Electronic Research Archive

ERA, 32(3): 1708–1730.
DOI: 10.3934/era.2024078
Received: 06 November 2023
Revised: 15 January 2024
Accepted: 22 January 2024
Published: 27 February 2024

http://www.aimspress.com/journal/ERA

Research article

Research and optimization of YOLO-based method for automatic pavement defect detection

Hui Yao[1,*], Yaning Fan[1], Xinyue Wei[1], Yanhao Liu[1], Dandan Cao[1] and Zhanping You[2]

[1] Beijing Key Laboratory of Traffic Engineering, College of Metropolitan Transportation, Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology, Beijing 100124, China

[2] Department of Civil and Environmental Engineering, Michigan Technological University, Houghton, MI 49931-1295, USA

* **Correspondence**: Email: huiyao@mtu.edu.

Abstract: According to the latest statistics at the end of 2022, the total length of highways in China has reached 5.3548 million kilometers, with a maintenance mileage of 5.3503 million kilometers, accounting for 99.9% of the total maintenance coverage. Relying on inefficient manual pavement detection methods is difficult to meet the needs of large-scale detection. To tackle this issue, experiments were conducted to explore deep learning-based intelligent identification models, leveraging pavement distress data as the fundamental basis. The dataset encompasses pavement micro-cracks, which hold particular significance for the purpose of pavement preventive maintenance. The two-stage model Faster R-CNN achieved a mean average precision (mAP) of 0.938, which surpassed the one-stage object detection algorithms YOLOv5 (mAP: 0.91) and YOLOv7 (mAP: 0.932). To balance model weight and detection performance, this study proposes a YOLO-based optimization method on the basis of YOLOv5. This method achieves comparable detection performance (mAP: 0.93) to that of two-stage detectors, while exhibiting only a minimal increase in the number of parameters. Overall, the two-stage model demonstrated excellent detection performance when using a residual network (ResNet) as the backbone, whereas the YOLO algorithm of the one-stage detection model proved to be more suitable for practical engineering applications.

Keywords: pavement engineering; pavement distress; object detection; optimization strategy; YOLO algorithms

1. Introduction

As a result of vehicle load and natural factors such as temperature, pavement tends to manifest a series of distresses. Common pavement defects mainly include cracks, potholes, etc. The initial distresses affect road capacity and users' road experience. The gradual development of the distress will bring irreparable damage to the overall structure of the road. Thus, timely pavement detection can greatly enhance driving comfort and reduce the total maintenance cost. Faced with the current vast highway traffic system, relying only on traditional manual inspection is completely insufficient to meet the needs of the industry. In contrast to the automated pavement detection approach, artificial visual detection is time-consuming and laborious, is subjective in its assessment of the deterioration of pavement, and has a certain impact on the traffic flow. Early pavement condition assessments used image data obtained from line-scanning or area-scanning cameras, such as the DHDV (Digital Highway Data Vehicle) detection system for American highways [1]. The appearance of CCD (Charge Coupled Device) digital imaging has significantly changed pavement crack detection.

In general, pavement distress detection has undergone the image processing stage, the machine learning stage, and the current stage based on deep learning. Image processing can reduce the impact of noise such as illumination, and it can be roughly classified into threshold segmentation, edge detection, and region-based segmentation. Threshold segmentation [2] distinguishes foreground and background by judging whether the feature attributes meet the threshold standard. The threshold segmentation method contains single thresholding and variable thresholding [3]. Obtaining a credible threshold value is the key to this approach. Edge detection is used to determine the boundary of the object in the image by finding the place where the gray value jumps greatly, so as to segment different objects. Commonly used edge detection algorithms include Sobel [4], Prewitt [5], Roberts, Canny [6], etc. Different edge detection algorithms have different characteristics and ranges of applications. For example, the Sobel operator can effectively suppress noise, but the location of the edge is not accurate enough.

Based on whether or not labeled data is used, machine learning can be categorized into two major methods, supervised learning and unsupervised learning. It can dig deep into features and accurately predict unknown data from existing data. Image processing techniques are commonly applied as preprocessing steps before feeding the data into machine learning models to enhance task completion. As an illustration, Marques et al. put forward a supervised learning algorithm based on a support vector machine (SVM) for automatic pavement crack detection [7]. The crack image was preprocessed to enhance the existing crack features, and the image was segmented into non-overlapping blocks to extract the feature vector of each block. However, automatic detection technology based on machine learning struggles to recognize objects with complex shapes and textures in images, and thus it is necessary to manually select suitable features and use pre-processing operations such as data cleaning.

Deep learning is a special type of machine learning that employs multi-layered neural networks to process data. As a result, deep learning models usually require more data to train and tend to require higher computational resources than traditional machine learning models. Object detection is one direction of deep learning, and the task of pavement detection mainly includes distress location and classification. With the continuous improvement of big data availability and basic computing power, many structures of CNN (Convolutional Neural Networks) with excellent performance have been proposed, such as VGGnet [8], GoolgeNet [9], ResNet [10], etc. This method of automatic extraction of high-dimensional feature information by stacking the convolution layer, pooling layer, and nonlinear

activation layer has made a surprising breakthrough in the realm of computer vision.

Object detection algorithms using deep learning can be classified into two categories: two-stage detection relying on candidate regions and one-stage detection based on regression. The two-stage algorithm generates the region proposal and then sends the region proposal to the classifier for classification and location correction. The typical two-stage algorithm R-CNN [11] integrates AlexNet [12] with selective search to complete the detection task in several independent steps. Through the continuous proposal and improvement of algorithms such as SPP-Net [13], Fast R-CNN [14], and Faster R-CNN [15], an end-to-end two-stage object detection model paradigm has been formed. One-stage detection algorithms directly regress the predicted object from the image. Typical one-stage detection algorithms include RetinaNet [16], YOLO series algorithms [17–22], and SSD series algorithms [23,24]. One-stage object detection algorithms have advantages such as fast running speed and high inference efficiency and are therefore widely used.

The YOLO series of object detection models, starting with the YOLOv1 algorithm in 2016, has since developed to the latest YOLOv7 model. The YOLOv5 model, released by Ultralytics, has a complete system architecture. YOLOv7 is currently the latest proposed network and has demonstrated state-of-the-art performance on the MS COCO dataset. Previously, Mandal et al. [25] established a real-time automatic pavement damage analysis system based on the YOLOv2 model and combined it with the distress analyzer. In this paper, the YOLO model is also applied to the field of pavement detection to further promote the application research of deep learning-based technology in automated pavement detection. First, the characteristics of the modules used in YOLOv5 and YOLOv7 are introduced, followed by an explanation of the overall structure of the models. The experiment aimed to explore an appropriate IOU loss function that accommodates the diverse characteristics of pavement distress. Additionally, the study attempted to optimize model performance by incorporating several mainstream attention mechanisms, namely simAM, CBAM, and SA. By analyzing the model parameters and detection results of YOLOv5 and YOLOv7, we conducted research on and discussed improvement strategies and proposed subsequent improvement strategies and research directions.

2. Model structures

The two-stage object detection model is not the focus of this chapter because of its long training time and difficulty in meeting the requirements of real-time detection. This section focuses on introducing the distinctive structures applied in YOLOv5 and YOLOv7 and analyzing the functional characteristics of each part.

2.1. YOLOv5 model highlights

Since the inception of the first generation of YOLO models in 2016, the model architecture has gradually matured and has been widely applied in industries such as agricultural pest recognition [26], marine water pollution, and transportation [27]. As depicted in Figure 1, the YOLOv5 model, introduced in 2020, is a powerful real-time object detection model. In this study, the YOLOv5s model with 16.0 GFlops and 270 layers was used.

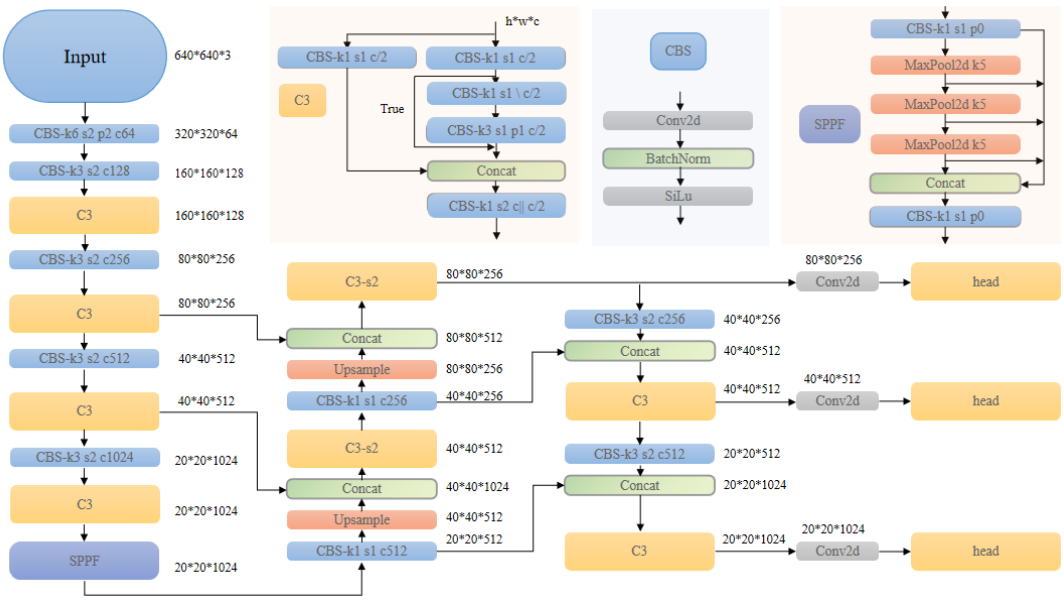


Figure 1. The original structure of the YOLOv5.

2.1.1. Focus module

The Focus structure is a way of downsampling while maintaining information losslessly, and it is used to extract features in the input feature map. As indicated in Figure 2, this layer operates similarly to the scaling layer by dividing the input feature map into four sections and subsequently combining them in a particular sequence to generate a novel feature map. This design better utilizes the correlation between channels, has higher resolution and richer detail information, can better capture subtle features of the target, and has a significant effect on the detection of small targets. The feature maps of 3 channels are transformed to 12, and the size is changed from H × W to H / 2 × W / 2. In the pursuit of efficiency, the Focus module has been replaced by a 6 × 6 convolutional kernel.

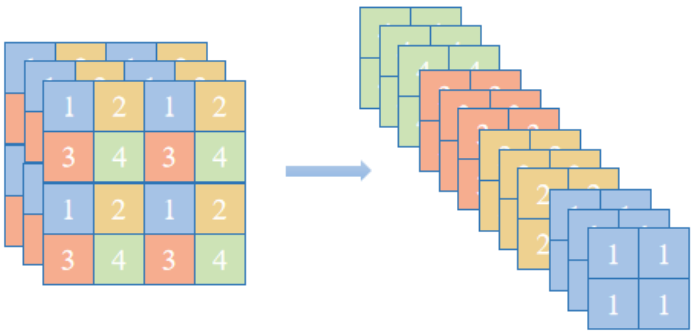


Figure 2. Focus diagram and module structure.

2.1.2. Improvement of SPP module to SPPF network module

The SPP (Spatial Pyramid Pooling) structure is a network structure that facilitates multi-scale

feature fusion and is employed to extract features from input images. The main idea is to perform pooling on different receptive fields to preserve feature information at different levels. As shown in Figure 3, by concatenating features of different scales, a higher dimensional feature vector is obtained. This feature vector contains feature information of different scales and can better capture target details. The backbone network of YOLOv5 includes a CSPDarknet53 structure and an SPP structure. The CSPDarknet53 structure is utilized to capture low-level features, whereas the SPP is applied to capture high-level features and merge features of different scales.

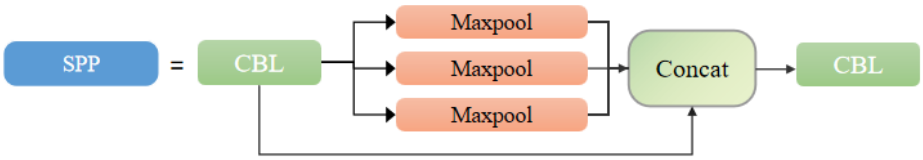


Figure 3. Spatial pyramid pooling structure.

The SPPF module was introduced in the later version of YOLOv5 (6.0). This structure is equivalent to improving the parallel structure of the model into a serial processing mode (Figure 4). Two 5 × 5 pooling layers are serialized together equivalent to a 9 × 9 pooling layer. The calculation efficiency of the model is greatly improved after replacement.

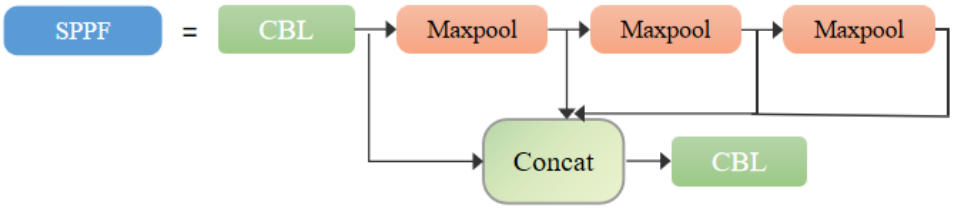
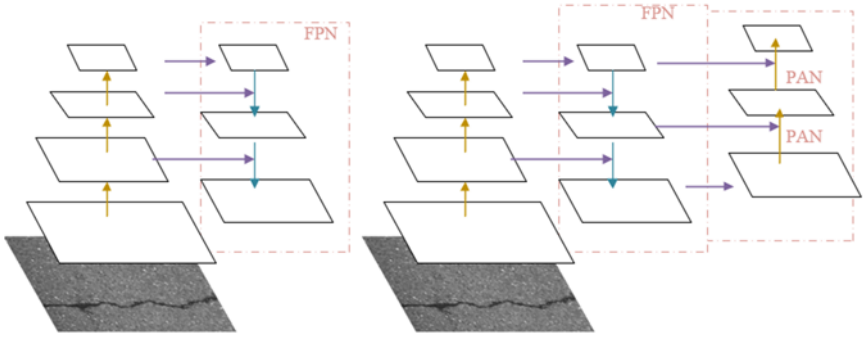


Figure 4. Spatial pyramid pooling fast structure.

2.1.3. PAN structure

Unlike the FPN used in YOLOv3, the neck part of YOLOv5 uses a PAN (Path Aggregation Network) structure [28], where features from different levels are fused through different paths. Illustrated in Figure 5, FPN utilizes a top-down methodology that combines high-level with low-level features via upsampling to obtain the feature map for subsequent prediction. This structure mainly transmits the strong semantic features of the high level and enhances the whole pyramid. However, it merely amplifies the semantic data and fails to transmit the positional information. The PAN structure mainly adds a bottom-up pyramid behind the FPN (feature pyramid network) to transfer the positioning characteristics of the lower level to augment the accuracy and robustness of the model.



(a) (b)

Figure 5. YOLOv5 backbone network upgrade: (a) the original feature pyramid network (FPN), (b) path aggregation network (PAN).

2.1.4. The setting of the loss function

The loss of YOLOv5 is comprised of class loss, objectness loss, and location loss. Class loss uses BCE loss (Binary CrossEntropy Loss) to calculate the classification loss of positive samples only, while objectness loss utilizes BCE loss but calculates the loss of all samples. Location loss is used, and CIOU (Complete IoU) loss is used to calculate only the location loss of positive samples. Different weights are applied to objectness loss in the three prediction feature layers of the model. Small targets are prone to problems such as false detection and missing detection, so greater weights are assigned.

$$L_{obj} = 4.0 \cdot L_{obj}^{small} + 1.0 \cdot L_{obj}^{medium} + 0.4 \cdot L_{obj}^{large}$$

2.2. YOLOv7 model highlights

In 2022, YOLOv7 was proposed (Figure 6). This model continues to follow the idea of the YOLO series, which considers the problem of object detection as a regression task and mainly introduces new modules of the E-ELAN module, SPPSPC module, and REP module.

ELAN (Effective Long-Range Aggregation Network) modules are designed to allow deep networks to converge quickly and learn feature information efficiently. The SPPSPC (Spatial Pyramid Pooling and Fully Connected Spatial Pyramid Convolution) module uses the idea of a pyramid pool and residual link for feature fusion. The RepVGG network is proposed based on the VGG (Visual Geometry Group) network in the REP (Re-parameterization) module. The RepVGG network has surpassed the ResNet, EffcientNet [29], and ResNeXt networks in both accuracy and speed. In the model training, the multi-branch model is used to strengthen the representation ability. In the actual test, the transformed single-line model is used to accelerate the reasoning speed and reduce memory consumption.

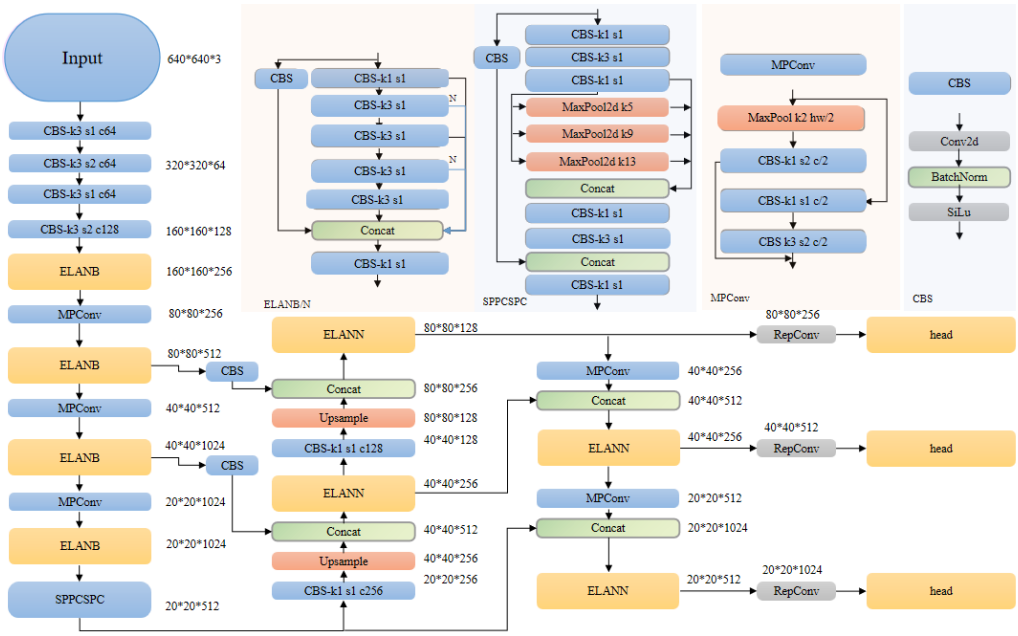


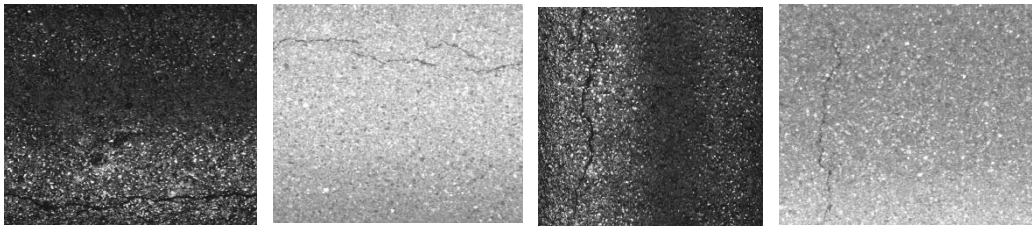
Figure 6. The original structure of the YOLOv7.

3. Improvement strategy of pavement detector

The experiment primarily focused on optimizing the dataset and model structure. It used data images that are more suitable for pavement preventive maintenance and improved the model structure accordingly. The YOLOv7 model, consisting of 407 layers, has a parameter size of 105.2 Gflops, which is much larger than that of the YOLOv5s model (16.0 Gflops). Considering the practical deployment of the model on mobile devices, YOLOv5 was chosen as the primary model for experimentation, improvement, and evaluation.

3.1. Optimization strategy for the dataset

To better meet the practical needs of pavement maintenance, the experiment collected a significant number of pavement cracks on newly constructed roads and combined data augmentation techniques to make the cracks more in line with the characteristics of pavement distress. Therefore, the model can be used for crack identification in road preventive maintenance, especially for cracks that are difficult to detect through manual inspection on newly constructed roads. As shown in Figure7, there are two types of crack distresses on asphalt pavement: one is longitudinal cracks caused by factors such as foundation settlement and traffic load, and the other is transverse cracks commonly caused by shrinkage stress exceeding tensile strength. This paper also collected images of early cracks on concrete pavement.



(a) (b) (c) (d)

Figure 7. Early minor pavement cracks (a: transverse crack on asphalt pavement, b: transverse crack on concrete pavement, c: longitudinal crack on asphalt pavement, d: longitudinal crack on concrete pavement).

3.2. Attention mechanism

The attention mechanism is a deep learning technique that helps models more accurately predict the categories and bounding boxes of objects in images. The primary function of the attention mechanism is to assist the model in highlighting vital characteristics in the image, which enhances the model's performance. The attention mechanism is a subnetwork that is encapsulated into a module during actual model improvement.

3.2.1. CBAM attention mechanism

The Convolutional Block Attention Module (CBAM) network was proposed in 2018 [30]. The innovation lies in that feature maps not only contain rich attention information in channels but also contain a significant amount of attention information between pixels in the feature map. Previous attention mechanisms only focused on attention information in channels, which wastes attention information in space. As shown in Figure 8, CBAM adjusts the weights of each channel in the feature map by building a Channel Attention Module (CAM) to learn the importance of each channel. By constructing a Spatial Attention Module (SAM) to learn the importance of each spatial position, the network places greater emphasis on spatial positional information and improves the quality of feature representation. These two sub-modules are combined to obtain more comprehensive and reliable attention information, providing more reasonable guidance for the allocation of computing resources.

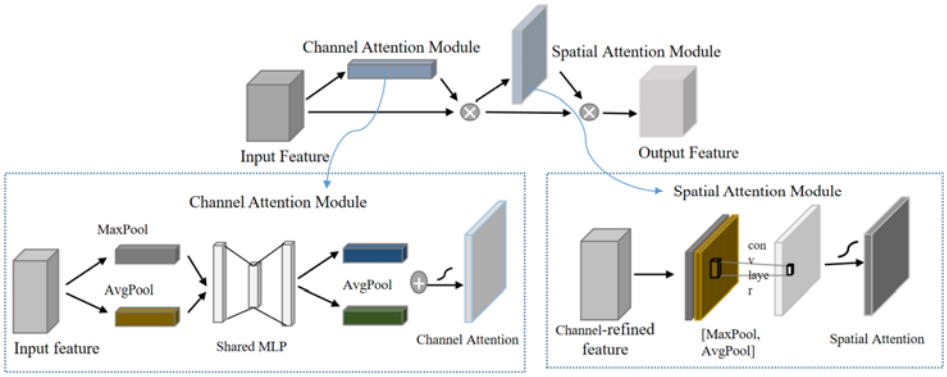


Figure 8. CBAM network structure.

3.2.2. Shuffle attention mechanism

The shuffle attention (SA) module [31] groups channels into multiple sub-features to effectively establish a multi-branch structure, thereby parallel processing different branches. For the grouped sub-features, SA utilizes channel shuffle network units to construct both channel attention and spatial attention simultaneously, fully leveraging their correlation to improve efficiency. As shown in Figure 9, the feature map (C * H * W) is segmented into G groups along the channel dimension. Each sub-feature is further divided into two branch dimensions of $\frac{C}{2G}$ * H * W, where one branch focuses on the inter-channel relationships and the other branch generates a spatial attention map. This processing enables the model to give better attention to both the target and its spatial position. Afterward, sub-branches are processed, aggregated, and then subjected to a channel shuffle operation similar to that of ShufflenetV2. This enables inter-group information interaction between channels. Currently, computer vision research mainly focuses on capturing pixel-level pairwise relationships and channel dependencies, i.e., spatial attention and channel attention. The SA module addresses the issue of increased computational costs resulting from the fusion of these two mechanisms by combining both types of attention using a channel shuffle operation.

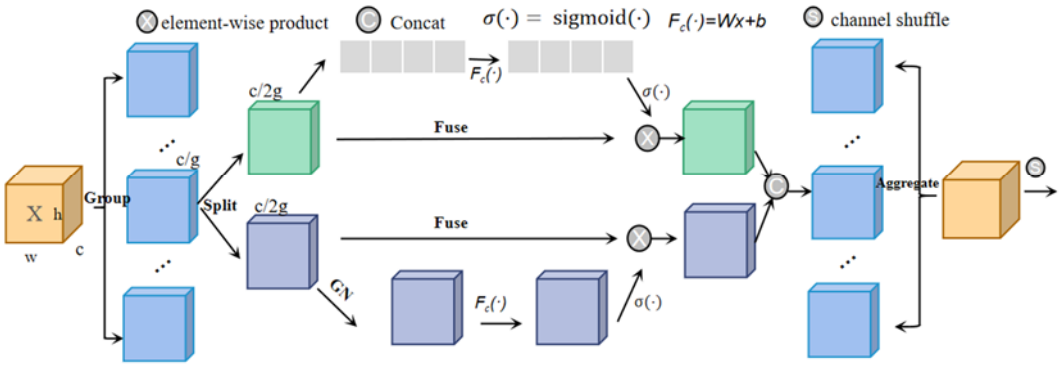


Figure 9. Shuffle attention (SA) network structure.

Compared with CBAM, which integrates both spatial and channel attention structures in its module, the SA module achieves higher efficiency by borrowing the lightweight unit design from ShufflenetV2. In this study, both SA and CBAM modules were applied to pavement defect detection from pavement data for comparison.

3.2.3. SimAM attention mechanism

To allow the neural network to learn distinguishable neurons fully, the authors suggest inferring three-dimensional weights from the current neuron. Compared with most 1D channel attention mechanisms or 2D spatial attention mechanisms, the SimAM attention mechanism infers the 3D weights of a feature map using an energy function that does not require adding parameters to the original network [32]. In neuroscience, active neurons can affect neighboring neurons' activities, indicating that neurons with obvious spatial inhibition effects should be more important in visual

processing. Based on the findings in neuroscience, an energy function defining what each neuron should possess is determined, and the energy function in Eq (1) is obtained through optimization by regularization.

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(w_t x_i + b_t))^2 + (1-(w_t + b_t))^2 + \lambda w_t^2 \quad (1)$$

Theoretical analysis suggests that each channel has an energy function, with t representing the neuron and λ serving as a hyperparameter. The result obtained through analysis is as follows:

$$w_t = -\frac{2(t-\mu_t)}{(t-\mu_t)^2+2\delta_t^2+2\lambda} \quad (2)$$

$$b_t = -\frac{1}{2}(t+\mu_t)\omega_t \quad (3)$$

μ_t and σ_t^2 are as follows:

$$\mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i \quad (4)$$

$$\sigma_t^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}(x_i - u_t)^2 \quad (5)$$

The minimum energy is calculated by using the following equations:

$$e_t^* = \frac{4(\sigma_t^2+\lambda)}{(t-\overline{\mu})^2+2\overline{\sigma}^2+2\lambda} \quad (6)$$

$$\overline{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i, \quad \overline{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \overline{\mu})^2 \quad (7)$$

The module ends with a sigmoid function that does not affect the relative importance between neurons.

$$\tilde{X} = sigmoid(\frac{1}{E})\odot X \quad (8)$$

Unlike the aforementioned attention mechanism, SimAM directly evaluates the importance of independent neurons, assigning higher weights to important neurons without augmenting the number of model parameters.

3.3. Loss function

The loss function can help the model more accurately predict the category and bounding box of objects in an image, directly affecting the performance of the model. Traditional object detection regression losses characterize the four variables of rectangular boxes, meaning that these four variables are independent of each other. Considering their correlations, the IOU is introduced to the regression loss [33]. IOU only focuses on the overlapping area, while GIoU [34] better reflects the degree of overlap between the predicted box and the ground truth by solving the ratio of the intersection area and the union area of the two bounding boxes. DIoU [35] considers the distance factor between the predicted and ground truth boxes on the basis of GIoU, directly minimizing the distance between the

two boxes and greatly speeding up the convergence rate of the model. The loss function of the DIOU (L_{DIOU}) is calculated as follows:

$$\mathrm{DIOU} = \mathrm{IOU} - \frac{\rho^2(\mathrm{b},\mathrm{b}^{\mathrm{gt}})}{\mathrm{C}^2} \tag{9}$$

$$\mathrm{L}_{\mathrm{DIOU}} = 1 - \mathrm{DIOU} \tag{10}$$

CIOU considers that the aspect ratio in bounding box regression increases the scale loss of the detection box, thus making the aspect ratio of the predicted box more consistent with the ground truth [36].

$$\mathrm{L}_{\mathrm{CIOU}} = 1 - \mathrm{IoU} + \frac{\rho^2(\mathrm{b},\mathrm{b}^{\mathrm{gt}})}{\mathrm{c}^2} + \alpha\mathrm{v} \tag{11}$$

$$\alpha = \frac{\mathrm{v}}{1-\mathrm{IoU}+\mathrm{v}} \tag{12}$$

$$\mathrm{v} = \frac{4}{\pi^2}\left(\arctan\frac{\mathrm{w}^{\mathrm{gt}}}{\mathrm{h}^{\mathrm{gt}}} - \arctan\frac{\mathrm{w}}{\mathrm{h}}\right)^2 \tag{13}$$

Taking into account that the definition of v in CIOU reflects the difference in aspect ratio between the ground truth box and predicted box, rather than the actual width and height, it is considered that the design of CIOU loss for aspect ratio loss is not reasonable. The CIOU loss may optimize the similarity measure in a way that is not suitable, which prevents the model from smoothly transitioning from (w,h) to (wgt, hgt). As a result, EIOU [37] is optimized by dividing the loss function into IOU loss, distance loss, and orientation loss.

$$\mathrm{L}_{\mathrm{EIOU}} = \mathrm{L}_{\mathrm{IOU}} + \mathrm{L}_{\mathrm{dis}} + \mathrm{L}_{\mathrm{asp}} \tag{14}$$

$$\mathrm{L}_{\mathrm{EIOU}} = 1 - \mathrm{IoU} + \frac{\rho^2(\mathrm{b},\mathrm{b}^{\mathrm{gt}})}{(\mathrm{w}^{\mathrm{c}})^2+(\mathrm{h}^{\mathrm{c}})^2} + \frac{\rho^2(\mathrm{w},\mathrm{w}^{\mathrm{gt}})}{(\mathrm{w}^{\mathrm{c}})^2} + \frac{\rho^2(\mathrm{h},\mathrm{h}^{\mathrm{gt}})}{(\mathrm{h}^{\mathrm{c}})^2} \tag{15}$$

In addition, imbalanced training samples are a common problem in experiments, as there are significantly fewer high-quality samples with small regression errors than there are low-quality samples. Therefore, actual training needs to focus on learning the contributions of high-quality samples. In regression problems, outliers (samples with large differences between predicted values and true label values, i.e., low-quality samples) may cause the model's training process to be unstable and even result in overfitting. In practical applications, some techniques are typically used, such as replacing the MSE loss function with SmoothL1 loss. Dynamic R-CNN re-weights the bounding boxes based on the SmoothL1 loss [38], but this improvement only strengthens the contribution of high-quality samples and does not weaken the influence of low-quality samples on the gradient. The Focal loss, which was introduced in this experiment, dynamically reduces the weights of easily distinguishable samples during the training process through a dynamic scaling factor, thereby quickly focusing on the difficult-to-distinguish samples.

$$\mathrm{L}_{(\mathrm{f})} = \begin{cases} -\frac{\mathrm{ax}^2(2\ln(\beta\mathrm{x})-1)}{4}, 0 < \mathrm{x} \le 1; \ \ 1/\mathrm{e} \ \le \beta \le 1, \\ -\alpha\ln(\beta)\mathrm{x} \ + \ \mathrm{C}, \ \mathrm{x} > 1; \quad 1/\mathrm{e} \ \le \beta \le 1 \end{cases} \tag{16}$$

C is a constant. To ensure that $L_{(f)}$ in Eq (16) is continuous at x = 1, $C = (2\alpha \ln\beta + \alpha)/4$. In practical applications, Focal-EIOU does not directly replace x in the equation with EIOU Loss. To avoid small gradient problems caused by EIOU that affect parameter optimization, the value of IOU was used to re-weight it, and the expression is as follows:

$$L_{Focal-EIOU} = IOU^{\gamma} L_{EIOU} \tag{17}$$

This section is mainly based on the pavement distress images to optimize the loss function suitable for pavement distress characteristics through training model performance. The Focal loss idea is not only used for the improvement of EIOU but also analyzed with CIOU to study the influence of the loss function. By optimizing the loss function, the model parameters can be optimized to better fit the training data.

4. Experiment research

4.1. Data set preparation

The data used for model training was collected from various road sections. By merging it with the publicly available dataset, Crack Forest, a total of 7710 images depicting pavement defects are obtained. To enhance the robustness of the model, data augmentation techniques were utilized to extend the dataset, including angle rotation, horizontal flip, and vertical flip. The data was divided into the training set, validation set, and test set in a ratio of 6:2:2 before being fed into the model. LabelImg was used to label the distresses. According to Figure 10, four types of distresses were labeled based on the collected pavement data: transverse cracks, longitudinal cracks, alligator cracks, and potholes. The dataset incorporates the mentioned pavement micro-cracks.

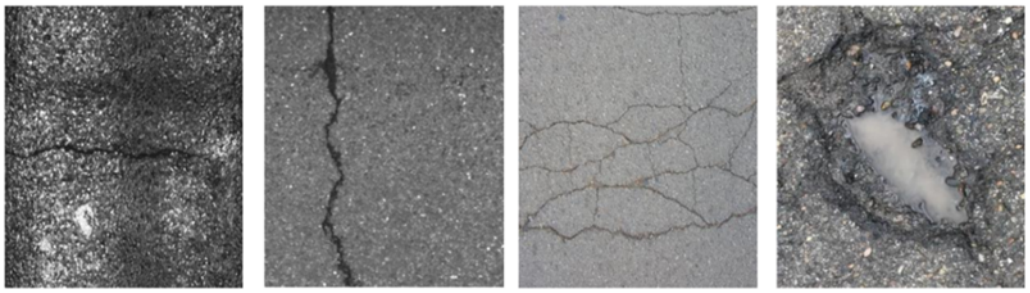


Figure 10. Raw images: (a) transverse crack, (b) longitudinal crack, (c) alligator crack, (d) pothole.

4.2. Evaluation metrics

To explain the metrics, it is necessary to introduce the concept of Intersection over Union (IOU), where the IOU value represents the ratio of the intersection between the ground truth and predicted bounding boxes to their union. In other words, a higher IOU value indicates a closer agreement between the prediction and the actual annotation. Generally, the IOU threshold can be adjusted; when the IOU value between the predicted and ground truth bounding boxes exceeds the threshold, the

prediction is categorized as True Positive (TP); otherwise, it is classified as False Positive (FP) if the IOU value falls between 0 and the threshold. Figure 11 illustrates this process, taking a threshold of 0.5 as an example.

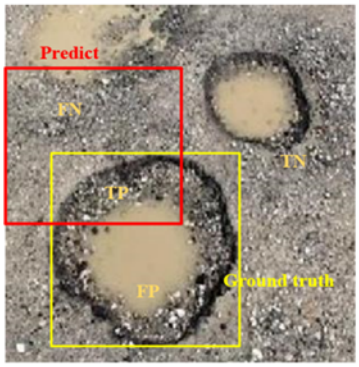


Figure 11. Interpretation of the TP, FP, FN, TN.

$IOU = \frac{TP}{TP+FP+FN} > 0.5$, the predicted result is TP.

$IOU = \frac{TP}{TP+FP+FN} < 0.5$, the predicted result is FP.

Precision (the abbreviation "P" will be referred to later) represents the ratio of the predicted result that is correct:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (the abbreviation "R" will be referred to later) refers to the proportion of correctly predicted instances to the ground truth annotations:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Multiple sets of P and R can be obtained by adjusting different IOU thresholds. The P-R curve is plotted with the vertical coordinate P and the horizontal coordinate R. The area enclosed by the curve and the coordinate axis is the AP (Average Precision). The calculation of AP involves the computation of the area under the precision-recall curve, which provides a measure of the balance between precision and recall values and ranges from 0 to 1. The AP of each type of distress is summed and averaged to obtain the mAP (mean Average Precision). The mAP is obtained by aggregating the AP scores across all categories, thereby providing a comprehensive performance evaluation metric for the model across different classes.

4.3. Model training and result description

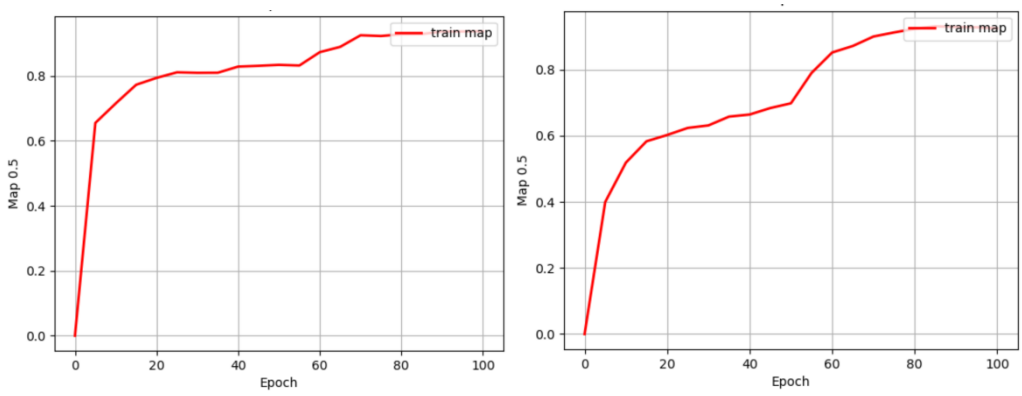
4.3.1. Model operating environment

The experimental setup featured Windows 10 as the operating system, an Intel(R) Core (TM) i9-10885H CPU@2.40GHz, and an NVIDIA Quadro RTX 5000 with Max-Q Design GPU having a

memory capacity of 16 GB. After testing, it was found that setting the parameter batch size to 16 achieved the desired result. The deep learning framework used was Pytorch 1.10.0, with CUDA 11.3.

4.3.2. Results and analysis

The model training utilized pre-trained weights. The training of Faster R-CNN consisted of two stages. In the first stage, the backbone network was frozen, and only the parameters of the feature extraction network were updated. This stage involved training for 50 epochs. Subsequently, the entire model was trained in the second stage, also for 50 epochs. The figure below illustrates the trend of the mAP curve during the model training process, indicating that the curve gradually stabilizes around epoch 80.



(a) (b)

Figure 12. mAP curve: (a) Resnet, (b) VGG.

The experimental data for the two-stage object detection algorithm Faster R-CNN are presented in Table 1.

Table 1. Comparison data of the different algorithms.

	AP (%)				mAP (%)	Model weight file size
	Transverse crack	Longitudinal crack	Alligator crack	Pothole		
faster-rcnn-resnet50	88.38	88.57	98.44	99.87	**93.81**	108 M
faster-rcnn-vgg	87.17	86.27	97.99	**100**	92.86	521 M
YOLOv4	83.85	83.71	85.41	73.69	81.66	245 M
YOLOv5	86.8	**89.4**	89.0	98.9	91.0	**13.7 M**
YOLOv7	**88.8**	86.6	**99.4**	98.1	93.2	284 M

The experiments demonstrate that the two-stage object detection algorithm exhibits better performance. Models employing ResNet as the backbone network outperform utilizing the VGG architecture. Although the YOLOv5 object detection algorithm may have slightly inferior detection

effectiveness, it possesses a significant advantage in terms of lightweight design compared to Faster R-CNN-VGG and YOLOv7. The YOLOv5 models occupy much less memory, with a size not exceeding 20MB. On the other hand, YOLOv7 has a higher parameter count (105 Gflops) compared to YOLOv5 (16.5 Gflops). YOLOv3, an earlier version of this algorithm, was applied in the industrial domain. YOLOv4, built upon its foundations and optimized, has achieved significant improvements. However, this model performs poorly in detecting pothole distress during experiments, resulting in an overall detection performance that falls short of expectations. Therefore, future improvements are inclined toward deploying models that cater to real-time detection requirements.

To better understand the impact of the various improvement measures on network performance, subsequent improvement experiments were conducted, focusing on attention mechanism and loss function optimization. The models were compared based on these enhancements, and the results are presented in Table 2. The model resulting from the incorporation of the CBAM module and replacing the boundary box positioning loss (L_{bbox}) with the Focal CIOU loss function was designated as V5-CBAM-FC. The YOLOv5 model integrates the Shuffle Attention (SA) module and is improved based on the Focal EIOU (FE) technique, resulting in the designated name V5-SA-FE. This naming convention was consistently followed throughout the table to ensure uniformity and clarity in the presentation of the models.

Three types of attention mechanisms (CBAM, simAM, Shuffle Attention) and different IOU loss functions (CIOU, EIOU, Focal CIOU, Focal EIOU) were applied in this experiment. The number of layers in the YOLOv7 model is approximately 1.5 times that of the YOLOv5 model, so no more structural layers were added. The batch size is uniformly configured as 16 for both the training and testing phases.

Table 2. Comparisons of model improvement effects.

Number	Model	Parameter	Precision	Recall	mAP@0.5	Inference
1	V5-CBAM-C	7063284	0.93173	0.89127	0.93027	**27.9 ms**
2	V5-CBAM-E		0.9051	0.89314	0.9259	28.5 ms
3	V5-CBAM-FC		0.91093	0.88488	0.92326	28.4 ms
4	V5-CBAM-FE		0.89782	0.89902	0.92341	28.3 ms
5	V5-SimAM-C	7030417	0.93228	0.88278	0.92915	28.0 ms
6	V5-SimAM-E		0.92882	0.89015	0.92859	28.3 ms
7	V5-SimAM-FC		0.9382	0.88251	0.92838	28.3 ms
8	V5-SimAM-FE		0.92386	0.86795	0.91748	28.3 ms
9	V5-SA-C	7030609	0.90892	0.91812	0.93057	28.8 ms
10	V5-SA-E		0.9005	0.8954	0.92339	29.1 ms
11	V5-SA-FC		0.9243	0.89179	**0.93233**	29.0 ms
12	V5-SA-FE		0.90062	0.88479	0.92327	29.1 ms
13	V7-C	37622682	0.9165	0.9134	**0.93280**	60.0 ms
14	V5-C	7020913	0.90826	0.89626	0.92407	27.0 ms

Overall, adding attention mechanisms can improve the performance of the original model to varying degrees. However, the SimAM attention mechanism is not very effective, despite the fact that this module avoids excessive adjustments to the model structure by influencing weights through energy functions. Both the CBAM and SA attention mechanisms perform significantly better than the original

model (V5-C) and achieve nearly the same level of mAP as YOLOv7. From the perspective of the IOU loss function, EIOU is not generally suitable for pavement crack distresses, but the Focal loss idea proposed by the authors can improve CIOU. In combination with the SA attention mechanism, the mAP of the model has been improved by 0.826% compared with the original.

In terms of model inference speed, the YOLOv5-based improved models are significantly more suitable for real-time detection, with detection speeds about twice that of YOLOv7. The YOLOv7 model has a parameter size about five times larger than the YOLOv5 model, which is not conducive to model deployment. The best-performing models among the above are V5-CBAM-C and V5-SA-FC. The model using the SA attention mechanism achieves mAP@0.5 of 0.93233, achieving a good balance between precision and recall.

To further demonstrate the effectiveness of the model improvements, the experiment evaluated the accuracy of models with different attention mechanisms and predicted the performance of the models on specific categories, as shown in Figure 13. Tests have shown that the SimAM attention mechanism exhibits significantly superior performance on pothole distress, but the performance on crack distresses is not outstanding. The added attention mechanism modules have all improved the original model. Testing provides a reliable guarantee for the practical application of the model. In future experiments, different combinations of attention mechanisms will be attempted to improve detection accuracy for different distress types.

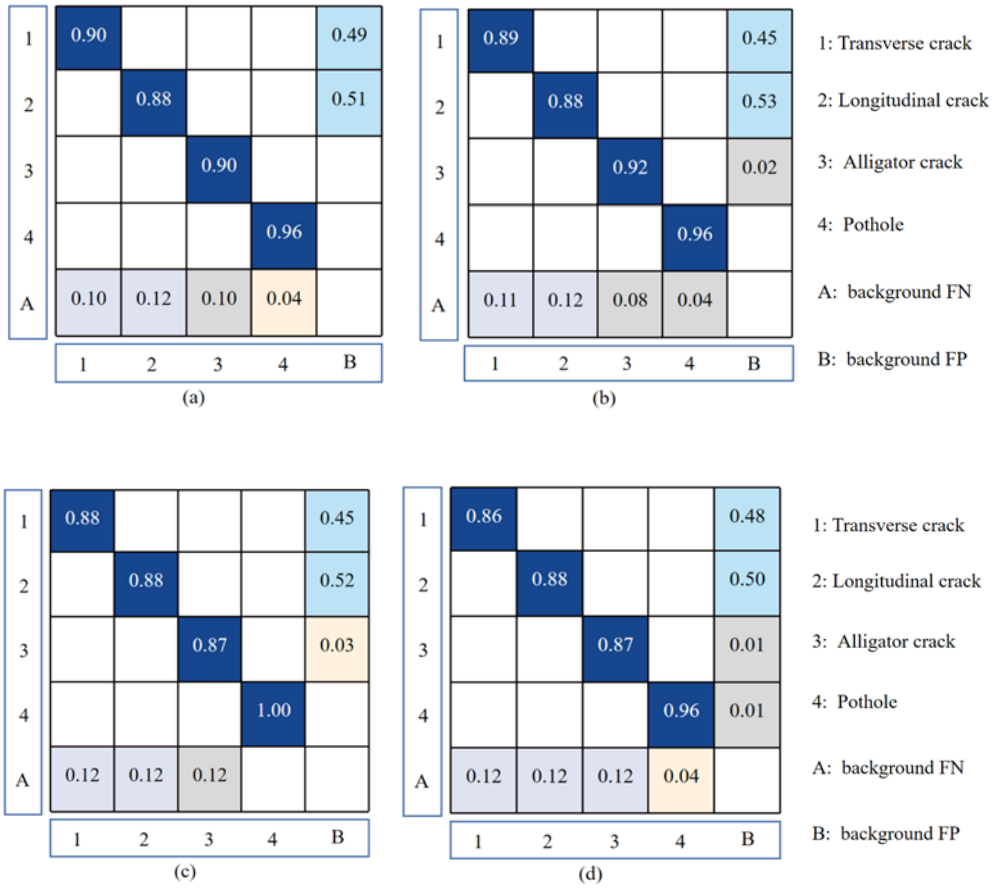


Figure 13. YOLOv5 confusion matrix of different attention mechanisms: (a) CBAM, (b) SA, (c) SimAM, (d) original model.

5. Discussion

5.1. Further discussion based on pavement detection

The data set of pavement distresses includes longitudinal cracks, transverse cracks, alligator cracks, and potholes. It is not difficult to find that in the target characteristics, the length and width of crack distresses are relatively large, and the marking frame is slender, while the size changes of mesh cracks and pits are not obvious. Replacing traditional convolution operation fixed sampling position with deformable convolution might have unexpected effects. YOLOv3 and YOLOv4 introduced the BN layer, high-resolution classifier, multi-scale training, and other methods into the model structure, which improved the accuracy and recall rate. In this experiment, YOLOv5 and YOLOv7 with high precision and high speed were compared to explore the improvement method of the applicable model for pavement distress detection. The results show that YOLOv7 does have some improvement based on YOLOv5, but it is much worse than YOLOv5 in terms of model weight. The weight file of YOLOv4 (245.8M) is about ten times larger than that of YOLOv5. The weight file of the best.py model trained by YOLOv7 reaches 284 M. Therefore, the focus of promoting this technology and deploying it on mobile devices is to carry out lightweight processing such as pruning and distillation on the model. Besides enhancing the IOU loss and incorporating attention mechanisms, the effectiveness of new training strategies and transfer learning methods has been notable. Notably, studies focusing on improving the YOLOv3 model have successfully employed a straightforward semi-supervised label distillation (SSLD) technique to acquire pre-trained models [39]. This implementation demonstrates the advantages of SSLD in the domain of model pre-training.

Furthermore, the performance of the model largely depends on the quality of the dataset, so other approaches need to be considered in the research. To address the challenges in data collection and pavement distress identification techniques, Ma et al. employed a road-surface generative adversarial network for image generation [40]. They proposed an optimized YOLO-MF model based on YOLOv3, which achieved accurate crack tracking and counting through accelerated algorithms and median flow algorithms. Therefore, future research should focus more on addressing the issue of poor pavement distress detection in complex environments during actual detection. Previous studies have addressed the challenges of noise interference and redundant information extraction in practical applications by employing the YOLO model and an omni-scale network (OSNet) to achieve accurate localization of pavement defects [41]. The utilization of unmanned aerial vehicles (UAVs) for detection purposes has emerged as a mainstream direction [42], which is expected to gain even greater practical significance in the future.

5.2. The use of attention mechanisms

By emphasizing the salient aspects of the target, the attention mechanism offers a means to augment the model's focus and minimize the impact of extraneous factors, ultimately enhancing the discriminative power of the network. In practical applications, incorporating attention mechanisms such as SE, ECA, CBAM, and SK is currently a mainstream technique. The recognition of diverse objects necessitates continuous exploration of appropriate augmentation methods and attention modules. Yao et al. developed the Pyramid Region Attention Module (PRAM) to achieve precise extraction of crack information [43], enabling efficient global multi-scale context integration and

capturing long-range dependencies with reduced computational cost. Li et al. improved the YOLOv5 model for rice field spike detection by introducing the CBAM attention mechanism to solve the problem of gradient disappearance during training [44]. The recall rate of the improved algorithm reached 98.1%, and the mAP reached 94.3%, effectively solving the problem of missing small object detection, which meets the needs of crop detection and counting. For pedestrian detection, the AS-YOLO improved model was proposed based on YOLOv4 [45]. The introduction of residual networks and SE attention mechanisms with small parameters improved the model's performance by 2.02% in terms of mAP compared with YOLOv4, while increasing the detection speed by 16.63%. This provided good technical support for pedestrian behavior analysis. In the field of autonomous driving, researchers have added channel attention mechanisms and spatial attention mechanisms to YOLOv3 to correctly conduct pavement object detection, and the improved model performance is significantly better than the YOLOv3 algorithm [46]. In addition, attention mechanisms have found extensive application in the domain of semantic segmentation [47], face recognition [48], medical image processing [49], 3D vision [50], and other fields.

For pavement defects with large-scale changes and data sets containing strong noise interference, experiments have shown that not all attention mechanisms can improve the detection performance of distresses, and additional research is needed to explore attention mechanisms that can recognize the characteristics of such distress targets. The attention mechanism in this study is added to the model backbone. In the context of relevant studies, Yao et al. conducted extensive experimental research to investigate the impact of attention module placement, quantity, and integration methods on model performance [51]. Their research findings revealed a 6.7% improvement in performance within the enhanced model. This study explored the impact of attention mechanisms on distress category detection, but further research is needed on whether the combined use of different attention mechanisms will affect the model performance.

5.3. Optimization of IOU loss function

IOU loss is usually used to calculate the matching degree between the network's output bounding boxes and the annotated boxes. It has various advantages such as scale invariance, relatively smooth gradients, and diagonal ratio invariance. During model training, the IOU loss function in object detection can also be used as an important component of the backpropagation algorithm. This experiment mainly introduced the idea of Focal loss to strengthen the learning of difficult-to-distinguish samples. It is found that this idea does not achieve a good effect in the combination of all attention mechanisms, such as the combination of the CBAM attention mechanism even leading to the degradation of model performance. In line with studies focusing on pavement defects, Du et al. endeavored to address the challenge of class imbalance in YOLOv5 by implementing Varifocal Loss as a solution [52]. To expedite convergence and enhance the detection capability of potholes under wet and dry conditions, Wang et al. [53] considered the IOU loss of the YOLOv3 model, which comprises class, confidence, and localization losses. They introduced the CIOU loss, which incorporates aspect ratio considerations, by accounting for the proportion consistency factor between predicted and ground truth boxes. This approach strengthens the model's ability to detect potholes on pavement. Nie et al. [54] provided a detailed explanation of the IOU loss utilized in their road crack detection model based on YOLOv3. By optimizing the loss function, the issue of gradient vanishing can be alleviated, resulting in accelerated model convergence and improved training efficiency [55].

The improvement effect should be related to the characteristics of the distress data, and the pavement distresses are not difficult to distinguish. SIOU was also applied to the YOLOv5 model, but the effect of mAP@0.5 at only 0.92545 was not significant. Considering that the data of the training model are not entirely high-quality samples and that the penalty on the geometric factor is reduced when the prediction frame and the real frame coincide well, this newly proposed loss of WIOU was applied to the model, but the model performance was much worse than before (mAP@0.5:0.92194). It is still necessary to explore the loss of adapting to the characteristics of pavement distress in the actual model application, and the latest ideas introduced into the model improvement cannot always achieve ideal results.

6. Conclusions

This article compares the advanced single-stage object detection models YOLOv5 and YOLOv7, introducing the widely used two-stage object detection algorithm Faster R-CNN for reference and comparison. A new crack detection method based on the YOLOv5 model is proposed, which combines a new attention mechanism and an improved IOU loss function. The experiment involved 7710 data instances, which were partitioned in a ratio of 6:2:2 to form the training, validation, and testing sets, respectively.

The study focused on locating and classifying four types of pavement distresses: longitudinal cracks, transverse cracks, alligator cracks, and potholes. The YOLOv7 model does indeed have better performance than YOLOv5, but in terms of parameter comparisons, YOLOv7 (105.2 GFLOPs) has over six times the number of parameters of YOLOv5 (16.5 GFLOPs). In terms of inference speed, YOLOv7 (49.7 ms) falls significantly behind YOLOv5 (17.7 ms). The experiment involved comparing CBAM, SA, and simAM added at the same positions in the network backbone, using the YOLOv5 model as the baseline. The concept of focal loss was also introduced, and Focal CIOU, Focal EIOU, EIOU, and CIOU were individually tested. It was observed that the combination of Focal CIOU and the SA attention module yielded the best testing performance. As a kind of defect automatic recognition model based on deep learning, the model provides an effective auxiliary means for pavement maintenance. Meanwhile, the experiment further verified the effectiveness of automated pavement detection and different improvement strategies. This advances research in computer vision-based pavement defect detection and enables further exploration of possibilities for optimizing real-time detection models. The experiment included additional early distress images of asphalt and concrete pavement, making it applicable for distress detection before preventive maintenance of newly constructed roads. However, it is limited by the lack of specifically collected data for weather conditions such as rain or snow, which makes it more challenging to recognize distresses in more complex detection scenarios. In the future, there will be a greater emphasis on quantitative research regarding pavement distress, aiming to assess the extent of pavement damage [56] and predict the health condition of pavement [57].

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors appreciate the financial support from Hunan Expressway Group Co. Ltd and the Hunan Department of Transportation (No. 202152) in China. The authors also appreciate the funding support from the project of Beijing High-Level Overseas Talents. Any opinion, finding, and conclusion expressed in this paper are those of the authors and do not necessarily represent the view of any organization.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. K. Wang, Z. Hou, W. Gong, Automation techniques for digital highway data vehicle (DHDV), in *7th International Conference on Managing Pavement Assets*, Citeseer, 2008.
2. S. Zhu, X. Xia, Q. Zhang, K. Belloulata, An image segmentation algorithm in image processing based on threshold segmentation, in *2007 Third International IEEE Conference on Signal-Image technologies and Internet-Based System*, (2007), 673–678. https://doi.org/10.1109/sitis.2007.116
3. S. S. Al-Amri, N. V. Kalyankar, Image segmentation by using threshold techniques, preprint, arXiv:1005.4020. https://doi.org/10.48550/arXiv.1005.4020
4. N. Kanopoulos, N. Vasanthavada, R. L. Baker, Design of an image edge detection filter using the Sobel operator, *IEEE J. Solid-State Circuits*, **23** (1988), 358–367. https://doi.org/10.1109/4.996
5. W. Dong, Z. Shisheng, Color image recognition method based on the prewitt operator, in *2008 International Conference on Computer Science and Software Engineering*, **6** (2008), 170–173. https://doi.org/10.1109/CSSE.2008.567
6. L. Er-Sen, Z. Shu-Long, Z. Bao-shan, Z. Yong, X. Chao-gui, S. Li-hua, An adaptive edge-detection method based on the canny operator, in *2009 International Conference on Environmental Science and Information Application Technology*, **1** (2009), 465–469. https://doi.org/10.1109/ESIAT.2009.49
7. A. Marques, P. L. Correia, Automatic road pavement crack detection using SVM, in *Lisbon, Portugal: Dissertation for the Master of Science Degree in Electrical and Computer Engineering at Instituto Superior Técnico*, 2012.
8. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556
9. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., Going deeper with convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 1–9. https://doi.org/10.1109/CVPR.2015.7298594
10. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
11. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2014), 580–587. https://doi.org/10.1109/cvpr.2014.81

12. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Adv. Neural Inf. Process. Syst.*, (2012), 25. https://doi.org/10.1145/3065386
13. K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, **37** (2015), 1904–1916. https://doi.org/10.18280/ts.370620
14. R. Girshick, Fast R-CNN, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 2380–7504. https://doi.org/10.1109/ICCV.2015.169
15. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *Adv. Neural Inf. Process. Syst.*, (2015), 28. https://doi.org/10.1109/TPAMI.2016.2577031
16. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 2980–2988. https://doi.org/10.1109/TPAMI.2018.2858826
17. J. Redmon, S. K. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 779–788. https://doi.org/10.48550/arXiv.1506.02640
18. C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, preprint, arXiv:2207.02696. https://doi.org/10.48550/arXiv.2207.02696
19. C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, et al., YOLOv6: A single-stage object detection framework for industrial applications, preprint, arXiv:2209.02976. https://doi.org/10.48550/arXiv.2209.02976
20. A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, YOLOv4: Optimal speed and accuracy of object detection, preprint, arXiv:2004.10934. https://doi.org/10.48550/arXiv.2004.10934
21. J. Redmon, A. Farhadi, YOLOv3: An incremental improvement, preprint, arXiv:1804.02767. https://doi.org/10.48550/arXiv.1804.02767
22. J. Redmon, A. Farhadi, YOLO9000: Better, faster, stronger, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 7263–7271. https://doi.org/10.1109/CVPR.2017.690
23. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, et al., Ssd: Single shot multibox detector, in *Computer Vision–ECCV 2016. ECCV 2016. Lecture Notes in Computer Science*, **9905** (2016). https://doi.org/10.1007/978-3-319-46448-0_2
24. A. Womg, M. J. Shafiee, F. Li, B. Chwyl, Tiny SSD: A tiny single-shot detection deep convolutional neural network for real-time embedded object detection, in *2018 15th Conference on Computer and Robot Vision (CRV)*, (2018), 95–101. https://doi.org/10.1109/CRV.2018.00023
25. V. Mandal, L. Uong, Y. Adu-Gyamfi, Automated road crack detection using deep convolutional neural networks, in *2018 IEEE International Conference on Big Data (Big Data)*, (2018), 5212–5215. https://doi.org/10.1109/BigData.2018.8622327
26. S. Dong, J. Zhang, F. Wang, X. Wang, YOLO-pest: a real-time multi-class crop pest detection model, in *International Conference on Computer Application and Information Security (ICCAIS 2021)*, **12260** (2022), 12–18. https://doi.org/10.1117/12.2637467
27. L. Liu, C. Ke, H. Lin, H. Xu, Research on pedestrian detection algorithm based on MobileNet-YOLO, *Comput. Intell. Neurosci.*, **2022** (2022). https://doi.org/10.1155/2022/8924027

28. S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 8759–8768, https://doi.org/10.1109/CVPR.2018.00913
29. M. Tan, Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, preprint, arXiv:1905.11946v2. https://doi.org/10.48550/arXiv.1905.11946
30. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 3–19. https://doi.org/10.48550/arXiv.1807.06521
31. Q. L. Zhang, Y. B. Yang, SA-Net: Shuffle attention for deep convolutional neural networks, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2021), 2235–2239. https://doi.org/10.1109/ICASSP39728.2021.9414568
32. L. Yang, R. Y. Zhang, L. Li, X. Xie, Simam: A simple, parameter-free attention module for convolutional neural networks, in *International Conference on Machine Learning*, (2021), 11863–11874.
33. J. Yu, Y. Jiang, Z. Wang, Z. Cao, T. Huang, Unitbox: An advanced object detection network, in *Proceedings of the 24th ACM International Conference on Multimedia*, (2016), 516–520. https://doi.org/10.1145/2964284.2967274
34. H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 658–666. https://doi.org/10.1109/CVPR.2019.00075
35. Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-IoU loss: Faster and better learning for bounding box regression, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 12993–13000. https://doi.org/10.48550/arXiv.1911.08287
36. Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, et al., Enhancing geometric factors in model learning and inference for object detection and instance segmentation, *IEEE Trans. Cybern.*, **52** (2021), 8574–8586. https://doi.org/10.48550/arXiv.2005.03572
37. Z. Yang, X. Wang, J. Li, EIoU: An improved vehicle detection algorithm based on vehiclenet neural network, in *Journal of Physics: Conference Series*, **1924** (2021), 012001. https://doi.org/10.48550/arXiv.2005.03572
38. H. Zhang, H. Chang, B. Ma, N. Wang, X. Chen, Dynamic R-CNN: Towards high quality object detection via dynamic training, in *Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science*, (2020), 260–275. https://doi.org/10.1007/978-3-030-58555-6_16
39. Z. Liu, X. Gu, H. Yang, L. Wang, Y. Chen, D. Wang, Novel YOLOv3 model with structure and hyperparameter optimization for detection of pavement concealed cracks in GPR images, *IEEE Trans. Intell. Transp. Syst.*, **23** (2022), 22258–22268. https://doi.org/10.1109/TITS.2022.3174626
40. D. Ma, H. Fang, N. Wang, C. Zhang, J. Dong, H. Hu, Automatic detection and counting system for pavement cracks based on PCGAN and YOLO-MF, *IEEE Trans. Intell. Transp. Syst.*, **23** (2022), 22166–22178. https://doi.org/10.1109/TITS.2022.3161960
41. J. Li, C. Yuan, X. Wang, Real-time instance-level detection of asphalt pavement distress combining space-to-depth (SPD) YOLO and omni-scale network (OSNet), *Autom. Constr.*, **155** (2023), 105062. https://doi.org/10.1016/j.autcon.2023.105062
42. Q. Qiu, D. Lau, Real-time detection of cracks in tiled sidewalks using YOLO-based method applied to unmanned aerial vehicle (UAV) images, *Autom. Constr.*, **147** (2023), 104745. https://doi.org/10.1016/j.autcon.2023.104745

43. H. Yao, Y. Liu, H. Lv, J. Huyan, Z. You, Y. Hou, Encoder-decoder with pyramid region attention for pixel-level pavement crack recognition, *Comput.-Aided Civil Infrastruct. Eng.*, 2023. https://doi.org/10.1111/mice.13128
44. R. Li, Y. Wu, Improved YOLO v5 wheat ear detection algorithm based on attention mechanism, *Electronics*, **11** (2022), 1673. https://doi.org/10.3390/electronics11111673
45. J. Sun, H. Ge, Z. Zhang, AS-YOLO: an improved YOLOv4 based on attention mechanism and SqueezeNet for person detection, in *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, **5** (2021), 1451–1456. https://doi.org/10.1109/IAEAC50856.2021.9390855
46. J. Li, H. Wang, Y. Xu, F. Liu, Road object detection of YOLO algorithm with attention mechanism, *Front. Signal Process.*, (2021), 9–16. https://doi.org/10.22606/fsp.2021.51002
47. Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, J. Wang, Ocnet: Object context network for scene parsing, preprint, arXiv:1809.00916. https://doi.org/https://doi.org/10.48550/arXiv.1809.00916
48. Q. Wang, T. Wu, H. Zheng, G. Guo, Hierarchical pyramid diverse attention networks for face recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 8326–8335. https://doi.org/10.1109/CVPR42600.2020.00835
49. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, et al., Attention u-net: Learning where to look for the pancreas, preprint, arXiv:1804.03999. https://doi.org/10.48550/arXiv.1804.03999
50. M. H. Guo, J. X. Cai, Z. N. Liu, T. J. Mu, R. R. Martin, S. M. Hu, Pct: Point cloud transformer, *Comput. Visual Media*, **7** (2021), 187–199. https://doi.org/10.1007/s41095-021-0229-5
51. H. Yao, Y. Liu, X. Li, Z. You, Y. Feng, W. Lu, A detection method for pavement cracks combining object detection and attention mechanism, *IEEE Trans. Intell. Transp. Syst.*, **23** (2022), 22179–22189. https://doi.org/10.1109/TITS.2022.3177210
52. F. J. Du, S. J. Jiao, Improvement of lightweight convolutional neural network model based on YOLO algorithm and its research in pavement defect detection, *Sensors*, **22** (2022), 3537. https://doi.org/10.3390/s22093537
53. D. Wang, Z. Liu, X. Gu, W. Wu, Y. Chen, L. Wang, Automatic detection of pothole distress in asphalt pavement using improved convolutional neural networks, *Remote Sens.*, **14** (2022), 3892. https://doi.org/10.3390/rs14163892
54. M. Nie, C. Wang, Pavement crack detection based on yolo v3, in *2019 2nd International Conference on Safety Produce Informatization (IICSPI)*, (2019), 327–330. https://doi.org/10.1109/IICSPI48186.2019.9095956
55. D. Zhou, J. Fang, X. Song, C. Guan, J. Yin, Y. Dai, et al., IoU loss for 2d/3d object detection, in *2019 International Conference on 3D Vision (3DV)*, (2019), 85–94. https://doi.org/10.1109/3DV.2019.00019
56. C. Han, T. Ma, L. Gu, J. Cao, X. Shi, W. Huang, et al., Asphalt pavement health prediction based on improved transformer network, *IEEE Trans. Intell. Transp. Syst.*, **24** (2022), 4482–4493. https://doi.org/10.1109/TITS.2022.3229326
57. Z. Tong, T. Ma, W. Zhang, J. Huyan, Evidential transformer for pavement distress segmentation, *Comput.-Aided Civil Infrastruct. Eng.*, 2023. https://doi.org/10.1111/mice.13018