Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 3, 2023

Research on Expressway Pavement Crack Detection based on Improved YOLOv5s

Chunlin He [1], Jiaye Wu [2, *], Yujie Yang [1]

[1] Sichuan University of Science & Engineering, Zigong, Sichuan, 643000, China
[2] Sichuan Central Inspection Technology Inc., Chengdu, Sichuan, 610045, China

*** Corresponding author:** Jiaye Wu

Abstract: In order to address the issues of missed detection, false detection, and low accuracy of current road cracks, we propose a road crack recognition model based on improved YOLOv5. Firstly, add a CBAM attention module to the backbone network to enhance feature extraction capabilities; Then, a weighted bidirectional feature pyramid (BiFPN) is incorporated into the model for multi-scale feature fusion, replacing the traditional feature pyramid (FPN)+pixel aggregation network (PAN) structure to enhance feature fusion. The experimental results indicate that the improved model outperforms the traditional YOLOV5 model in terms of mAP@0.5 By 17.3%, the improved YOLOv5 algorithm performs well in detecting road cracks and can quickly and accurately identify and locate cracks on the road.

Keywords: Road Crack Detection; YOLOv5s Algorithm; CBAM Attention Mechanism; Feature Fusion.

1. Introduction

With the development of highways in China and the increase in car ownership, the load on highways has also increased, and the problem of road diseases has become increasingly prominent. Road cracks are one of the most common diseases [1], threatening driving safety and shortening the service life of roads. At present, there are two main methods for detecting road cracks: digital image processing and deep learning.

Traditional digital image processing methods extract and classify the color, shape, edge and other features of cracks in images. Oliveira et al. [2] used the difference in grayscale values between cracks and the background for detection, but the results were not satisfactory in actual road conditions. Wang Xing et al. [3] used wavelet transform to identify complex road cracks, but there is a large amount of noise in data collection during high-speed driving, which makes it difficult to detect discontinuous cracks effectively.

With the continuous development of deep learning technology, detection methods based on neural networks are gradually being applied to the industrial field[4]. You Jiangchuan et al. [5] proposed an improved RCNN asphalt pavement crack detection method with an accuracy of 91.25%. However, the detection speed is slow and cannot be applied to fast form scenarios; Xu Kang et al. [6] proposed an improved Faster RCNN method for asphalt pavement crack detection, with an accuracy of 85.64%. However, it requires a large number of parameters and floating-point operations and is not suitable for resource constrained mobile deployment platforms; Gu Shuhao et al. [7] proposed a crack automatic detection algorithm that enhances the fusion of semantic information and multi-channel features. By adding an expansion convolution module and attention mechanism, the model's ability to extract feature details is improved; Yang et al. [8]proposed the FPHBN network, which uses a feature pyramid structure to fuse deep and shallow feature information, and balances different samples through layered weighting to effectively improve the accuracy of crack detection. The YOLO series of object detection algorithms have been continuously updated, and their detection speed and model size are more advantageous compared to other detection models. J. Zhang et al. replaced the feature extraction network in YOLOv4 with some lightweight networks to improve detection speed, reducing the number of parameters and backbone network layers, resulting in better model accuracy and speed, and a lighter model [9]. In order to improve the recognition accuracy of complex road surfaces, some scholars have made variant improvements to the YOLO series of algorithms. Peng Yunuo et al. proposed YOLO-lump and YOLO-crack for extracting multi-scale features of sparse expressions, which can reduce information loss and achieve the goal of improving detection accuracy and response speed.

This article uses the YOLOv5s model to detect road cracks and makes structural improvements to improve the extraction and utilization of deep feature information, ensuring real-time detection while improving detection accuracy.

2. The Principle of YOLOv5s

2.1. Target Detection

Object detection is an important research direction in the field of computer vision. Object detection mainly completes three tasks: detecting the position of targets in an image and the possibility of multiple detection targets in the same image; Detect the size range of the target; Identify and classify the detected targets. The category and position coordinates of the objects in the final output image. The two-stage algorithm mainly determines target candidate regions through multiple detection boxes of different scales, and then performs classification and regression operations. The YOLO series is a one-stage object detection algorithm that can obtain the position and classification information of multiple regression boxes at once based on multiple initial error boxes, achieving network training for single stage object detection. Compared to the two-stage algorithm, the YOLO series algorithm has a faster detection speed. Representative one-stage algorithms include SSD, ATSS, and RepPoints based on error points.

2.2. The Structure of YOLOv5s

YOLOv5s consists of four parts: input, backbone, Neck network, and Head output, as shown in Figure 1. At the input end, the methods of Mosaic transformation, adaptive anchor box, and image scaling enrich the training data, improve the inference speed of the algorithm, and enhance the robustness of the model. Backbone extracts refined feature information through an efficient combination of CBS, CSP, and SPPF modules. To further enhance contextual information, the Neck part composed of Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) achieves multi-scale information fusion. In the prediction stage, the position of the prediction box is calibrated through CIOU Loss, and then the optimal prediction box is obtained through Non Maximum Suppression (NMS).

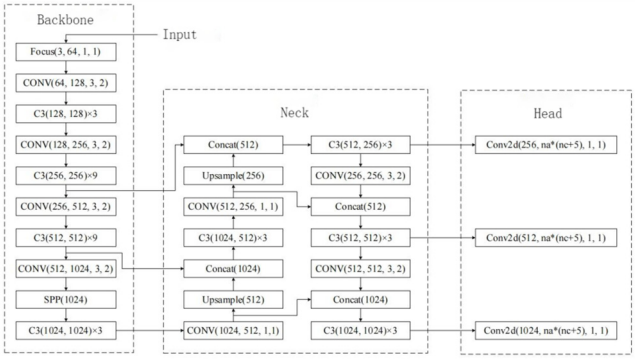


Figure 1. Schematic diagram of the YOLOv5 structure

3. Improvement of YOLOv5s algorithm

3.1. CBAM Attention Module

This article introduces the CBAM (Convolutional Block Attention Module) attention module, which is a lightweight convolutional attention module suitable for mobile deployment. The interior includes channels and spatial attention modules, which can locate, recognize targets from both dimensions of spatial channels, and refine the extracted features, avoiding the problem of redundant information flooding targets caused by convolution. In addition, to avoid premature addition of attention mechanisms that may lead to deviation in network focus, this module is added to the last layer of the backbone network. CBAM is shown in Figure 2.

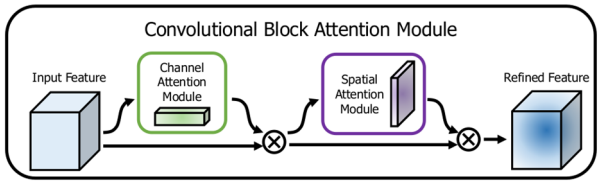


Figure 2. Schematic diagram of the CBAM module

The "plug and play" structure of the CBAM module provides great convenience for the use of the module, without clear regulations on the insertion position. In principle, the CBAM module can be inserted into any position in the network. In common application examples, attention modules are often combined with feature fusion operations, but further experiments are needed to determine which position or positions in the network can achieve better results. If all feature fusion operations in the network are integrated with the CBAM module, the computational complexity of the network will greatly increase. Perhaps the network may also have overfitting, which can actually reduce the performance of the network model and run counter to the purpose of embedding attention modules in the network. Therefore, CBAM modules should be embedded in critical positions in the network. The following is a detailed introduction to the experimental design of the insertion position of the CBAM module.

3.1.1. Insertion Position of CBAM Module

The input end performs preprocessing such as Mosaic data augmentation and adaptive scaling on the image of the input network, with the aim of enriching the dataset, enabling the model to have better learning ability on the dataset, and improving the generalization ability of the network model. After processing, the input image is still image data and no operations such as convolution or pooling have been performed to extract image features. Therefore, there is no need to introduce an attention module for the input end. Therefore, the CBAM model can be embedded into the other three major network modules to study the impact of the integration of CBAM modules with the three parts on the overall network performance. The network models generated by the fusion of CBAM model and YOLOv5s' Backbone, Neck, and Head modules are denoted as YOLOv5s_B, YOLOv5s_N and YOLOV5s_H.

(1) YOLOv5s_B

YOLOv5s_B is a new network model obtained by integrating the CBAM module with the Backbone master network. Backbone, as the backbone network of YOLO5, is used to extract the high, middle, and low level features of an image through deep convolution operations, and then aggregate them to form the overall features of the image. This section includes four structures: Focus slicing operation, CONV convolution layer, C3 cross stage local network structure, and SPP spatial pyramid pooling structure.

The Focus layer was first proposed in YOLOv5 and is a slicing operation for feature maps. The structural diagram is shown in Figure 3.This operation utilizes the conversion of flat data information on width and height to the channel level, and then uses convolutional methods to extract different features to achieve the purpose of down-sampling. The down-sampling operation in neural networks is generally used to reduce the number of parameters and achieve dimensionality reduction, while also increasing the range of local receptive fields. However, compared to using a convolution or pooling layer with a step size of 2, the Focus layer can effectively reduce the information loss caused by down-sampling while reducing computational complexity.

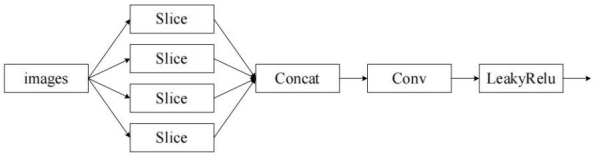


Figure 3. Schematic diagram of the Focus layer

The slicing operation principle is shown in Figure 4. Four positions with a distance of 2 are sliced and aggregated, and integrated into one channel. The original 4x4x3 size image is transformed into a 2x2x12 size feature map.

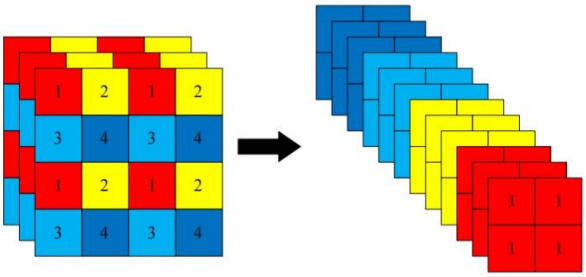


Figure 4. Schematic diagram of the slicing operation

As shown in Figure 5, the CONV convolution layer is composed of three structures: CONV convolution layer, Batch Normalization batch normalization, and SiLU activation function. CONV convolution layers are used for feature extraction, and multi-layer convolution layers can be used to obtain deeper features; The BN layer makes the model less sensitive to the parameters in the network, and makes the distribution of input data in each layer of the network relatively stable, accelerating the learning speed of the model while making the network learning more stable. It can also control gradient explosion, alleviate problems such as gradient disappearance and over-fitting. The function of SiLU activation function is to enable the network to have the ability to learn layered nonlinear problems, which is the key to the model being able to solve nonlinear problems.

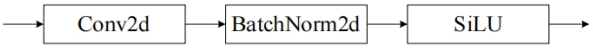


Figure 5. Schematic diagram of the CONV convolution layer

The C3 structure is a modification of the original Bottleneck CSP structure, which aims to solve the problem of high computational complexity caused by repeated gradient information in inference. The C3 structure divides the input feature map into two parts, and then merges them through the cross stage local network CSP Net [10] (Cross Stage Partial Network), which not only reduces computational parameters but also ensures high accuracy. As shown in Figure 6, this structure divides the input into two branches. One branch undergoes deep convolution and residual operations through the CONV convolution layer and the Bottleneck residual block, while the other branch only undergoes the CONV convolution layer. The two parts are then concatenated to ensure that the output of the C3 structure is the same size as the input, thus preserving the feature information of different branches more completely, allowing the model to learn more information.

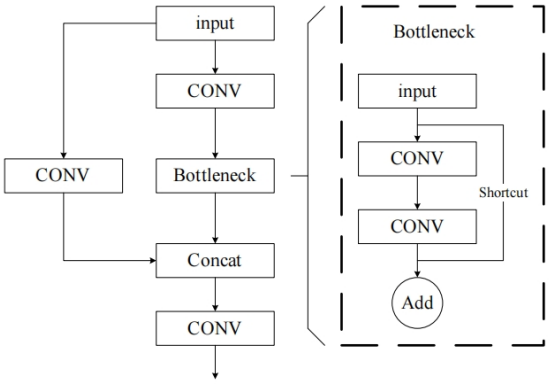


Figure 6. Schematic diagram of the C3 structure

The SPP structure [11], also known as spatial pyramid pooling, can convert feature maps of any size into fixed size feature vectors, mainly achieved by extracting feature vectors of the same size from feature maps of different sizes. The principle is shown in Figure 7. The input feature maps pass through three different sized maximum pooling layers, and then the input and feature maps obtained through the three pooling layers are concatenated. This structure also does not change the size of the input feature map, and the output and input sizes are consistent. This structure integrates local and global features, increasing the receptive field.

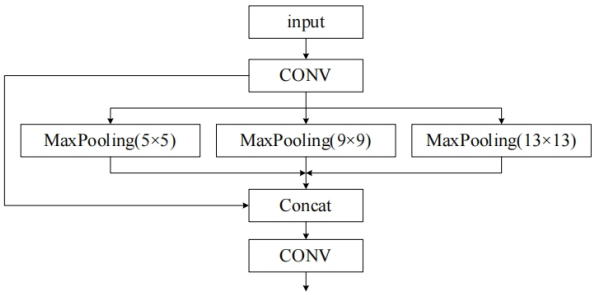


Figure 7. Schematic diagram of the SPP structure

Through studying the structures of the four components of the backbone network, it was found that although feature fusion operations exist in all four structures, the C3 structure plays a major role in achieving feature fusion, while the other three parts mainly assist in feature extraction and fusion of the backbone network. Therefore, the C3 structure, which plays a crucial role in the network, was selected for fusion with the CBAM module. After embedding the CBAM module into the Bottleneck structure of the C3 structure, the fusion results in C3_CBAM module is shown in Figure 8. Then add C3_CBAM module replaces the C3 structure in the backbone network to obtain YOLOv5s_B network model.

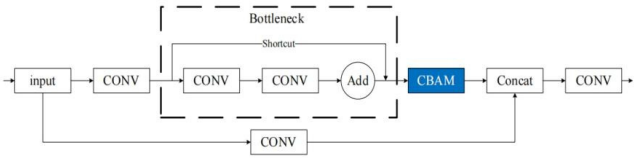


Figure 8. Schematic diagram of the YOLOv5s_B structure

(2) YOLOv5s_N

YOLOv5s_N is a new network model obtained by integrating the CBAM module with the Neck network. The Neck network layer adopts an FPN+PAN structure, which adds a bottom-up pyramid-Path Augmentation module-to the existing feature pyramid FPN [12] (Feature Pyramid Networks), as shown in Figure 9.

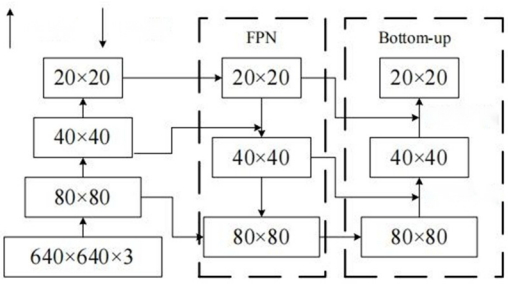


Figure 9. Schematic diagram of the FPN+PAN structure

By observing the structure of the Neck network, it is not difficult to find that there is also a C3 structure in the Neck

network. However, the FPN and PAN structures, especially the PAN structure, play a more important role in the Neck network. Therefore, after selecting to embed the CBAM module into the PAN[13] structure to complete the concat feature fusion operation, and before the C3 structure, the fusion method is shown in Figure 10.

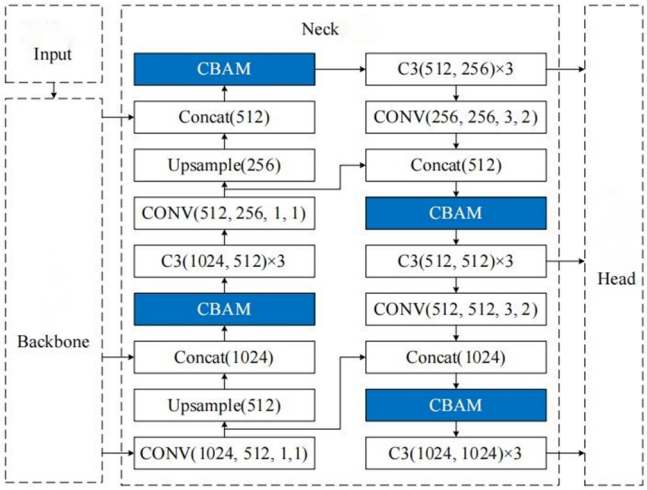


Figure 10. Schematic diagram of the YOLOv5s_N structure

(3) YOLOV5s_H

YOLOv5s_H is a new network model generated by the fusion of CBAM module and Head output terminal. The head output end is the final detection part. YOLOv5 adopts the YOLO series universal detection layer design, and the output end has three detection heads of different sizes to detect targets of different sizes based on feature maps of different sizes. Before embedding the CBAM module into three detection head convolution operations, the embedding method is shown in Figure 11.

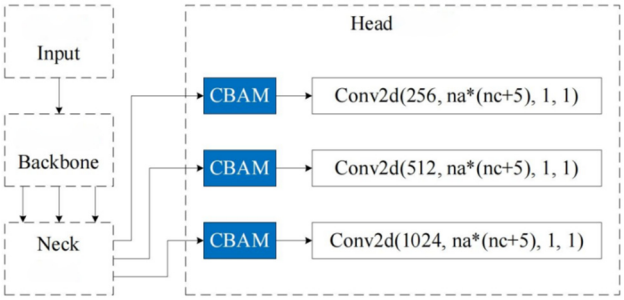


Figure 11. Schematic diagram of the YOLOv5s_H structure

3.1.2. Analysis of Experimental Results on the Fusion of CBAM Module and Three Major Network Modules

Compare the three network models in section 3.1.1 with YOLOv5s without embedded CBAM module in a comparative experiment. Using the data set crack_1 Composed of 3000 road crack images, divided into a training set, validation set, and testing set in a 7:1:2 ratio, with 100 rounds of training. The preparation of the dataset is explained in Section 4.2. Based on the evaluation index data of the crack extraction model, compare the applicability of the CBAM module and the three module fusion schemes. The detection results of the four models on the test set are shown in Table 1. The crack extraction performance of the four models is evaluated from two aspects: the average precision AP and the average intersection to union ratio mIoU.

Table 1. YOLOv5s and the measurement results data of models fused with CBAM module

Model name	average precision(AP)	mIoU
YOLOv5s	0.822	0.817
YOLOv5s_B	0.916	0.834
YOLOv5s_N	0.856	0.822
YOLOV5s_H	0.874	0.816

According to Table 1, YOLOv5s_ Model B is the best fusion scheme for the fusion experiment of CBAM module and three major network modules.

3.2. BiFPN

In the process of collecting road crack data, due to the different shapes and sizes of cracks, as well as the distance between collection, the size of the crack input data varies greatly. After being processed by multiple C3 modules in the YOLOv5s backbone network, the image size will continue to decrease by half and the underlying position information will be partially lost. Insufficient utilization of features between different scales results in limited detection accuracy of the network model.

BiFPN (Weighted Bidirectional Feature Pyramid Network) [14] is a module with efficient bidirectional feature fusion, skip connections, and weighted feature fusion mechanisms, which can simultaneously obtain global features containing high-level and low-level semantic information. The weights of different input layers are different and can be automatically updated according to the network. Its structure is shown in Figure 12 (b), It ensures that the network retains global contextual feature information to the maximum extent possible, enabling the target detection network to extract features of different scales, greatly improving the detection performance of small targets. BiFPN adds a lateral jump connection between the input and output nodes, which not only increases the input of the fusion node, but also preserves the feature information of the original node, enabling the model to fuse more features; At the same time, nodes without feature fusion on single input edges were removed, simplifying the network structure and reducing computational complexity. BiFPN can fuse more features through repeated stacking. This article introduces the BiFPN module in Neck to replace the original FPN+PAN structure, as shown in Figure 12 (a).

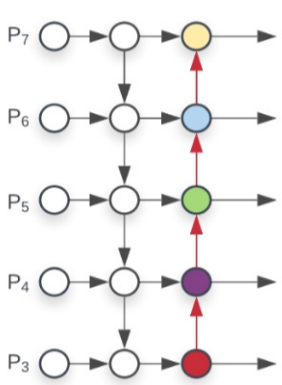


Figure 12(a): FPN+PAN

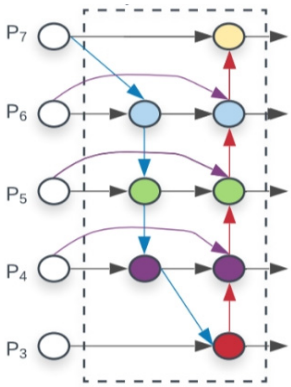


Figure 12(b): BiFPN

In summary, compared to the FPN-PAN structure that directly concatenates different inputs, the BiFPN structure has better feature fusion performance and can improve network expression ability. The specific weighting formula is shown in equation (1).

$$Out = \sum_i \frac{w_i \cdot fm_i}{\varepsilon + \sum_j w_j} \tag{1}$$

In the formula: w_j represents the learnable weight. As the model continues to train, the parameter value will change with the optimizer towards the direction of minimizing the loss function. The value is set to 1 during initialization; fm_i represents the input feature map in the network structure; ε= 0.0001; The SiLU function normalizes the weight refinement to 0-1.

The fusion method of a certain layer in the network is shown in Figure 4. According to equation (1), the feature fusion process and output are shown in equations (2) and (3).

$$P_4^{td} = Conv\left(\frac{w_1 \cdot P_4^{in} + w_2 \cdot \mathrm{Re}\,size\left(P_3^{in}\right)}{w_1 + w_2 + \varepsilon}\right) \tag{2}$$

$$P_4^{out} = Conv\left(\frac{w_3 \cdot P_4^{in} + w_4 \cdot P_4^{td} + w_5 \cdot \mathrm{Re}\,size\left(P_3^{out}\right)}{w_3 + w_4 + w_5 + \varepsilon}\right) \tag{3}$$

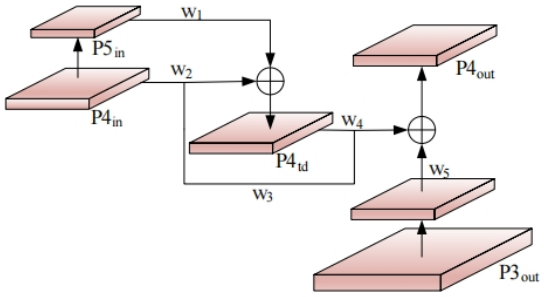


Figure 13. Schematic diagram of the BiFPN fusion method

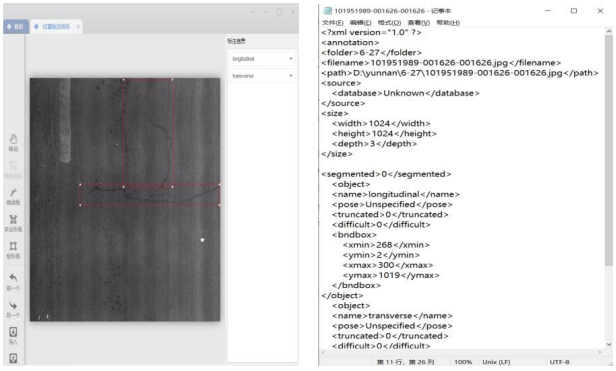
4. Experiment and Result Analysis

4.1. Experimental Environment and Parameters

The GPU of the experimental environment is NVIDIA GeForce GTX 1080Ti, and the CPU is a 6-core Intel (R) Xeon (R) CPU E5-2650v4@2.20GHz On a server with a memory of 15GB; The software part uses Pycharm as the IDE, and program design is carried out through the Python programming language and deep learning Python framework.

4.2. Experimental Data Set

The data set used in this article is images of Yunnan roads taken by inspection carts provided by the internship company. Firstly, using an open-source road crack target detection model, conduct a preliminary screening of 20000 images that may contain cracks; Then fill in the 3000 images with cracks screened out; Then use the sprite annotation assistant to annotate. Two datasets were annotated, and the label information of the first data set is only crack, named data set crack_1. The second data set divides cracks into transverse cracks, longitudinal cracks, and block cracks, with label information of transverse, longitudinal, and block, and is named data set crack_2. Annotate the road surface disease data set in the figure.



(a) (b)

Figure 14(a): Sprite annotation assistant (b): annotation result xml file

4.3. Evaluating Indicator

This article uses precision P (Precision), recall rate R (Recall), average precision AP (Average Precision), and average precision mAp (Mean average Precision) as model evaluation indicators. The specific confusion matrix is shown in Table 2. Among them, the recall rate directly reflects the ratio of actual positive samples (TP) to all positive samples (TP+FN) in the identified correct targets, as shown in equation (4); The accuracy directly reflects the ratio of actual positive samples (TP) to all correctly identified targets (TP+TF), as shown in equation (5); The average accuracy is shown in equation (6); MAP is obtained by averaging the average accuracy (AP) of all categories, as shown in equation (7).

Table 2. confusion matrix table

confusion matrix	predictive value	
	1	0
true value	TP	FN
	FP	TN

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$AP = \int_0^1 P(R) \tag{6}$$

$$mAP = \frac{1}{c}\sum_{j=1}^{c} AP_j \tag{7}$$

In the formula, AP represents the average accuracy; AP_j represents the average accuracy of the jth class object detection; C represents the category of markers; mAP represents the average accuracy mean.

$$mAP@0.5 = \frac{\sum_{j=1}^{c} AP@0.5j}{c} \tag{8}$$

In the equation: AP@0.5j The average accuracy of the jth class target when the intersection to union ratio threshold is 0.5; C represents the category of markers; mAP@0.5 Represents the average accuracy mean when the intersection to union ratio threshold is 0.5.

4.4. Experimental Result

The optimizer for model training selects SGD, with an initial learning rate of 0.01 and a training round of 300. After the model design is completed, it is necessary to train the model and adjust the parameters to obtain a relatively optimal iterative model. The Python version used in this experiment is 3.8, and the CUDA version is 10.1. The network model was built using Python 1.7. The model optimization adopts the Stochastic Gradient Descent (SGD) optimization algorithm, and the loss function adopts GIOU Loss. The parameters used in the experiment are set based on prior experience, and the specific values are shown in Table 3.

Table 3. YOLOv5s and the measurement results data of models fused with CBAM module

parameter	Value
Input size	640*640
Learning rate	0.01
Momentum	0.9
Optimizer	SGD
Weigth decay	0.0005
Batch size	8
Epoch	300

The loss curve during the model training process is shown in Figure 15. In the early stage of model training, the loss value decreases relatively quickly. As the number of training rounds increases, the loss curve gradually decreases and tends to stabilize. After 300 Epochs of training, the loss of the training and validation sets gradually decreased and stabilized to a small numerical range, and the model reached convergence.

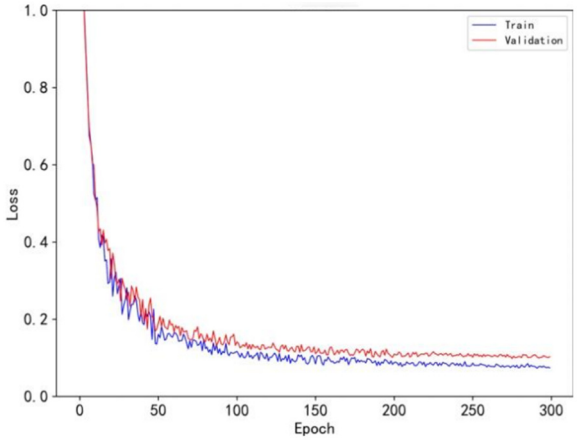


Figure 15. model training loss change curve

Compared with the traditional YOLOv5 model, the training effect of the YOLOv5 model improved by the method in this paper is better in Precision, Recall, and mAP@0.5 These three indicators have all improved, as shown in Tables 4 and 5.

Table 4. evaluation indicators of the traditional YOLOv5 model

cack type	precision	Recall	mAp@0.5
transverse crack	0.579	0.478	0.517
longitudinal crack	0.545	0.48	0.479
block crack	0.676	0.713	0.744
all	0.6	0.557	0.58

Table 5. evaluation indicators of the improved YOLOv5 model

cack type	precision	Recall	mAp@0.5
transverse crack	0.774	0.664	0.678
longitudinal crack	0.657	0.647	0.692
block crack	0.805	0.831	0.889
all	0.745	0.714	0.753

5. Conclusion

This article proposes an improved YOLOv5 road crack recognition model to address the issues of missed detection, false detection, and low accuracy of current highway cracks. Firstly, a CBAM attention module was added to the backbone network of the improved YOLOv5 model to obtain more detailed features; Then, in the feature fusion layer, BiFPN weighted bidirectional feature pyramid network is used for multi-scale feature fusion, replacing the traditional feature pyramid (FPN)+pixel aggregation network (PAN) structure to enhance feature fusion. The experimental results indicate that the improved YOLOv5s model has mAP@0.5 An increase of 17.3%.

However, this article only classifies and identifies three types of road disasters, and further improvement of the database is still needed. In addition, it is hoped to develop a road disaster classification system with a complete interface and software and hardware platform, in order to facilitate road inspection personnel to inspect and maintain the road.

Acknowledgments

This work was supported in part by Sichuan Central Inspection Technology Inc.

References

[1] CHEN F B. Analysis of the causes and hazards of cracks in concrete road pavement in my country[J]. Sichuan Cement, 2014(10): 18.

[2] OLIVEIRA H,CORREIA P L. Automatic road crack segmentation using entropy and image dynamic thresholding[C]// 2009 17th European Signal Processing Conference, 2009: 622-626.

[3] WANG X. Identification method of road cracks in complex environment based on wavelet transform directional component reconstruction [j]. Shanxi Science & Technology of Communications , 2021(4):52-55.

[4] LI Y D,HAO Z B,LEI H . Review of convolutional neural network re-search[J]. Journal of Computer Applications, 2016,36 (9) : 2508-2515 ,2565.

[5] YOU J C. Pavement crack detection based on improved Mask-RCNN[J]Video Engineering,2022,46(6) : 7-9,19.

[6] XU K, MA R G. Crack detection of asphalt pavement based on improved faster RCNN[J]. Computer Systems & Applications, 2022, 31 (7)341-348.

[7] GU S H,LI X X,WANG X Y, et al. Crack detection with enhanced se-mantic information and multi-channel feature fusion [J]. Computer Engi.neering and Application, 2021,57 (10): 204-210.

[8] YANG F, ZHANG L,YU S, et al. Feature Pyramidand Hierarchical Boosting Network for Pavement Crack DetectionJ. EEE Transactions on Intelligent Transportation Syst, 2020, (4):1525-1535.

[9] ZHANG J, QIAN S R,TAN C. Automated Bridge Crack Detection Method Based on Lightweight Vision Models [J].Complex and Intelligent Systems, 2023,9:1639-1652.

[10] Wang C. -Y, Mark Liao H. -Y, Wu Y. -H et al. "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," 2020 1EEE/CVF Conference on Computer Visionand Pattern Recognition Workshops (CVPRW), 2020, pp. 1571-1580.

[11] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Spatial Pyramid Pooling in DeepConvolutional Networks for Visual Recognition. (J. IEEE transactions on pattern analysis and machine intelligence,2015,37(9).

[12] Lin T. -Y.. Piotr D. Ross B. G., et al. Feature Pyramid Networks for Object Detection.[J]. CoRR. 2016. abs/1612.03144.

[13] Gao Huang. Zhuang Liu, Kilian et al. Densely Connected Convolutional Networks.jJCoRR,2016abs/1608.06993.

[14] Qiu Tianhao, Chen Shurong. Dual branch multi-scale joint learning pedestrian recognition based on EfficientNet [J]. Computer Applications, 2022,42 (7): 2065-2071.