Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 9, No. 3, 2024

Research on road Damage Detection Based on Improved YOLOv8

Jingwei Yang, Yuansong Li

Sichuan University of Science & Engineering, School of Computer Science & Engineering, Yibin 644000, China

Abstract: Aiming at the low accuracy of deep learning in road damage detection, a road damage detection method based on improved YOLOv8 is proposed. Firstly, SENetv2 attention module was added to the backbone network of YOLOv8 to improve the model's feature learning ability. Secondly, in the neck network, GSConv is used to replace the common convolutional module to reduce the complexity of the model and improve the accuracy. Finally, the loss function is changed to Wise-iou to reduce the impact of detection due to a small number of low-quality instances. The experimental results show that compared with the traditional YOLOv8, the mAP50 of this model is increased by 2.1 percentage points, and the detection effect is good, which can meet the requirements of accurate detection.

Keywords: Road detection; Deep learning;YOLOv8.

1. Introduction

Road damage detection is of great significance to traffic safety, road damage maintenance and automobile intelligent driving assistance. Traditional manual detection methods often need to close roads to affect traffic, and are restricted by off-site conditions such as weather, so the detection cost is high and the detection efficiency is low. The detection results are affected by subjective factors such as the detection personnel, and the accuracy is low. With the development of science and technology, road detection gradually began to use a large number of fully automated detection technology. The fully automated detection technology far exceeds the traditional detection technology in terms of speed and accuracy, and the detection content is more rich and detailed.

With the rapid development of deep learning, scholars in the field of computer vision have conducted research on pavement disease detection based on deep learning.Zhang et al. [1] collected 500 road images of 3264×2448 pixels taken by smart phones, divided them into 1 million road image blocks of 99×99×3 pixels (RGB color images), and trained 640,000 images by using a neural network containing four convolutional layers. 160,000 for training validation and 200,000 for testing. It is the first time to use deep learning to classify diseased images, and the image blocks can be quickly divided into two types with cracks and without diseases. The recognition accuracy rate is 86.96%, and the recall rate is 92.51%. The classification effect was significantly better than 93-dimensional SVM and Boosting method in control group.Based on YOLOv3, Wang et al.[2] combined low-level features and high-level features to enhance the description ability of the network, and improved the loss function according to the characteristics of transverse and longitudinal crack extension, achieving better results.Wan et al. [3] proposed a lightweight road damage recognition model by enhancing the YOLOv5s method. Their main work is to add ECA attention module to the lightweight model ShuffleNetV2, improve the detection accuracy while lightweight model, and also use BiFPN structure instead of feature pyramid structure to extract more rich feature information. In addition, in order to correct the imbalance of samples, generate higher quality anchor frame. They used Focal-EIOU as a positioning loss, which provides superior performance in both accuracy and efficiency.Cha et al. [4] proposed a concrete crack classifier based on CNN in order to overcome the influence of factors such as light and shadow changes. The classifier network model has 8 layers (4 convolutional layers, 2 pooling layers), and uses SGD algorithm to optimize the network. The output layer uses SoftMax function to classify the output detection results.Nie et al. [5] used the Faster R-CNN target detection model and the transfer learning method with parameter fine-tuning to complete the task of road surface damage detection.By optimizing the loss function of the SSD algorithm and applying hierarchical convolution to pedestrian detection, Yang et al.[6] greatly accelerated the detection speed and improved the accuracy.By optimizing the Faster-RCNN model and combining the advantages of VGG16, ZFNet and Resnet50 networks, Sun et al. [7] proposed an improved Faster-RCNN model based on pavement pot-sealing crack detection method, thus significantly improving the positioning accuracy and the accuracy of test results.Gu [8] et al. proposed an automatic crack detection algorithm, which enhances the fusion of semantic information and multi-channel features, and introduces extended convolution module and attention mechanism to improve the model's ability to extract feature details.Deep learning has shown absolute advantages in the field of image recognition. Deep convolutional neural network can extract high-level semantic features of images and realize automatic recognition of input images without preprocessing input features. Compared with traditional image recognition methods, the effect is better.

This paper proposes a road damage detection method based on improved YOLOv8n:

(1) Based on the traditional YOLOv8n model, SENetv2[9] attention mechanism is introduced in backbone to improve detection accuracy.

(2) GSConv[10] is added to Neck to replace ordinary convolution, reducing the complexity of the model and improving the accuracy.

(3) Use Wise-IoU[11] loss function instead of CIoU function to reduce the impact of a small number of low-quality data instances on detection.

2. Algorithm based on YOLOv8

YOLO series target detection algorithm has been widely used in pavement damage target detection because of its advantages of high accuracy and fast detection speed. At present, YOLO algorithm has been upgraded to YOLOv8, which further optimizes the network structure and enhances the performance of the model. This paper takes YOLOv8n as the base model. In order to improve the detection accuracy of the model, a road surface damage detection algorithm based on improved YOLOv8 is studied

2.1. YOLOv8 Network structure

YOLOv8 is an object detection model composed of four main components: input, backbone, neck and head. YOLOv8 network integrates many advantages of target detection models, retains the CSP idea of YOLOv5, and still adopts the feature fusion methods of FPN-PAN and SPPF. Compared with YOLOv5, there are mainly improvements in the following aspects: First, in order to meet the needs of more scenarios, more size models are designed. A target detection network with a resolution of 640×640 P5 and 1280×1280 P6 is provided. Secondly, we design a C2f (Convolution block) module similar to ELAN (High Efficiency Layer aggregation network) structure, and replace all C3 modules with C2f modules. Two convolutional connection layers in the Neck module were removed. The Head part is changed from the coupling detection head to the structure detection head, using Anchor-Free instead of Anchor-Based. The YOLOv8 structure is shown in the figure.

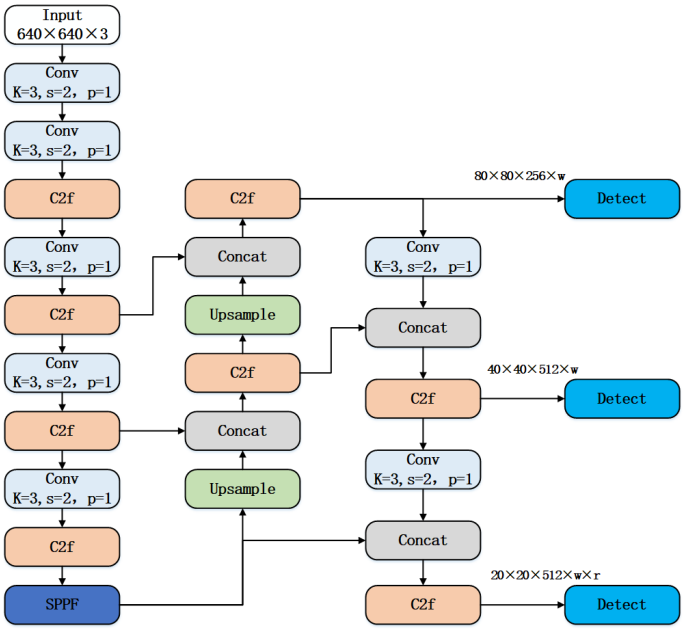


Figure 1. YOLOv8 structure diagram

2.2. Improved algorithm based on YOLOv8

2.2.1. Attention mechanism

Attention mechanisms have been widely used in deep learning convolutional neural networks. By introducing the attention mechanism, the network can obtain the importance of each feature map and generate the corresponding weight. Using the detection results to guide the feature map weights in reverse, the model can put more emphasis on the useful Spaces in the image and focus attention on these areas when processing the task. At the same time, irrelevant features are suppressed, which improves processing efficiency. This approach allows the network to focus more on critical information, improving the efficiency and accuracy of task processing.

SENetv2 is an improved SENet module, which improves the expression capability of the network by introducing Squeeze aggregated excitation (SaE) module. The SaE module combines the two operations of squeezing and excitation, and enhances the learning ability of the network through multiple branches. SENetv2 can greatly improve the detection accuracy and effectively improve the performance of the model while slightly increasing the number of model parameters. The Fig. 2 shows the comparison of the three network modules.

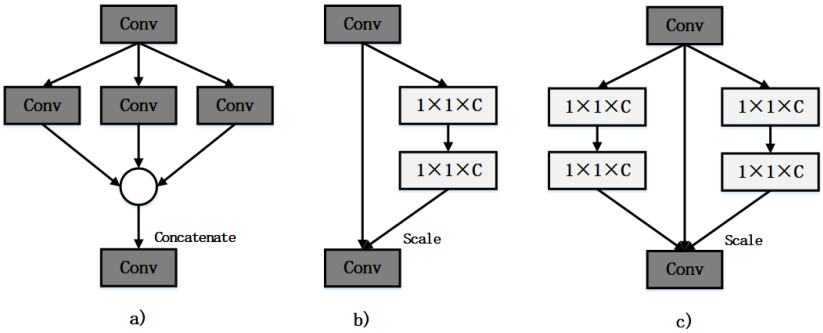


Figure 2. Attention Mechanism Module

a)ResNeXt module: a multi-branch convolutional neural network structure is adopted. The feature graphs of several branches are combined after convolution processing, and finally the combined results are convolved again.

b)SENet module: After the first standard convolution operation, the feature is extruded by global average pooling, then two 1×1 fully connected layers are used to get the channel weight, and finally the feature is scaled.

c)SENetV2 module: Combine the features of ResNeXt module and SENet module. Firstly, squeeze and excite the features through the multi-branch fully connected layer, and finally carry out a feature scaling operation.

SENetV2 is designed to use a multi-branch structure to further enhance feature representation and global information integration.

After extruding the output, it is fed into a fully connected layer with multiple branches, and then stimulated, and the split input is finally passed back to its original shape. SaE modules are designed to enable networks to learn features more efficiently and take into account the interdependencies between different channels during feature transformation. The Fig .3 shows the network structure of SaE.

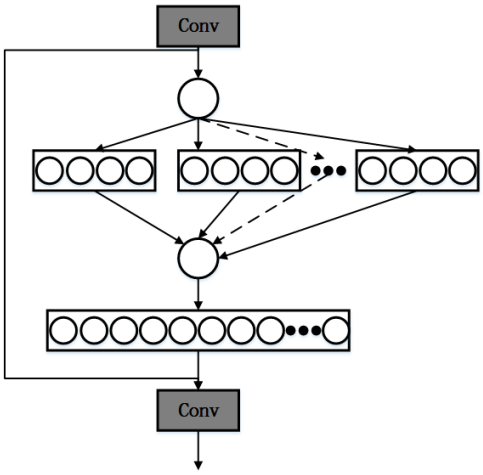


Figure 3. SaE structure

In this paper, SENetV2 is used to add an attention module on the SPPF structure of YOLOv8 backbone network to enhance the feature extraction capability of the backbone network

2.2.2. GSConv

A common convolution module extracts features by doing a dot product mapping of each channel in the input feature graph, where the number of convolution cores is the same as the number of input features. However, when a large number of convolutional layers are stacked, redundant feature graphs will appear, which consumes too much computation and parameter number. The GSConv can ensure the detection performance of the model while reducing the computational load.

In general, in order to improve the inference speed, the images in the convolutional neural network must transmit spatial information step by step to the channel. GSConv preserves the connections between each channel as much as possible at a low cost, protecting the semantic information of the feature map. Using GSConv instead of standard convolution can effectively reduce the computational cost and maintain the learning ability of the model. Then, a GSbottleneck module is constructed on the basis of GSConv, and its structure is shown in Figure 4. GSConv is first subsampled by a conventional convolution and then deep convolution by DWConv. The results processed by the two are concatenated. Finally, the corresponding channels of the two convolution are adjacent by shuffle. Fig .4 shows the GSConv structure.

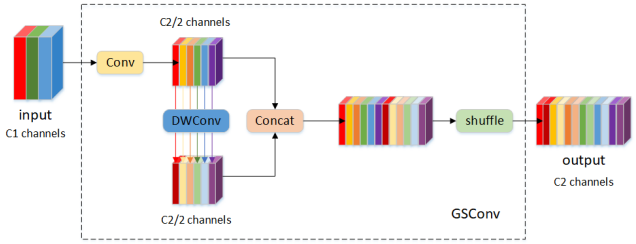


Figure 4. GSConv structure

Assume that the input channel of the module is C1 and the output channel is C2. First, the number of channels is adjusted to C2/2 by standard convolution, and then the convolution can be separated by the depth of convolution kernel size of 5×5 and the number of channels is kept constant. Finally, the feature graphs obtained from the two convolution operations are splicted and mixed. The mixing operation can evenly shuffle the output feature graphs, so that the information generated by standard convolution can penetrate into the information generated by deep separable convolution to enhance the channel information fusion and improve the representation ability of the network.

In this paper, GSConv is introduced into the Neck part of YOLOv8n to replace two common convolution modules. After improvement, the feature fusion capability of the Neck part is enhanced, the receptive field is enlarged, the parameter number and calculation amount are reduced, and the calculation speed of the network is improved.

2.2.3. Wise-IoU

In the process of model training, each data sample will get a predicted value after passing through the model, and the gap between the predicted value and the real value is called the loss value. The loss function is the function used to calculate the gap between the predicted value and the real value, and also serves as the learning criterion for optimizing the model problem. The loss function plays a decisive role in the target detection network model. As a penalty measure, the loss function needs to be continuously minimized during the training process, and ideally match the target prediction frame with the real prediction frame.

Intersection over Union (IOU) is a distance measure used in object detection tasks to measure the overlap between the predicted bounding box and the real bounding box. The calculation formula of IOU is as follows:

$$IOU = \frac{A \cap B}{A \cup B} \tag{1}$$

A represents the predicted bounding box, B represents the true bounding box, A∩B represents the intersection between two boxes, and A ∪ B represents the union between two boxes. The larger the IOU value, the closer the prediction frame is to the real frame, the higher the accuracy of the algorithm, and the better the performance of the model. The IOU calculation

diagram is shown in the figure.

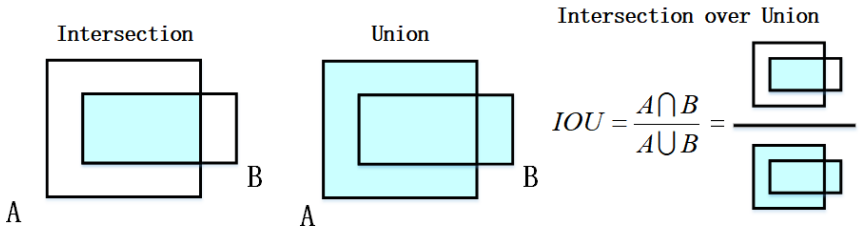


Figure 5. IoU calculation process

When IOU is used as a loss function, its calculation formula is as follows:

$$L_{IOU} = 1 - IOU = 1 - \frac{A \cap B}{A \cup B} \tag{2}$$

In the process of road damage model training, due to the large amount of training data, complex and diverse scenes, and because of the characteristics of road cracks themselves, the shape of target instances is long and varied, and it is inevitable that there will be a small number of low-quality target instances. However, geometric measures such as distance and aspect ratio will magnify the penalty for low-quality examples, which will reduce the generalization performance of the model. The high performance loss function should take into account the penalty of weakening the geometric measure when the anchor frame and the target frame coincide well, so as to maintain the generalization ability of the model. CIOU calculates the boundary frame loss and adds the aspect ratio calculation, but does not consider the balance of the dataset sample itself. Therefore, Wise-IoU is introduced as a bounding box loss function. The formula for Wise-IoU is as follows:

$$L_{WIoU} = \frac{\beta}{\delta \alpha^{\beta - \delta}} R_{WIoU} L_{IoU} \tag{3}$$

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \tag{4}$$

In the formula, α and δ are used as hyperparameters, which are generally set to 1.9 and 3.0. W_g and H_g are the size of the minimum surrounding box, (x, y) and (x_{gt}, y_{gt}) are the coordinates of the center point of the anchor box and the target box, respectively.

β is defined as the outlier, used to describe the mass of the anchor frame, expressed as:

$$\beta = \frac{L_{IoU}^*}{\overline{L_{IoU}}} \in [0, +\infty) \tag{5}$$

In the formula, L_{IoU}^* is the gradient gain of the monotone focusing coefficient, where * represents the dynamic change in the training process according to the specific situation, $\overline{L_{IoU}}$ is the moving average of the momentum m, and the dynamic updating of the normalization factor can keep the gradient gain at a higher level as a whole, solving the problem of slow convergence in the late training period. The formula for calculating momentum m is as follows:

$$m = 1 - \sqrt[tn]{0.05} \tag{6}$$

In the formula, t represents the value of the training rounds epoch, and n represents the value of batchsize during the training process. After T-wheel training, WIoU assigns small gradient gains to low-quality frames to reduce harmful gradients, while focusing on average-quality frames to improve the positioning performance of the model.

3. Experiments and Analysis

3.1. Experimental environment configuration

In this paper, the experimental operating system is Windows10, the graphics card is RTX 2080Ti, and the CPU is Intel Xeon Platinum 8255C CPU@2.50GHz. Pycharm is used as the IDE and Pytorch is used as the deep learning framework in the experimental environment. The Python version is 3.8. In this paper, the stochastic gradient descent (SGD) optimizer was adopted for network model training. The initial learning rate was set to 0.01, the momentum factor was set to 0.937, and the weight attenuation coefficient was set to 0.0005. The total number of training rounds was set to 300 epoch and 32 batchsize. workers is set to 8. During the training of the network model, the input image size was set to 640×640. The data of the training set was input to the network after mosaic data enhancement. All the models were trained and tested on the same equipment.

3.2. Experimental evaluation index

In this paper, Precision, Recall, mAP, parameter number and calculation amount are used as evaluation indexes of the model. Accuracy refers to the correct proportion of all detected objects; Recall rate refers to the proportion of targeted cases correctly identified by the model in all correct cases. The mAP is the average accuracy of multiple categories and is used to evaluate how good the model is on all categories. The calculation formula of evaluation index is as follows:

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$AP = \int_0^1 P(R) \tag{9}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{c} AP_i \tag{10}$$

In the formula, TP represents the correct number of positive samples detected; FP represents the number of positive samples for detecting errors; FN represents the negative sample number of detection errors; n indicates the number of categories.

3.3. Dataset

The data set adopted in this paper is the open data set RDD2022(Road Damage Detection-2022) [12]. It includes 47,420 road images from six countries, including China, India, the United States, Japan, the Czech Republic and Norway, and contains more than 50,000 examples of road damage. The data set mainly divides road damage into four common types of road damage, namely D00(longitudinal cracks), D10(lateral cracks), D20(mesh cracks) and D40(potholes). In addition, the data set also includes some other types of road damage, such as pedestrian crossing blur D43, white line blur D44, etc., but they do not belong to the detection objects in this paper, so after data cleaning, redundant data annotations are removed.

In this paper, YOLO series algorithms are used as the benchmark model. In the process of training, xml annotation files in PASCAL VOC format need to be converted to txt annotation files in YOLO format first. In the original data set, there are some pictures that do not meet the requirements, so it is necessary to filter these pictures.

After data cleaning, this paper randomly selected 11600 pictures from the RDD2022 data set as the data set of this experiment, and divided the training set, verification set and test set according to the ratio of 8:1:1, including 9280 pictures in the training set, 1160 pictures in the verification set and 1160 pictures in the test set.

In order to more intuitively show the different types of road damage examples in the dataset, four data categories are shown, namely (a) longitudinal cracks, (b) transverse cracks, (c) mesh cracks, and (d) potholes.

Figure 6. Road damage examples in the dataset

3.4. Experimental result

3.4.1. Comparative experiment of attention mechanism

The experiment also uses several other attentional mechanism modules, such as SE attention module,CA attention module and CBAM attention module. The experimental results are shown in Table 1.

Table 1. Comparison results of multiple attention mechanisms

Numble	Attention module	mAP50	mAP50-95	Params/M	GFLOPs
1	YOLOv8n	56.4	27.9	3.0	8.1
2	YOLOv8n+SE	56.1	27.5	3.0	8.1
3	YOLOv8+CBAM	56.7	28.1	3.1	8.1
4	YOLOv8+CA	56.9	28.2	3.0	8.1
5	YOLOv8+SENetv2	57.5	28.5	3.0	8.1

As can be seen from Table 1, the addition of SENetv2 attention mechanism greatly improves the accuracy compared with other attention mechanisms

3.4.2. Comparison of ablation experiments

In order to verify the accuracy of the improved algorithm in this paper, several models need to be established to conduct ablation experiments. As shown in Table 2, the improved algorithm adopts a more efficient network structure to improve the network structure of YOLOv8n, and each module plays a role in promoting the model. Through the final experiment, the accuracy is improved while the number of parameters and calculation amount of the model are reduced slightly.

Table 2. Comparison results of ablation experiments

Numble	SENetv2	GSConv	Wise-IoU	mAP50	mAP50-95	Params/M	GFLOPs
1				56.4	27.9	3.0	8.1
2	√			57.5	28.4	3.0	8.1
3		√		57.1	28.2	2.9	8.0
4			√	56.9	28.1	3.0	8.1
5	√	√		58.1	28.6	2.9	8.0
6	√		√	57.2	28.5	3.0	8.1
7	√	√	√	58.5	28.9	2.9	8.0

3.4.3. Visualization of experimental results

The comparison between the improved model in this paper and the YOLOv8n model training process mAP50 is shown in the Fig. 7.

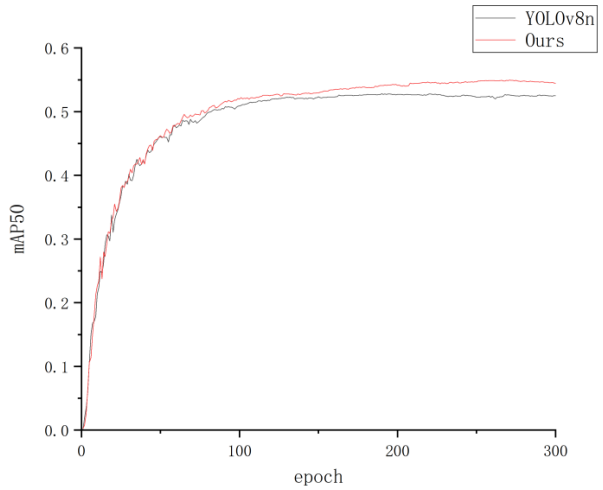
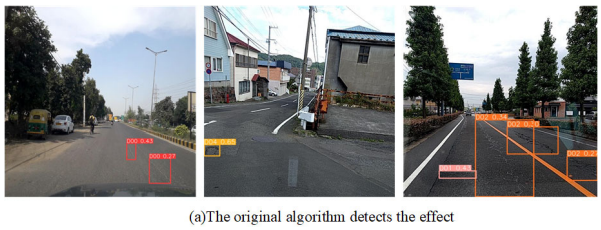


Figure 7. Training process mAP50 comparison

As can be seen from the figure, mAP50 of the two algorithms grows faster in the first 100 rounds; Between 100-200 rounds, the mAP50's growth rate gradually slows down, and between 200-300 rounds, it begins to level off. It can also be seen from the figure that the mAP50 curve of the improved model in this paper is basically above YOLOv8n, and the fluctuation is relatively gentle, indicating that the proposed algorithm is superior to YOLOv8n in road damage detection accuracy and feature learning effect.

In order to more intuitively show the effect of the improved model, Figure 8 is a comparison diagram of the two model detection.It can be seen that the accuracy of the improved model detection has been significantly improved



(a)The original algorithm detects the effect

(b)Improved algorithm detection effect

Figure 8. Comparison figure of the two model detection

4. Conclusions

In this paper, an improved YOLOv8n network model is proposed to solve the problem of low accuracy in road damage detection. The network is optimized by introducing SENetv2 attention mechanism, replacing a part of the Conv in the Neck part of the model with GSConv, and using Wise-IoU loss function instead of the original CIoU loss function. The experimental results show that on the RDD2022 data set, the mAP50 of the improved algorithm in this paper reaches 58.5%, which is 2.1% higher than the original YOLOv8n model. The performance is better than the original YOLOv8n model, which can effectively extract feature information and meet the requirements for accurate detection of road damage targets.

References

[1] Zhang, L., Yang, F., Zhang, Y. D., et al. (2016). Road crack detection using deep convolutional neural network. In 2016 IEEE International Conference on Image Processing (ICIP) (pp. 3708-3712). IEEE.

[2] Wang, Q., Mao, J., Zhai, X., et al. (2021). Improvements of YoloV3 for road damage detection. In Journal of Physics: Conference Series (Vol. 1903, No. 1, p. 012008). IOP Publishing.

[3] Wan, F., Sun, C., He, H., et al. (2022). YOLO-LRDD: a lightweight method for road damage detection based on improved YOLOv5s. EURASIP Journal on Advances in Signal Processing, 2022(1), 98.

[4] Cha, Y. J., Choi, W., & Büyüköztürk, O. (2017). Deep learning-based crack damage detection using convolutional neural networks. Computer-Aided Civil and Infrastructure Engineering, 32(5), 361-378.

[5] Nie, M., & Wang, K. (2018). Pavement distress detection based on transfer learning. In 2018 5th International Conference on Systems and Informatics (ICSAI) (pp. 435-439). IEEE.

[6] Yang, D. M., Zhang, J. G., Xu, S. B., et al. (2018). Real-time pedestrian detection via hierarchical convolutional feature. Multimedia Tools and Applications, 77(19), 25841-25860.

[7] Sun, Z., Pei, L., Li, W., et al. (2020). Pavement sealed crack detection method based on improved Faster R-CNN. Journal of South China University of Technology (Natural Science Edition), 48(02), 84-93.

[8] Gu, S., Li, X., Wang, X., et al. (2021). Crack detection based on enhanced semantic information and multi-channel feature fusion. Computer Engineering and Applications, 57(10), 204-210.

[9] Narayanan, M. (2023). SENetV2: Aggregated dense layer for channelwise and global representations. arXiv preprint arXiv:2311.10807.

[10] Li, H., Li, J., Wei, H., et al. (2022). Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. arXiv preprint arXiv:2206.02424.

[11] Tong, Z., Chen, Y., Xu, Z., et al. (2023). Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. arXiv preprint arXiv:2301.10051.

[12] Arya, D., Maeda, H., Ghosh, S. K., et al. (2022). Crowdsensing-based Road Damage Detection Challenge (CRDDC'2022). In 2022 IEEE International Conference on Big Data (Big Data) (pp. 6378-6386). IEEE.