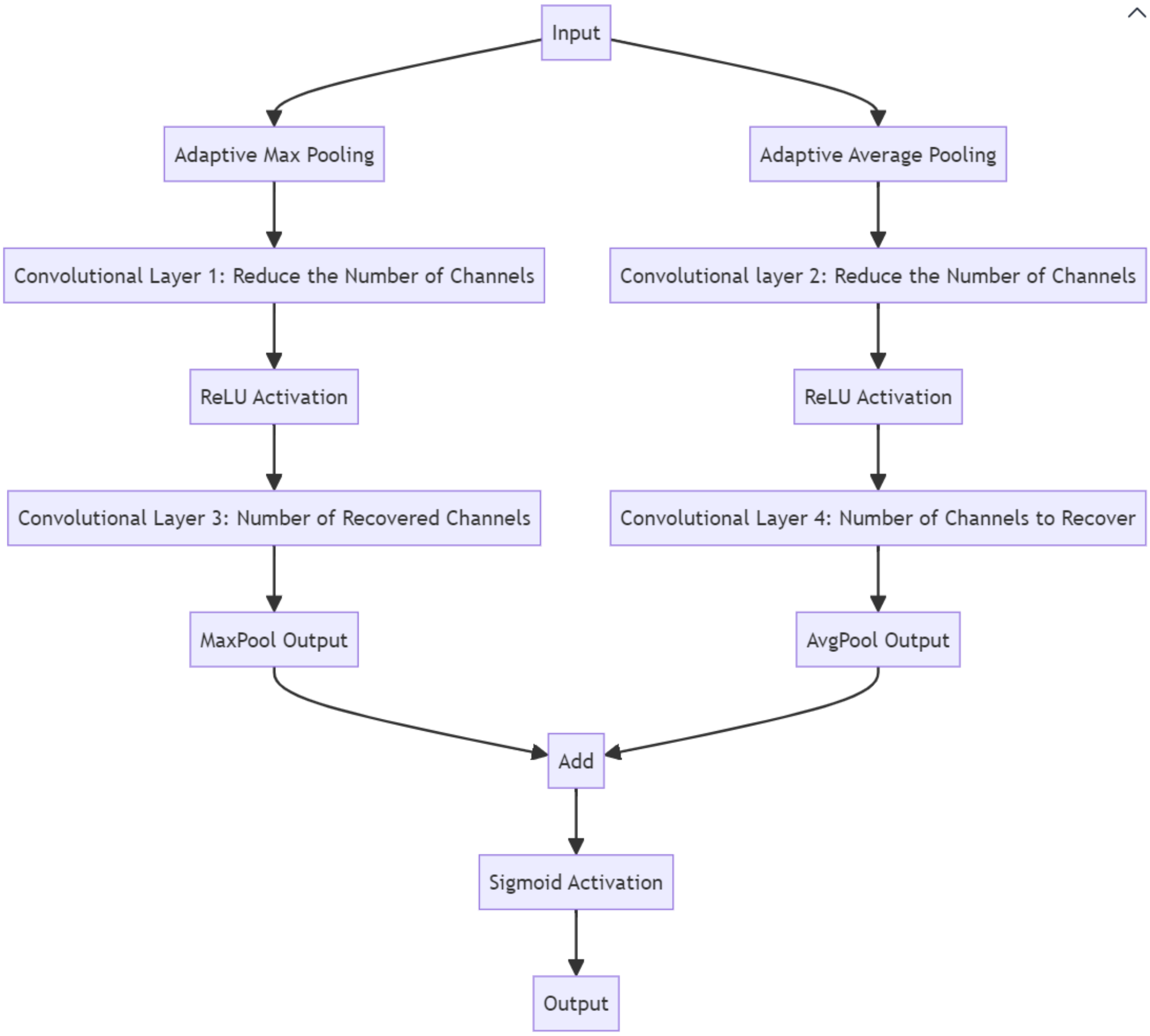
Input
Adaptive Max Pooling
Adaptive Average Pooling
Convolutional Layer 1: Reduce the Number of Channels
Convolutional layer 2: Reduce the Number of Channels
ReLU Activation
ReLU Activation
Convolutional Layer 3: Number of Recovered Channels
Convolutional Layer 4: Number of Channels to Recover
MaxPool Output
AvgPool Output
Add
Sigmoid Activation
Output

Figure 3. Channel attention module fundamentals.

As shown in Figure 3, this is the basic principle of the channel attention module. The input here refers to the feature map that was originally fed to the C3 layer. After input, the feature map is divided into two parts for adaptive Max pooling and adaptive average pooling respectively. The role of this step is to make the feature map retain the information of the global maximum value and the global average value respectively. Then the two parts enter the convolution layer and the activation layer respectively, which reduces the number of channels and extracts key information while introducing nonlinearity to increase model's expression ability. Convolution is then performed to recover the original number of channels. Finally, the results of the processed maximum pooling and the results of the average pooling were added to fuse the features of the two pooling strategies, and then the Sigmoid activation function was used to compress the output range to between [0, 1] to generate the channel attention weight. In this way, the importance of each channel can be obtained when performing the output. In general, the channel attention module enhances the expression ability of important channels by concatenating the convolutions of maximum pooling and average pooling results [8].

2.5. Introduction to Spatial Attention Module

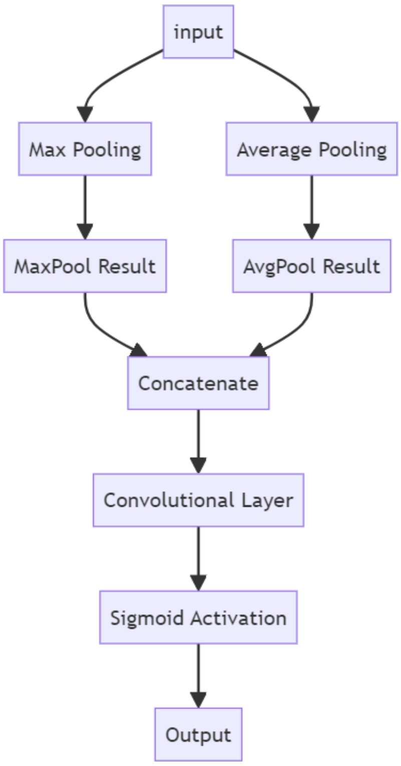

Figure 4. Spatial attention module fundamentals.

As depicted in Figure 4, this illustrates the fundamental principle of the spatial attention module. Once again, the output refers to the feature map that was originally fed to the C3 layer, not the output of the channel attention module. First, the research needs to perform Max pooling and average pooling operations on the input feature maps, which can extract the global maximum and average value in the spatial dimension. Then, the results obtained by Max pooling and average pooling are concatenated to obtain a feature map with increased dimensions and containing the results of both types of pooling. Finally, convolution and Sigmoid activation are also performed to enhance the key spatial location features and compress the output range. In general, the spatial attention module performs convolution and activation on the concatenation of the maximum pooling and average pooling results to achieve the effect of extracting important information from the spatial dimension[9]. In comparison, the main difference between the spatial attention module and the channel attention module is the difference in the pooling operation, and the difference in the method of combining the two parts of the pooling results during feature fusion.

2.6. Introduction to the dataset

The experiment uses the CrowdHuman dataset, which contains a total of 15,000 dense crowd images, in which a total of 339,565 objects have been marked for recognition. Due to the small size and fuzzy contour of a considerable part of targets in the dataset, this paper believes that this dataset is very suitable for the training and verification of dense crowd recognition[10]. Since the label specification of the dataset itself does not conform to the training scheme of YOLOv5, and the image sizes in the dataset are different, the format of the label should be adjusted before training, so that the label meets the label format of YOLOv5 and can adapt to different image sizes. In this paper, only the first part of the three parts of the data set of the training set is used in the training, and some images are too simple to recognize, such as images of several people standing in a row to take group photos, etc., so this paper removes these images in the verification.

3. Experimental results

3.1. Introduction to SENet

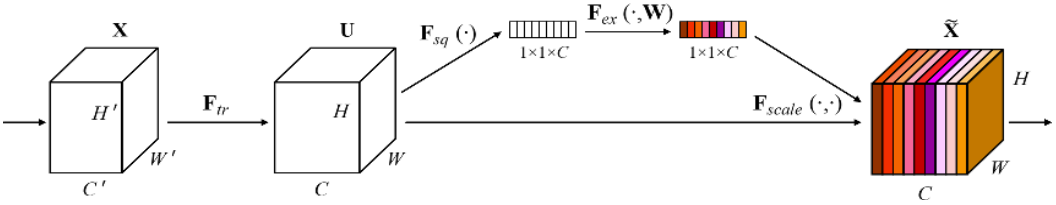

Figure 5. SENet Fundamentals

The effect of the SENet module is similar to this paper, which is a module that enhances the feature representation power to make the neural network perform better. As shown in Figure 5, the SENet module first performs the convolution operation on the input feature map to change the original size C '×H' ×W 'feature map into C×H×W. Next, Squeeze operation, namely Fsq, is applied to the obtained new feature map, and a 1×1×C vector is obtained by global average pooling. This vector is then fed into the fully connected layer Fex learns the importance of each channel. Finally, through Fscale. The product of the weight vector and the original feature map is the final output feature map [11].

3.2. Comparison of target detection effects

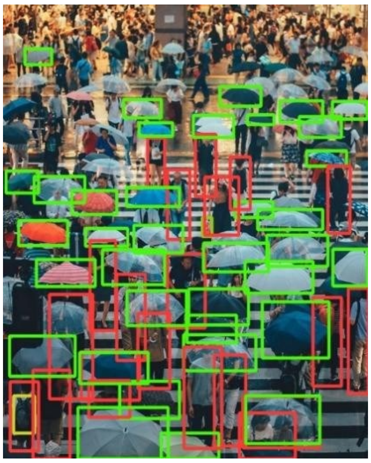

Figure 6. YOLOv5.

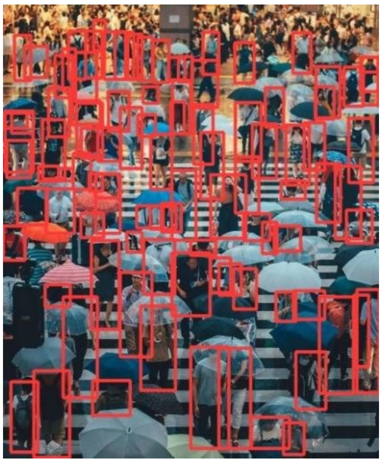

Figure 7. YOLOv5(aftertraining).

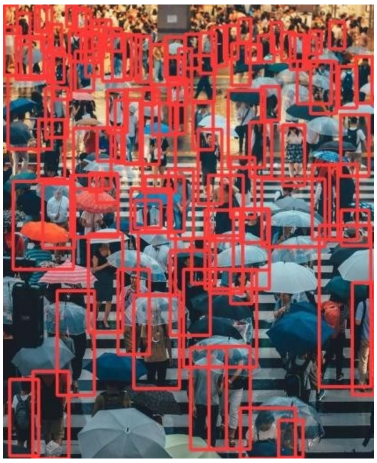

Figure 8. SENet.

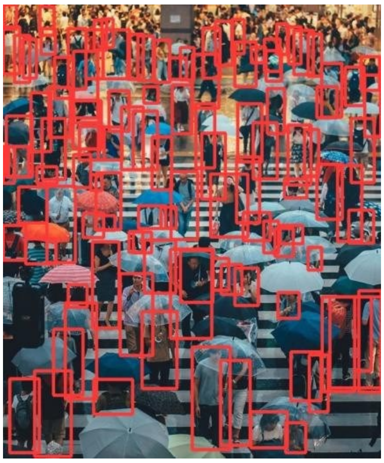

Figure 9. CBAM Improvement.

As shown in the figure 6, it can be seen that the untrained YOLOv5 identifies a lot of non-human parts and many people fail to recognize them, mostly because there are many and sparse detection targets in the images of the training set. YOLOv5 has a good recognition effect after training with the dense crowd training set CrowdHuman, as figure 7 shows, it lacks the ability to recognize long-distance targets. Figure 8 illustrates SENet has a strong recognition ability for distant targets, but its recognition ability for medium and near ranges is not as good as the original YOLOv5. The improved model after CBAM shown as figure 9, can not only accurately identify the medium and close targets, but also improve the ability of distant target recognition, which really improves the recognition accuracy.

3.3. Comparison of experimental data

Table 1. Validation data comparison.

Index	YOLOv5 CBAM	YOLOv5
F1 maximum	0.72 (when the confidence is 0.436)	0.69 (when the confidence is 0.427)
Confidence (when the precision is 1)	0.768	0.985
Accuracy at a confidence level of 0.5	0.109	0.723
Recall	0.76	0.71

As shown in Table 1, the F1 maximum value of YOLOv5 improved by CBAM is slightly improved, and its recognition accuracy is improved from 0.723 to 0.768 when the confidence level is 0.5. The most significant improvement was in recall, which went from 0.71 to 0.76. This shows that the improved model has a great improvement in the recognition of positive samples compared with the original model.

4. Conclusion

By improving the forward propagation function of YOLOv5 model and the complex convolutional layer C3 in the backbone network, this paper significantly improves the recognition accuracy of target detection. The model can learn features more effectively during training. By adjusting the structure and optimizing the parameters of the C3 convolution layer, the robustness and accuracy of the model in dealing with complex scenes are enhanced. The results show that after improved YOLOv5 model, it has achieved significant performance improvement on multiple public data sets, which proves the effectiveness of the proposed method.

However, this study has some limitations. Firstly, although the accuracy of the enhanced model has increased, its computational complexity and inference time have also increased, which may bring certain challenges in practical applications. Moreover, the improvements in this paper are mainly aimed at specific dense crowd detection tasks, and for other types of visual tasks (such as image segmentation or pose estimation), the effect is uncertain and needs further verification. Future research can be further explored in the following aspects: First, to further optimize the computational efficiency of the model and reduce resource consumption; Second, the proposed method is extended to other types of deep learning models and tasks to verify its universality. The third is to combine other advanced technical means, for example, incorporating attention mechanisms and multi-scale feature fusion, to further improve the performance and adaptability of the model. By improving the forward propagation function of YOLOv5 model and C3 convolution layer, this project successfully improved the accuracy of target detection and provided new ideas and methods for subsequent research. Expect to see more relevant innovations and breakthroughs in practical applications and other fields in the future.

Authors contribution

All authors contributed equally, regardless of the order of authorship.

References

[1] Jiang, X. K., Liao X. L. and Li Y. B. (2019). Regional crowd flow statistics based on Deep learning. Digital Users, 29(21), 165-167

[2] Zhan, W. W. (2022). Design and implementation of crowd counting and anomaly detection system. Beijing: Beijing University of Technology.

[3] Chen, B., Dai, S, L. and Ye, B, Y. (2023). Yolo-based social distancing detection method for people in public areas. Artificial intelligence and robotics research, 12(3), 10.12677/AIRR. 2023.123023

[4] Cong, X, H., Li, S, X., Chen, F, K. and Meng, Y. (2023). An improved dense pedestrian detection algorithm based on YOLOv5. Computer science and applications, 13(6), 1199-1207

[5] Wang, X., Dong, Q., Yang, G. Y. (2023). Crops diseases and insect pests recognition based on optimized CBAM improvement YOLOv5. Computer system application, 32 (7), 261-268. 10. 15888 / j.carol carroll nki. Csa. 009175.

[6] Li, X. P., Zhang, Y. B., Li Y. P., et al. (2023). An improved algorithm for infrared image target detection based on YOLOv5s. Laser & Infrared, 53(7), 1043-1051. 10.3969/j.issn.1001-5078. 2023.07.010.

[7] Pei, Y. H., Xu, L. M., & Zheng, B. C. (2022). Improved YOLOv5 for Dense Wildlife Object Detection. *BiometricRecognition:16thChineseConference*, 569-578.

[8] Ji, D. J., and Cho, D. H. (2021). ChannelAttention: Utilizing Attention Layers for Accurate Massive MIMO Channel Feedback. IEEE Wireless Communications Letters, 10(5), 1079-1082. https://doi.org/10.1109/LWC.2021.3057934

[9] You, C. (2021). Research on Smoke and Flame Image Classification Algorithm Based on BAN. Zhejiang Sci-Tech University.

[10] Xu, H. H., Wang, X. Q., Wang, D., et al. (2023). Object detection in crowded scenes via joint prediction. *Defense Technology*, 21(3), 103-115.https://doi.org/10.3969/j.issn.2214-9147. 2023.03.008

[11] Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2019). Squeeze-and-Excitation Networks. Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.1709.01507