scientific reports

Check for updates

OPEN

Object detection model design for tiny road surface damage

Chenguang Wu, Min Ye✉, Hongwei Li & Jiale Zhang

Road surface damage detection is crucial in highway maintenance and traffic safety maintenance. However, existing detection methods generally suffer from insufficient generalization capability, poor detection of tiny damage, and difficulty balancing detection accuracy and computational cost. This study proposes a novel road surface damage object detection model (RSDD) to address these challenges. Firstly, a backbone applied to road surface damage feature extraction is designed to solve the problems of feature loss and insufficient extraction of tiny damage during feature extraction. Second, to achieve efficient feature fusion, multiple attention is introduced to optimize features at different stages. Then, a bi-directional feature fusion path is proposed to realize the information exchange between features of different stages, and an enhanced feature pyramid is constructed. Finally, a multi-scale decoupled detection head is adopted to realize the accurate detection of different sizes of damage. Additionally, this study built a road dataset containing rich samples of tiny damage. Extensive comparative experiments are conducted on the collected dataset and a public dataset to validate the generalization performance of RSDD. The experimental results show that RSDD has significant advantages in tiny damage detection while having excellent trade-offs in terms of accuracy, scale, and speed. Specifically, the model achieves 70.8% and 61.2% mAP_{50} on the two datasets with an inference latency of only 4.5 ms under the condition that the number of parameters is 16.5 M. Compared with YOLOv8s, which has a similar number of parameters, RSDD achieves 5.5% and 3.3% improvement in the detection accuracy, respectively, and speeds up the inference by 0.6 ms.

Keywords Road surface damage, Object detection, Tiny damage, Feature extraction, Feature fusion

As a key transportation infrastructure, the road network is of strategic importance for national economic development. However, during the service period, road surface damage is prone to cracks, potholes, and other damages due to the combined effects of vehicle loads and environmental factors, which seriously affect road performance and safety. Therefore, road surface damage detection technology plays an irreplaceable role in road preventive maintenance and safety assurance[1].

Traditional manual inspection methods have limitations of high cost, poor efficiency, and high subjectivity, and the implementation process affects traffic flow and safety hazards[2]. Therefore, the development of intelligent automatic road surface damage detection technology has an important practical significance for improving detection efficiency, reducing maintenance costs, and ensuring traffic safety. Recently, the development of deep learning technology has provided a new technical paradigm for intelligent pavement detection. Currently, deep learning has made breakthroughs in many fields[3–5], and in the field of computer vision, which has formed three core task systems of image classification[6], object detection[7], and semantic segmentation[8]. This study focuses on the problem of object detection in road surface damage detection and proposes a novel network architecture for road surface damage object detection.

Object detection technology can synchronize the localization and recognition function of the target object, and has been widely researched and applied in road surface damage detection. According to the differences in the recognition process, object detection models can be categorized into two-stage models and one-stage models. Among the two-stage models, the most representative ones include classical architectures such as R-CNN[9], Fast R-CNN[10], and Faster R-CNN[11]. Based on these models, researchers have carried out several innovative works. For example, Cha et al.[12] proposed a multi-damage detection method based on Faster R-CNN, which realized the simultaneous detection of five different types of damage. Ju et al.[13] developed a CrackDN network model for the crack detection task, which was improved based on Fast R-CNN. Li et al.[14] utilized the R-CNN model to successfully achieve effective detection of six road surface damages. Sun et al.[15] proposed a pavement crack detection network using different networks such as VGG16[16], ZFNet[17], and Resnet50[18] as the backbone and optimizing the aspect ratio of the crack candidate frame. Hascoet et al.[19] systematically investigated the road

Key Laboratory of Road Construction Technology and Equipment of MOE, Chang'an University, Xi'an 710065, Shaanxi, People's Republic of China. ✉email: mingye@chd.edu.cn

damage detection method of Faster R-CNN and deeply discussed the impact of different depth backbones and multiple testing techniques on the detection performance. However, such algorithms generally suffer from the inherent defect of slow inference speed, which limits their application in practical engineering to some extent.

One-stage detection models effectively overcome the limitation of slow inference speed. Although early one-stage models have a gap in detection accuracy compared with two-stage models, the detection accuracy of one-stage models has improved significantly today. Among them, the YOLO series of models[20–23], as a representative network that combines inference speed and detection accuracy, has attracted wide attention in road surface damage detection. Specifically, Zhang et al.[24] realized road surface damage detection based on UAV-collected data by improving YOLOv3. Xiang et al.[25] innovatively combined Transformer with YOLOv5 to enhance the network model's ability to extract contextual information about cracks. Diao et al.[26], based on YOLOv5, designed a lightweight road damage detection model LE-YOLOv5. Guo et al.[27] used a lightweight backbone instead of YOLOv5's backbone and introduced the attention mechanism. Wang et al.[28] proposed an improved YOLOv8 model, which achieves the dual goals of accuracy enhancement and lightweight by optimizing the network modules. In addition, in various road surface damage detection competitions, numerous researchers[29–32] made innovative improvements based on the YOLO model, which significantly improved the model's detection accuracy of complex road surface damage. These research works fully demonstrate the great potential of one-stage inspection models in practical applications.

Although significant progress has been made in road surface damage detection, due to the wide distribution, large number of damages, and complex detection environment, the existing models are mostly limited to a single scenario, which lacks robustness and generalization ability. Moreover, tiny damages (e.g., cracks 1–2 pixels wide and potholes less than 20 pixels wide) are prone to insufficient feature extraction and feature loss in traditional models, and at the same time, they are susceptible to the interference of noisy features, which leads to high leakage and misdetection rates. In addition, considering the real-time requirements of actual detection scenarios, the computational efficiency of the model still needs to be improved. To address the challenges, a novel road surface damage object detection model (RSDD) is proposed in this study, based on the structural characteristics of road surface damage. Overall, the contributions of this study include:

(1) To address the lack of a large-scale image dataset of road surface damage, this study conducted comprehensive image acquisition of various types of road surfaces in a variety of climatic regions, using multiple camera angles and covering different weather conditions and periods. Through a rigorous data collection and annotation process, a total of 10,440 road surface images were acquired, and the refined annotation of 36,579 road surface damage areas was completed. Compared with the publicly available dataset, the scenarios in this dataset are more complex, including more weather conditions and more occluding interfering objects, and involve more fine-grained and small-scale damage objects, which provides data support for the research of road surface damage detection algorithms.
(2) Aiming at the leakage and misdetection problems in road surface damage detection, especially the challenges of insufficient characterization, feature loss, and noise interference in tiny damage extraction, this study proposes a novel feature extraction backbone based on the structural properties of the damage. Enhancing the model's understanding of the road surface detection environment while improving the extraction of tiny damage features effectively improves the detection accuracy of road surface damage, while reducing the number of parameters and computation of the model.
(3) The variability of feature information at different stages in the feature pyramid generated by the backbone is addressed, as well as the key issues faced in multi-scale damage object detection. We propose a neck network that combines feature selection and feature fusion to achieve more efficient feature fusion, and a multi-scale detection head is used to realize effective detection of multi-size damage. Experiments demonstrate that the method effectively optimizes the detection performance of the model.

Road surface damage detection network

Figure 1 illustrates the general structure of the RSDD, which consists of a backbone, a neck, and a detection head. Among them, the backbone is enhanced by dual scale convolutional downsampling module (DSCD), multi-strategy feature extraction module (MFE), and sensory field enhancement module (SPPF)[33] effectively avoids the problems of feature information loss and feature extraction insufficiency, extracts rich detail information, semantic information and contextual information, and realizes the full understanding of the road environment and damage. The feature fusion neck enhances the information of different stages of the feature maps output from the backbone and enhances the expressive ability of each feature map in the output feature pyramid. Finally, multi-scale detection heads are used on the output feature pyramid for multi-size road damage detection.

Backbone for feature extraction

The backbone consists of three modules, whose structure is shown in Fig. 2, and the detailed implementation is shown in Table 1. Overall, the backbone consists of four stages, stage 1 consists of two DSCD modules and one MFE_SCR module. This stage is used for initial feature extraction to capture primary features in the input image that are essential for subsequent more advanced feature extraction. Stages 2 and 3 have the same structure, both consisting of a DSCD module and an MFE_PCR module. Stage 4 consists of a DSCD module, an MFE_PCR module, and an SPPF module. Stages 2, 3, and 4 utilize a multi-branch-multi-scale structure to capture multiple sensory field features in the input feature maps, allowing the backbone to extract rich detail, contextual, and global information. The input is downsampled at the beginning of each stage, and except for stage 1, which downsamples the input image twice, all other stages downsample the input feature map once, and the corresponding downsampling multiples of each stage are 4, 8, 16, and 32 times, respectively. Moreover, the channels of the feature maps increase sequentially after the completion of downsampling in each stage, and the

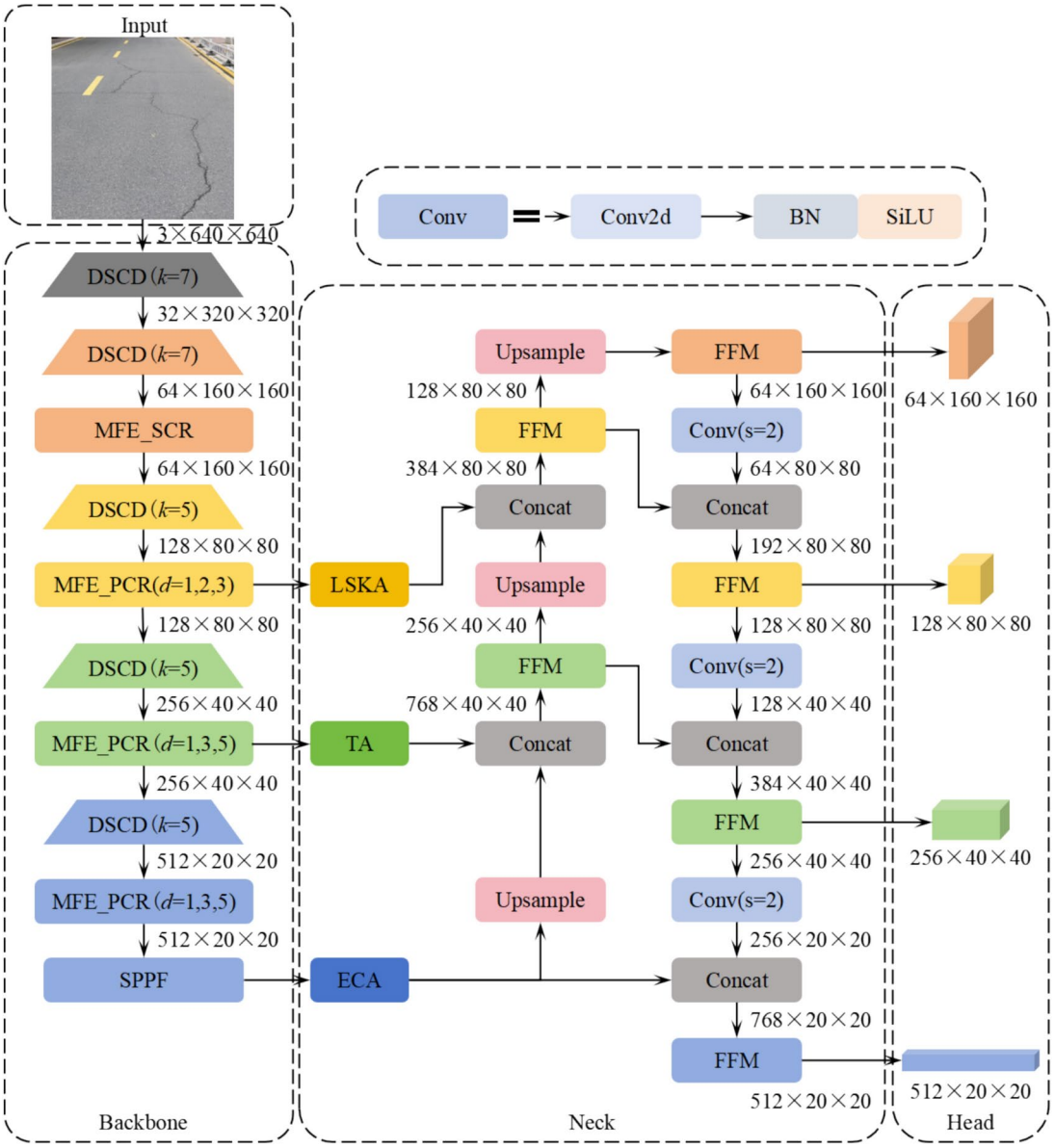


Fig. 1. The structure of RSDD.

channels of the feature maps extracted from the four stages of the backbone are 64, 128, 256, and 512 in that order.

Double scale convolution downsampling module

Downsampling is an essential component of the backbone, and the most widely used downsampling methods are pooling and convolution. Since the widths of cracks are usually smaller, the boundaries of potholes are fuzzy, and there are instances of small sizes, this leads to the fact that the feature information of road damage is easy to be lost during the traditional downsampling operation. Moreover, the distribution of cracks is widely spread, the proportion of large-size potholes in the image is also large, and the sensory field of using a 2×2 pooling or a 3×3 convolution is too small to obtain enough contextual information features, leading to insufficient extraction of features. Convolution using a large convolution kernel can effectively expand the sensory field, but simply utilizing a large kernel convolution can lead to the loss of tiny damage features.

Therefore, this study proposes a series-connected DSCD module, which is applied in down-sampling operation to simultaneously acquire both long-distance-dependent and short-distance-dependent information on road damage features and to efficiently acquire the features of the damage region. As shown in Fig. 3, DSCD adds a series of $k \times k$ depthwise convolution ($s = 1$) to the traditional 3×3 convolution ($s = 2$), where the value of k is greater than 3. The 3×3 convolution can obtain short-distance-dependence information to form the feature maps with a smaller sensory field while down-sampling. The $k \times k$ depthwise convolution uses the convolution of a larger convolution kernel on the feature map with a smaller sensory field to obtain the long-distance-dependence

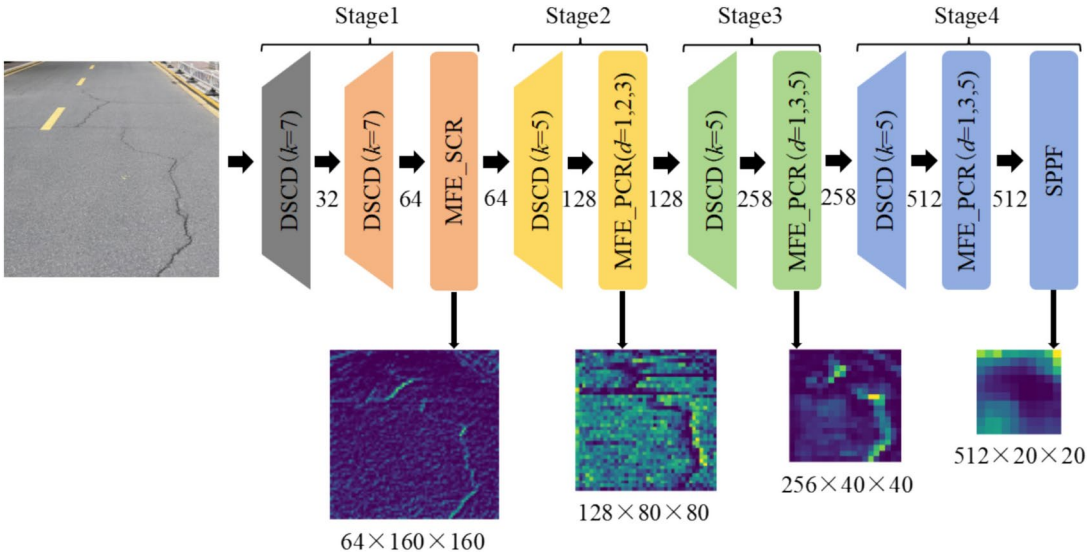


Fig. 2. The overall structure of the backbone.

Stage	Operator module	Input size	Output size	*s*	*k*	*d*	*n*
0	Input image	–	3×640×640	–	–	–	–
1	DSCD	3×640× 640	32×320×320	2	7	–	1
	DSCD	32×320× 320	64×160×160	2	7	–	1
	MFE_SCR	64×160× 160	64×160×160	1	–	–	1
2	DSCD	64×160× 160	128×80×80	2	5	–	1
	MFE_PCR	128×80× 80	128×80×80	1	–	1,2,3	1
3	DSCD	128×80× 80	256×40×40	2	5	–	1
	MFE_PCR	256×40× 40	256×40×40	1	–	1,3,5	1
4	DSCD	256×40× 40	512×20×20	2	5	–	1
	MFE_PCR	512×20× 20	512×20×20	1	–	1,3,5	1
	SPPF	512×20× 20	512×20×20	–	–	–	1

Table 1. Backbone realization details. In the table, "*s*" represents the stride of the first convolution in each module; "*k*" represents the kernel size of the second convolution in DSCD; "*d*" represents the dilation rate of parallel convolution in MFE_PCR; and "*n*" represents the number of modules.

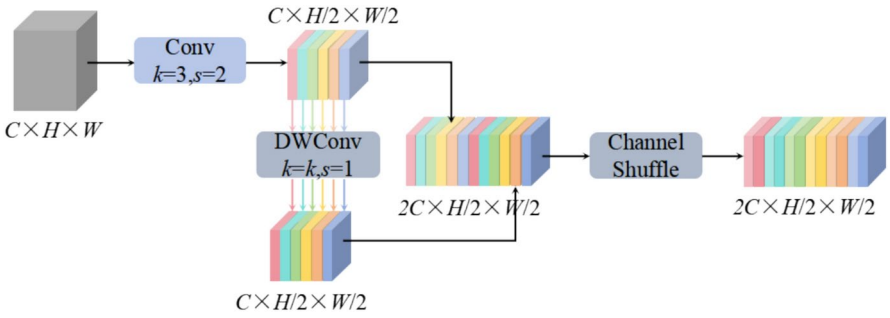


Fig. 3. Structure of DSCD.

information of the road damage to get the feature map with a larger sensory field and does not bring too much computational cost. Then, the different information feature maps are channel connection and the information interactions performed using channel shuffle. In DSCD, $k \times k$ depthwise convolution can effectively expand the sensory field, but the increase in convolution kernel size will lead to an increase in the computational complexity

of the network and the risk of overfitting. Therefore, in this study, an appropriate depthwise convolution kernel size k is selected according to the requirements of the road surface damage detection task.

Multi-strategy feature extraction module

By downsampling, the sensory field can be increased and the calculation cost can be reduced, but the spatial resolution, spatial information, and part of the detail information of the feature map are also reduced. In the road damage detection task, it is difficult to detect the damage accurately due to the high background noise and low contrast of the pavement. Therefore, a feature information extraction module will follow after each stage of downsampling of the network model is required. For the structural characteristics of cracks and potholes, a parallel multi-scale convolution residual module (PCR) is designed. In this module, multiple parallel convolutions are utilized for feature extraction, and an efficient residual feature extraction method is used. This module significantly enhances the capability of the model to extract multi-scale sensory field information, enabling the model to better understand object damage from multiple contextual information.

Figure 4b shows the specific structure of PCR, where an efficient two-step method is used to extract multiple contextual information inside the residuals. First, a 1×1 standard convolutional layer is used to implement the initial feature extraction and generate concise region feature maps. In the second step, 3×3 convolution with multiple dilation rates (d) is used to feature extraction with multiple sensory fields for different region features, respectively. After extracting the multi-scale sensory field features, all the features are channel-connected and subjected to a 1×1 pointwise convolution process to realize the information interaction in the channel direction and generate the final residuals. Finally, a more robust and comprehensive output feature is constructed by adding residuals to the input, which contains a rich set of detailed and contextual features.

However, the over-reliance on large-scale sensory fields to obtain more contextual information exceeds the practical requirements. Therefore, it is necessary to reasonably design the sensory fields of each stage module to optimize the feature extraction efficiency. To this end, this study synchronously designs the serial convolution residual module (SCR), whose structure, shown in Fig. 4c, the module utilizes a single branch serial convolution structure. Specifically, it uses two 3×3 convolutions for feature extraction and generates residuals, which are then combined with the input. Compared with the PCR module, these two modules have different sizes of sensory fields and are applied to the shallow and deep layers of the backbone, respectively.

In addition, to realize the lightweight of the network model, this study, from the perspective of network structure design, divides the input features into two for different processing. Figure 4a demonstrates the specific structure of MFE, firstly, a channel split is used to divide the input equally into two features of the same size, half of which is subjected to feature extraction by PCR or SCR, and then the extracted feature maps are concatenated with the original input in the channel direction. Finally, a 1×1 convolution is used to interact with the channel information and realize the compression of the channel dimension. This method significantly improves the

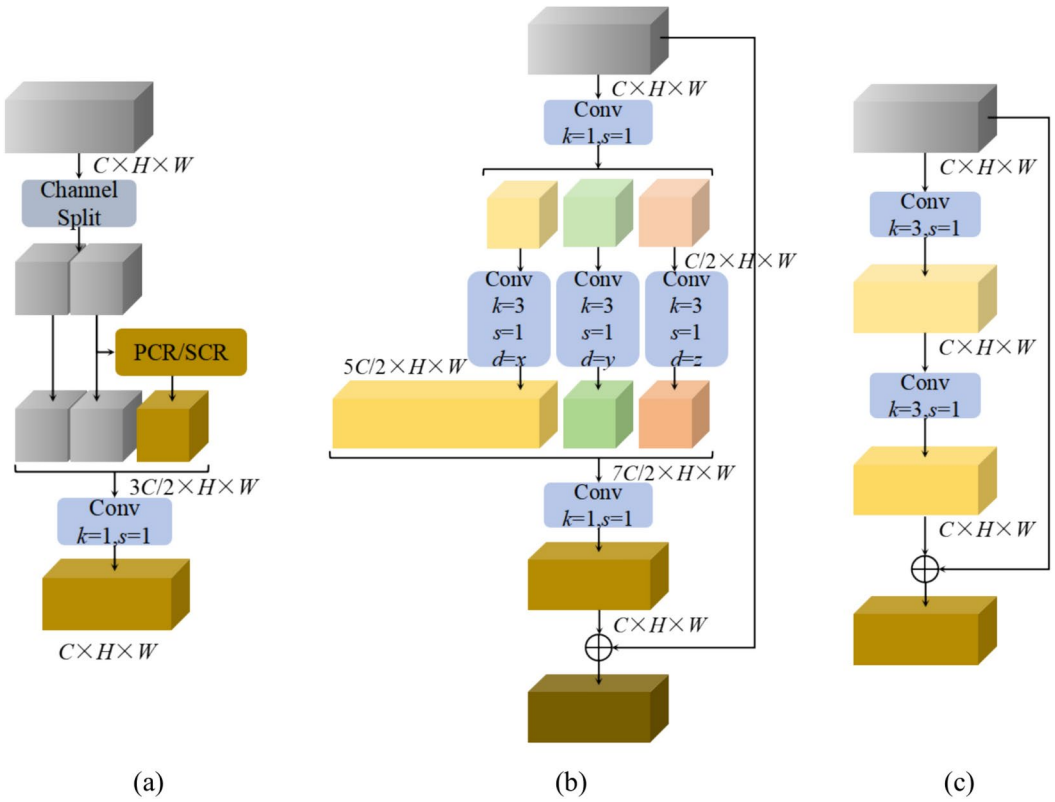


Fig. 4. Structure of the multi-strategy feature extraction module. (**a**) MFE; (**b**) PCR; (**c**) SCR.

efficiency of feature extraction and speeds up model training and inference. The MFE is named MFE_PCR and MFE_SCR according to the feature extraction module.

At the shallow stage of the backbone, the input feature maps are larger and contain more detailed information. At this point, the feature extraction of MFE_SCR is more effective than MFE_PCR. As the network deepens, the resolution of the extracted features will decrease step by step, and the feature information also changes from detail information to semantic information. To ensure efficient feature extraction efficiency, PCR with different dilation rates is used at other stages of the backbone to extract feature information. Deeper networks are also more receptive to larger sensory fields, as semantic enhancement allows the convolution to make larger spatial connections. Therefore, MFE_SCR is used in stage 1 of backbone and MFE_PCR is used for feature extraction in the last three stages of the network. Where the dilation rate (x, y, z) of PCR is 1, 2, 3 in stage 2. In stages 3 and 4, the dilation rate of PCR is 1, 3, 5 respectively. After research, it has been found that it is difficult for convolution to directly establish connections between spatially distant information, and long-distance connections need the help of short-distance connections, and small-size sensory fields are important in MFE_PCR. Therefore, the convolution branch in MFE_PCR with a dilation rate of 1 has five times as many output channels as the other branches.

Sensory field enhancement module

To extract the global information, the sensory field enhancement module uses SPPF, whose specific structure is illustrated in Fig. 5. Firstly, the feature channel is compressed by a 1×1 convolution, which is usually used with a compression ratio of 2. The compressed feature maps are then subjected to three series of pooling operations with the same stride, and each pooling operation extracts feature information that is different from the sensory field. Finally, the pooling results of different sensory fields are fused by channel concatenation and convolution operations to avoid information conflicts caused by simple splicing to aggregate feature representations at different scales. The SPPF effectively improves the global information extraction ability of the backbone, enabling the RSDD model to provide a more comprehensive and deeper understanding of the pavement environment and road surface damage.

Feature fusion neck

After the backbone extracts the features of different scales, it is also necessary to construct a feature fusion neck to realize feature enhancement. The neck includes a feature selection module (FSM) and a feature fusion module (FFM), in which the FSM applies different attention mechanisms to select the different scale features extracted by the backbone, and the FFM mainly realizes the information interaction between the different scale features through the operations of downsampling, upsampling, channel connection, and feature fusion basic block (FFB). In this study, the features C2, C3, and C4 extracted from stage 2, stage 3, and stage 4 of the backbone are utilized for selection and fusion, and the final output is a feature pyramid containing several scales.

Feature selection module

To achieve more efficient feature fusion, C2, C3, and C4 features are selected through FSM respectively. By selecting and extracting more significant road surface damage features and applying them to the feature fusion process, the interference information in the background can be effectively suppressed, reducing the impact of background noise on the accuracy of road surface damage detection. The FSM in this study is realized by multiple attention mechanisms, and the attention used in different stages of the backbone varies because the feature maps extracted in different stages focus on different information.

Channel attention C4 as the deepest feature has the smallest resolution and the richest semantic information, in which the features on each channel can be regarded as a specific category of responses, each of which affects the final semantic prediction to a different extent. Therefore, efficient channel attention (ECA)[34] is employed to

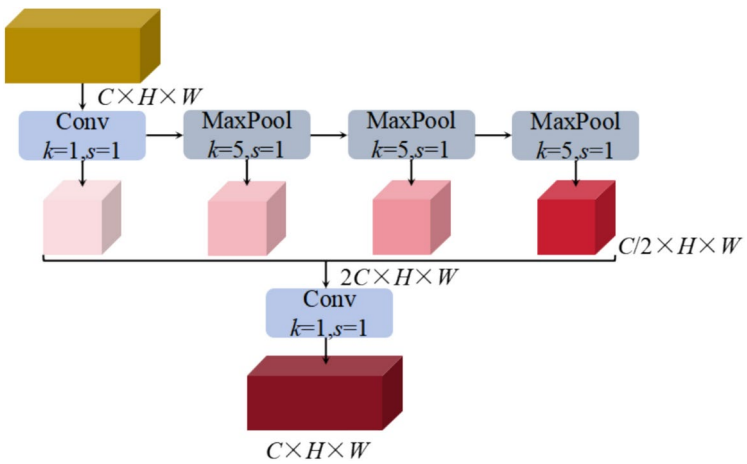


Fig. 5. Structure of the SPPF.

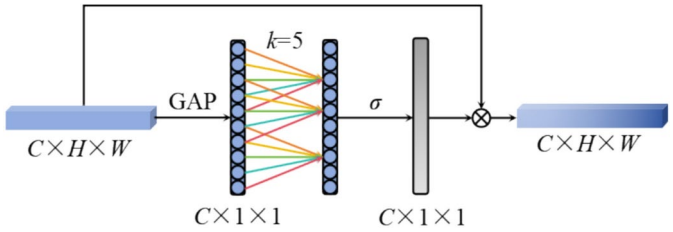


Fig. 6. Structure of the ECA.

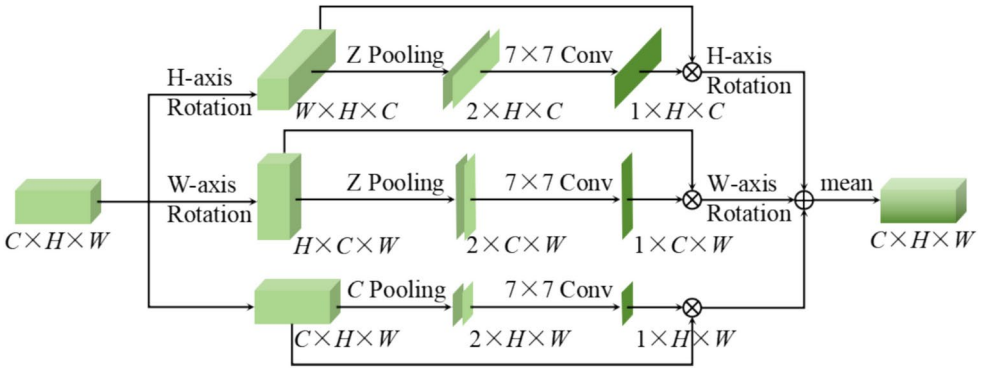


Fig. 7. Structure of the TA.

efficiently capture the relationships among the channels in a feature, thus enhancing the characterization of C4. As shown in Fig. 6, ECA obtains the global average of each channel through global average pooling (GAP), and then generates the channel weights through a 1×1 1D convolutional layer. The k in the figure represents the coverage of cross-channel interactions, and k is adapted indeed by the mapping of the channel dimensions. Finally, the features are normalized by a scaling factor (σ) to generate weights to be applied to each channel of the input, enabling the weighting of features from different channels.

Mixed attention C3 as an intermediate feature contains both considerable semantic and detail information, which can be efficiently selected by the TripleAttention (TA)[35] for both channel features and spatial features. Z-pooling in Fig. 7 reduces the tensor in the channel dimension to two dimensions, performs average pooling and maximum pooling on the tensor in the channel direction, respectively, and connects the obtained features in the channel direction. This operation preserves the rich representation of the actual tensor while reducing the depth and lightening the calculation. As shown in Fig. 7, input a $C \times H \times W$ feature map. In a branch of the TA, establish the relationship between dimension H and dimension C. First, the feature map is rotated 90° counterclockwise on the H-axis, and the shape of the rotated feature map is changed to $W \times H \times C$. Then a $2 \times H \times C$ feature is obtained by Z-pooling, subsequently, a $1 \times H \times C$ attention weight is generated by $k \times k$ standard convolution, normalization process, and activation function. Finally, the weights are weighted with the input and rotated 90° clockwise on the H-axis. In another branch, the relationship between dimension C and dimension W is established, the input is rotated 90° counterclockwise on the W-axis and then processed by Z-pooling, standard convolution, normalization, and activation function to obtain a $1 \times C \times W$ attentional weight, which is weighted to the input, and then it is rotated 90° clockwise along the W-axis. In the third branch of TA, an information interaction between the H and W dimensions is established.

Spatial attention C2 contains rich detail information, and through fully utilizing the shallow detail information can achieve effective fine-grained crack and small-scale pothole detection. However, only limited context information near the damage object is often considered in C2, and without reference to sufficient long-range context, it is possible that fine-grained but large-distributed damage objects such as cracks and fuzzy-boundary damage objects such as potholes may be incorrectly detected. Moreover, the required long-range contexts may vary from one damage to another, requiring different scales of sensory fields to be accurately localized and identified. This study addresses this problem through the large selective kernel attention (LSKA) mechanism[36], which consists of a large kernel convolution and a space kernel selection mechanism, as shown in Fig. 8. The large convolution kernels and the expansion rate ensure that the sensory field is sufficiently expanded, and multiple feature maps of the large sensory field are acquired by successive convolutions in LSKA. The space kernel selection mechanism automatically and adaptively assigns convolutional kernels to different objects based on multi-scale features, aiming to enhance the network's ability to capture the space contextual information associated with damaged objects. The mechanism uses a spatial selectivity strategy to filter out the most discriminative spatial regions from the feature maps generated by the multi-scale large convolutional kernel. Specifically, the feature maps of

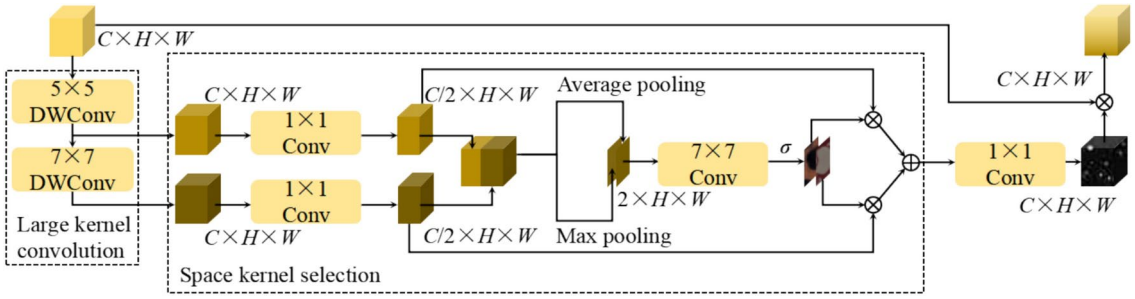


Fig. 8. Structure of the LSKA.

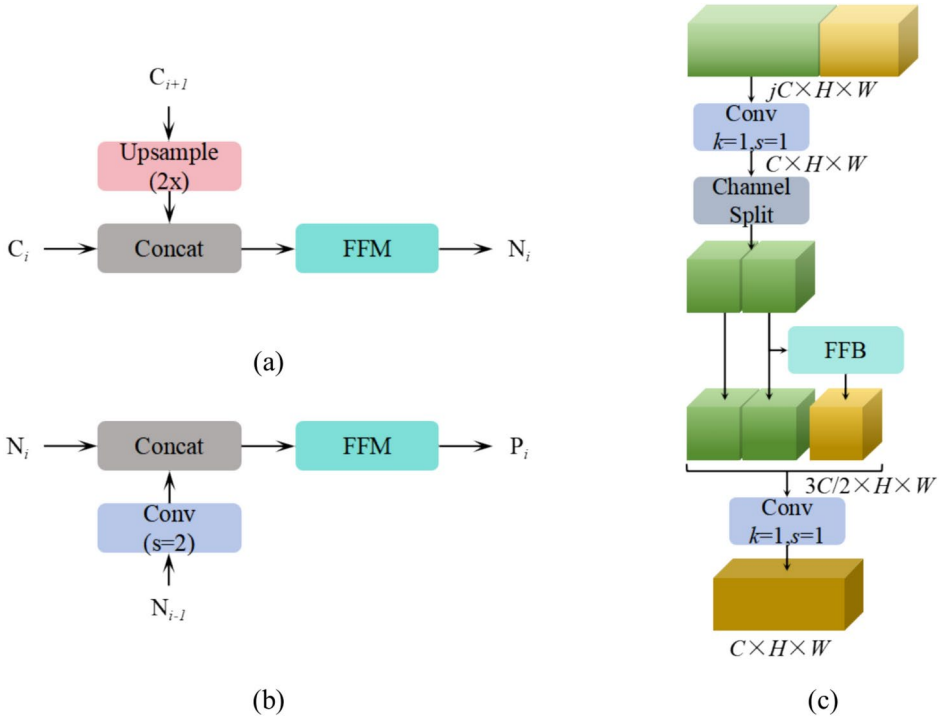


Fig. 9. Structure of the FFM. (**a**) Top-down feature fusion; (**b**) Bottom–up feature fusion; (**c**)FFM.

different sensory fields are first channel-concatenated, and then average pooling and maximum pooling in the channel directions are applied to extract spatial relationships efficiently. To realize the information exchange between different spatial descriptive information, the pooled spatial ensemble features are concatenated in the channel direction, and the ensemble features are transformed into separate spatial attention maps of different sensory field feature maps using convolutional layers and sigmoid activation functions. Then, they are weighted with the feature maps of different sensory fields and fused by point-wise summation to obtain features containing spatial attention. Finally, the initial input is elementwise multiplied with the spatial attention features to get the final output.

Feature fusion module

The features after feature selection have better characterization ability, but features with different resolutions have different information, to achieve an excellent detection effect, it is necessary to fuse and communicate the information of different features. However, there are semantic gaps between feature maps at different scales, and a simple fusion approach achieves very limited improvements. Therefore, the main body of feature fusion in this study adopts both top-down and bottom-up feature fusion architectures, as illustrated in Fig. 9, where high-level semantic information is first propagated into shallow features through the top-down feature fusion path to enhance the type recognition ability of all features. Then, the detailed information such as texture and edges is transferred to the deeper features through the bottom-up feature fusion path, which further enhances the localization ability and the detection ability of tiny damages in the whole feature hierarchical structure. In the process of information transfer, FFB is very important, which determines the effect of feature information fusion, for the study of this part of the structure we experimented with six structures as in Fig. 10. The structure d was

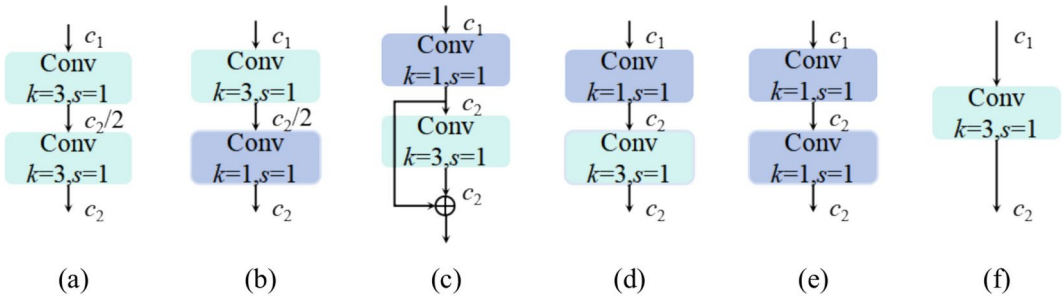


Fig. 10. FFB for different structures.

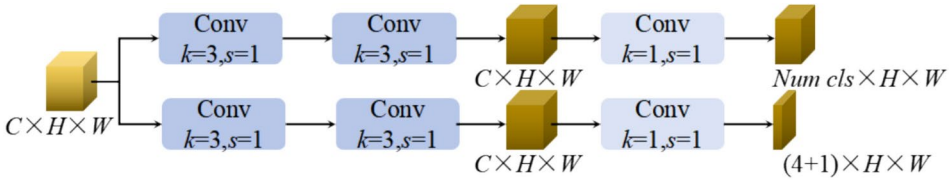


Fig. 11. Structure of the decoupled detection head.

finally determined as the final FFB, which first goes through a 1×1 convolution layer for channel information interaction, and then through a 3×3 convolution so that different information can be further communicated.

Multi-scale decoupled detection head

The detection head in this study performs detection in a decoupled manner so that the classification and regression tasks can focus on their respective features separately, effectively improving the performance of road surface damage detection. As shown in Fig. 11, two independent branches are designed to perform classification and regression separately, and the respective information is extracted by two 3×3 convolutions and one 1×1 convolution on the same output feature map, and finally the regression loss function and classification loss function are calculated separately. This design allows the model to better learn the features and laws of different tasks improve its ability to adapt to complex scenarios, and enhance the flexibility of the model during training and optimization. However, while decoupled detection heads separate tasks, they still share some features during the early stages of feature extraction to balance task independence and feature utilization efficiency. In addition, we perform detection on multiple scales of feature maps output from the neck to realize the detection of different sizes of road surface damage.

Dataset and evaluation indicators

Dataset

In this study, we use our own collected dataset for model training, as shown in Fig. 12, which contains road images from nine cities in China and Cambodia, involving highways, urban roads, and township roads. It includes a total of 10,440 road images covering diverse shooting angles, weather conditions, lighting environments, and pavement types, and is labeled with 36,579 instances of road surface damage. Figure 13 illustrates the manual annotation process of the collected road images using the LabelImg image annotation software. First, the target is selected by drawing a bounding box, and then the selected target is classified. The dataset records a total of four types of road surface damage: longitudinal cracks (D00), transverse cracks (D10), alligator cracks (D20), and potholes (D40). Compared to the publicly available dataset, the scenarios in this dataset are more complex, richer in interferences, and involve more tiny damage. Figure 14 summarizes the distribution of various detection objects in the dataset, and the final dataset is grouped into training, validation, and testing sets with a ratio of 7:2:1.

Evaluation indicators

The evaluation indicators for the object detection model mainly include F1 score, average precision (AP), number of parameters (Params), and computational complexity (GFLOPs). In addition, precision (P) and recall (R) as the basic indicators can also be used as the basis for the evaluation of the model, and the F1 score and AP calculated by P and R are the final indicators of the accuracy of the model detection.

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

Fig. 12. Partial presentation of this study's dataset.

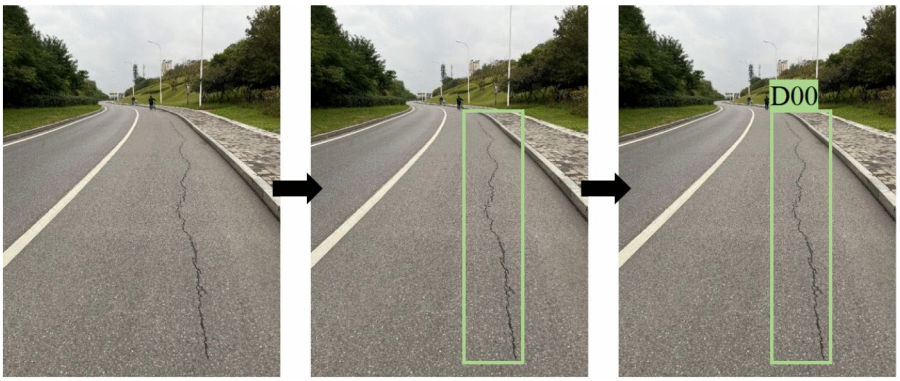


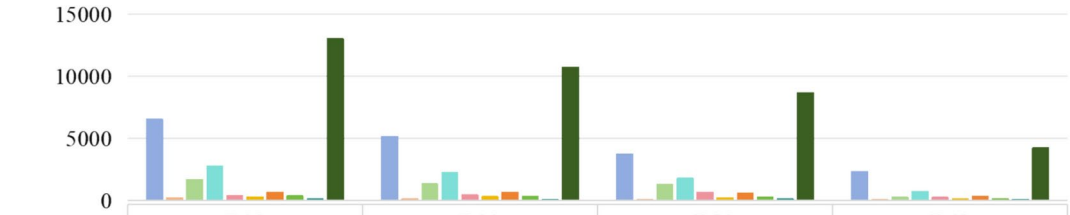
Fig. 13. Labeling process of road surface damage region.

$$\mathrm{AP} = \int_0^1 \mathrm{P}(\mathrm{R}) d\mathrm{R} \tag{3}$$

$$\mathrm{mAP} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{AP} \tag{4}$$

$$\mathrm{F1} = 2\frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{5}$$

where TP denotes that the prediction result is assigned as a positive sample and the true value is a positive sample, i.e., the number of positive samples that were correctly classified. FP denotes that the prediction result is



	D00	D10	D20	D40
Xi'an	6553	5134	3732	2327
Yan'an	189	120	77	80
Lan'zhou	1692	1321	1278	256
Bai'yin	2756	2276	1776	721
Hai'kou	377	450	667	236
San'ya	279	295	220	161
Wan'ning	656	678	612	314
Shang'hai	363	340	250	120
Cambodia	124	86	47	16
Total	12989	10700	8659	4231

Fig. 14. Number of examples of each type of damage in each city.

a positive sample but the true value is a negative sample, i.e., the number of negative samples that were incorrectly classified. FN denotes that the prediction result is a negative sample but the true value is a positive sample, i.e., the number of positive samples that were incorrectly classified. P is the percentage of correctly predicted positive samples out of all positively predicted samples. R is the percentage of positively predicted samples out of all positively predicted samples. AP denotes the area under the P-R curve. mAP denotes the average AP value of the *n* categories. the F1 scores are a combination of P and R, and are used as a measure of the overall performance and stability of the model.

Model training

In this study, the deep learning framework PyTorch is utilized as the network framework, and the software environment is Windows 10, Python3.9, CUDA11.3, and cuDNN8.2.1.32, and the hardware environment has an Intel i7 10700k as the CPU and an NVIDIA GeForce RTX3060 (12 GB RAM) as the GPU. During training, the input image is set to 640 × 640 and SGD is used as the optimizer for model training. The model training elapsed time is set to 300, the batch size is 16, and the initial learning rate is 0.01. Moreover, data enhancement techniques such as Mosaic data enhancement, random cropping, flipping, rotating, scaling, and brightness adjustment are applied to the training set to enhance the model's generalization ability, prevent overfitting, and improve its adaptability to various scenarios.

Results

Research on DSCD

In this experiment, the value of *k* in the DSCD was investigated and the experiment in Table 2 was designed, where the value of *k* in the DSCD was changed while the other units of the network model remained unchanged. In this case, YOLOv8s was used as the base model (model 0) and DSCD was used as the downsampling module in the backbone to replace the original downsampling. Different *k*-values were used in different stages of the network, which are given in the table; in the first stage there are two downsampling operations, so two corresponding *k*-values were set in stage 1. By comparison, it is found that model 8 has the best detection result when the *k* value of the first two DSCDs in the backbone is 7 and the *k* value of the last three DSCDs is 5. The average precision mAP_{50} and $mAP_{50:95}$ are 66.4% and 35.4%, respectively, and the F1 score reaches 66.5. From the analysis, it can be found that when the size of the input feature map is large, the *k*-value in DSCD should be larger, but not more than 7. When the size of the input feature map is small, the *k*-value in DSCD should be smaller, but not less than 5.

Research on MFE

This experiment identified the feature extraction strategies for the different stages in the backbone. The experiments in Table 3 show that when only MFE_PCR is used as the feature extraction module, it does not allow the network model to achieve the best detection results. In the first two stages of the network, the convolution using a sensory field that is too large is far beyond the sensory field size requirement. In the last two stages of the backbone, the acceptance of a larger feeling field is enhanced as the depth of the network deepens. Ultimately, the model is optimized for road surface damage detection when MFE_SCR is used in stage 1 of the backbone, MFE_PCR with dilation rates of 1, 2, and 3 in stage 2, and MFE_PCR with dilation rates of 1, 3, and 5 in stages 3 and 4, at which time the average accuracies of mAP_{50}, $mAP50_{50:95}$, and F1 are 66.1%, 35.0%, and 66.5, respectively.

Model No	The *k*-value in DSCD				P(%)	R(%)	mAP_{50}(%)	$mAP_{50:95}$(%)	F1
	Stage 1	Stage 2	Stage 3	Stage 4					
0	–	–	–	–	69.7	61.3	65.3	34.6	65.2
1	5,5	5	5	5	69.8	61.4	65.5	34.8	65.3
2	3,5	5	5	5	70.6	60.8	65.4	34.5	65.3
3	3,3	5	5	5	71.2	61.1	65.6	34.5	65.8
4	3,3	3	5	5	70.8	60.7	65.2	34.6	65.4
5	3,3	7	7	7	69.8	61.3	64.9	34.4	65.3
6	7,7	7	7	7	71.6	60.8	65.7	34.9	65.8
7	7,7	7	5	5	71.7	60.1	65.2	34.3	65.4
8	7,7	5	5	5	72.1	61.7	66.4	35.4	66.5
9	7,7	5	5	3	71.2	61.5	65.7	35.1	66.0
10	7,5	5	5	3	71.7	61.6	65.9	35.0	66.3
11	9,7	5	5	5	71.0	61.0	65.9	34.9	65.6

Table 2. The influence of *k*-value on model detection accuracy in different stages of the DSCD.

Multi-strategy feature extraction				P(%)	R(%)	mAP_{50}(%)	$mAP_{50:95}$(%)	F1
Stage 1	Stage 2	Stage 3	Stage 4					
MFE_PCR (1,3,5)	MFE_PCR (1,3,5)	MFE_PCR (1,3,5)	MFE_PCR (1,3,5)	71.9	59.7	65.2	34.5	65.2
MFE_PCR (1,5,9)	MFE_PCR (1,3,5)	MFE_PCR (1,3,5)	MFE_PCR (1,3,5)	70.8	60.6	65.2	34.4	65.3
MFE_PCR (1,3,5)	MFE_PCR (1,2,3)	MFE_PCR (1,3,5)	MFE_PCR (1,3,5)	69.5	60.9	65.0	34.4	64.9
MFE_PCR (1,2,3)	MFE_PCR (1,2,3)	MFE_PCR (1,3,5)	MFE_PCR (1,3,5)	70.9	61.5	65.6	34.5	65.9
MFE_PCR (1,2,3)	MFE_PCR (1,2,3)	MFE_PCR (1,2,3)	MFE_PCR (1,2,3)	70.7	60.0	64.9	34.0	64.9
MFE_PCR (1,2,3)	MFE_PCR (1,2,3)	MFE_PCR (1,2,3)	MFE_PCR (1,3,5)	69.8	61.3	64.7	34.2	65.3
MFE_SCR	MFE_PCR (1,2,3)	MFE_PCR (1,3,5)	MFE_PCR (1,3,5)	71.5	62.1	66.1	35.0	66.5
MFE_SCR	MFE_SCR	MFE_PCR (1,3,5)	MFE_PCR (1,3,5)	69.7	61.0	65.1	34.6	65.1
MFE_SCR	MFE_SCR	MFE_SCR	MFE_SCR	68.8	59.2	63.5	33.7	63.6

Table 3. The influence of different feature extraction strategies at each stage on model detection accuracy. In the table, "MFE_PCR (1,5,9)" represents the dilation rate of the parallel convolutional layers in MFE_PCR is 1,5,9 respectively, and "MFE_PCR (1,2,3)" represents the dilation rate is 1,2,3 respectively.

After the feature extraction strategies for each stage of the backbone were determined, the hyperparameters in the MFE_SCR and MFE_PCR modules also needed to be analyzed. It includes whether residual connections are used in MFE_SCR and MFE_PCR, the ratio of the number of output channels of the parallel convolutional layers in MFE_PCR, and the input–output ratio of the first 1×1 convolution in MFE_PCR. As shown in Table 4, where Model 1 and Model 2 identify residual connections that contribute to the detection performance of the model. Models 2 through 7 specify the ratio of the number of output channels of the parallel convolution in MFE_PCR, and since it is relatively difficult for the convolution to establish information connections directly over a large space, and long-distance connections require the help of short-distance connections, a small sensory field is important at all stages, the model performs best when the number of convolutional output channels with a dilation rate of 1 is five times the number of output channels of the other concurrent convolutional layers. Models 6, 8, and 9 determine the input–output ratio of the first 1×1 convolution in MFE_PCR, and the model has the best detection when its ratio is 2:1. With this experiment, we determined the specific structure of the MFE.

Research on FFB

In the feature fusion stage feature maps containing different information are fused through top-down and bottom-up structures, in this experiment we analyzed the effect of different feature fusion base blocks on road surface damage detection performance. Firstly, we used Darknet53 as a backbone to extract road surface image features, generate multi-scale feature maps through feature fusion, and finally detect and localize the damage on the output three-scale feature maps. In the experiments of Table 5, the specific structure of the FFB is discussed, and a variety of FFBs have been introduced in the above paper, it is found by comparison that the best detection

Model No	Residual connection	MFE_PCR		P(%)	R(%)	mAP_{50}(%)	$mAP_{50:95}$(%)	F1
		c_1:c_2:c_3	c_{in}:c_{out}					
1		2:1:1	2:1	71.2	61.6	65.8	34.9	66.1
2	✓	2:1:1	2:1	71.5	62.1	66.1	35.0	66.5
3	✓	1:1:1	2:1	71.6	61.2	65.6	34.6	66.0
4	✓	3:1:1	2:1	71.4	62.0	66.1	35.3	66.4
5	✓	4:1:1	2:1	71.4	61.8	66.2	35.1	66.3
6	✓	5:1:1	2:1	71.4	62.9	66.5	35.4	66.9
7	✓	6:1:1	2:1	70.3	63.3	66.1	35.1	66.6
8	✓	5:1:1	3:1	69.8	61.1	65.3	34.6	65.2
9	✓	5:1:1	3:2	72.1	61.9	66.1	35.3	66.6

Table 4. The influence of hyperparameters in MFE on model detection accuracy. In the table, "✓" indicates the use of residual connections in MFE_SCR and MFE_PCR, "c_1:c_2:c_3" represents the ratio of the number of output channels of the parallel convolutional layer in MFE_PCR, and "c_{in}:c_{out}" represents the input–output ratio of the first 1×1 convolution in MFE_PCR.

FFB	P(%)	R(%)	mAP_{50}(%)	$mAP_{50:95}$ (%)	F1	Params(M)	GFLOPs
a	69.7	61.3	65.3	34.9	65.2	21.4	28.4
b	68.0	59.0	63.6	33.7	63.2	19.8	26.8
c	69.6	61.0	65.1	34.6	65.0	19.8	26.8
d	71.5	60.4	65.8	35.0	65.5	19.8	26.8
e	69.2	59.1	63.9	33.5	63.8	18.3	25.1
f	71.0	59.2	64.1	33.8	64.6	19.6	26.6

Table 5. The influence of different FFB structures on the evaluation indicators of model performance. The structures corresponding to the FFBs in the table have been given above.

accuracy is achieved when the structure of the model is d. This structure can effectively realize the fusion of different feature information and improve the model's identification and localization accuracy of the damage.

Research on FSM

The FSM in this study mainly selects and enhances the feature maps extracted at different stages in the backbone through different attention, and this part of the study analyzed the effect of different attention on different information features. Experimental results are shown in Table 6, for deep features C4, both channel attention ECA and CAM can effectively enhance the accuracy of the model, and the characterization ability of the feature map can be enhanced by assigning different weights to different channels so that the model focuses on the features of the more important channels. For the intermediate feature C3 that contains both semantic and detailed information, TripleAttention can effectively enhance the characterization ability of C3 by cross-dimensional information interaction that captures the information exchange between spatial and channel dimensions. In contrast, the use of spatial or channel attention alone, as well as the simple combination of spatial and channel attention in tandem or in parallel to enhance the feature map, fail to significantly improve the characterization of the feature map. For shallow feature C2, MSCA and LSKblock can enhance the characterization ability of this feature map. Through the analysis, it can be found that the basic principles of MSCA and LSKblock are similar, both of them acquire multiple long-range contextual information of the object through the convolution of large size, and form different weights in the space. Considering the effect of attention on detection accuracy and model scale, the final choice was to use LSKA for C2, TripleAttention for C3, and ECA for C4 for feature map enhancement.

Comparison of detection performance of different necks

To validate the effectiveness of the neck proposed in this study in road surface damage detection performance, this paper compared the detection performance of different feature fusion neck network models, including FPN, BiFPN, GFPN, EVC, Gold-YOLO, HSFPN, and PAN. This experiment selected Darknet53 as the backbone, adopted a network structure with 3 decoupled detection heads, and performed transformations on the neck structure. As shown in Table 7, this study improved the neck network based on PAN by enhancing the fusion module and the selection model. Ultimately, the detection effect is significantly improved, resulting in a more comprehensive fusion of feature information at all stages. Compared to PAN, the improved method improves mAP_{50} by 2.4%, precision by 1.4%, and recall by 2.3%, while using less computation and number of parameters.

Feature Maps	Attention	mAP_{50}(%)	$mAP_{50:95}$(%)	GFLOPs
C2/C3/C4	–	66.9	35.5	24.4
C4	ECA	67.0(+0.1)	35.8(+0.3)	24.4(+0.0)
	GAM[37]	67.2(+0.3)	35.9(+0.4)	29.7(+5.3)
	CAM[37]	67.1(+0.2)	35.6(+0.1)	24.7(+0.3)
	PAM[37]	66.1(−0.8)	35.4(−0.1)	24.7(+0.3)
	TripleAttention	66.0(−0.9)	34.9(−0.6)	24.5(+0.1)
C3	ECA	66.6(−0.3)	35.0(−0.5)	24.4
	GAM	66.0(−0.9)	35.2(−0.3)	29.7(+5.3)
	CAM	66.5(−0.4)	35.4(−0.1)	24.6(+0.2)
	PAM	65.5(−1.4)	34.9(−0.6)	24.7(+0.3)
	TripleAttention	67.2(+0.3)	35.8(+0.3)	24.5(+0.1)
	CBAM[38]	65.2(−1.7)	34.6(−0.9)	24.5(+0.1)
	CoTAttention[39]	67.0(+0.1)	35.6(+0.1)	26.3(+1.9)
	Nonlocal[40]	66.6(−0.3)	35.6(+0.1)	25.3(+0.9)
	MSCA[41]	66.0(−0.9)	35.1(−0.4)	24.7(+0.3)
	SeaAttention[42]	67.0(+0.1)	35.5(–)	47.4(+23.0)
	LSKblock	67.1(+0.2)	35.8(+0.3)	25.2(+0.8)
C2	GAM	66.4(−0.5)	35.5(–)	29.7(+5.3)
	PAM	66.4(−0.5)	35.4(−0.1)	24.7(+0.3)
	TripleAttention	66.9(–)	35.6(+0.1)	24.5(+0.1)
	CBAM	66.3(−0.6)	35.3(−0.2)	24.5(+0.1)
	CoTAttention	66.6(−0.3)	35.9(+0.4)	26.3(+1.9)
	Nonlocal	65.9(−1.0)	35.1(−0.4)	25.3(+0.9)
	MSCA	67.2(+0.3)	35.5(–)	24.8(+0.4)
	SeaAttention	65.5(−1.4)	35.1(−0.4)	40.9(+16.5)
	LSKblock	67.2(+0.3)	35.9(+0.4)	25.3(+0.9)

Table 6. The influence of using different attention at different stages of the backbone on the model performance evaluation indicators. In the table, "–" indicates that the attention mechanism is not used at the end of each stage of the backbone.

Neck	P(%)	R(%)	F1	mAP_{50}(%)	GFLOPs	Parma(M)
FPN[43]	70.4	59.8	64.7	64.1	25.8	18.8
BiFPN[44]	66.7	59.3	62.8	62.5	28.6	19.8
GFPN[45]	69.8	60.5	64.8	65.1	29.5	23.6
EVC[46]	69.4	61.5	65.2	64.5	43.1	54.7
Gold-YOLO[47]	70.4	60.4	65.0	65.2	62.7	57.6
HSFPN[48]	67.5	57.5	62.1	62.2	24.3	14.0
PAN[49]	71.0	59.2	64.6	64.4	28.4	21.4
Ours	72.4	61.5	66.5	66.8	27.8	20.0

Table 7. Comparison of performance evaluation indicators of different necks on the test set of this study.

Ablation experiments

To analyze the enhancement of the road surface damage detection performance by each module proposed in this study, systematic ablation experiments were conducted in this part and the experiments are presented in Table 8. Each module was validated based on YOLOv8s, and the experiments show that DSCD improves the model's accuracy in detecting road surface damage and reduces the computational and parametric quantities of the model. MFE has a similar effect, and when both are applied simultaneously, better detection results can be achieved and the model can be further lightweight. SPPF can effectively enhance the global sensing capability of the model, enabling the detection model to have a more comprehensive understanding of the road surface environment and damage. The introduction of FSM can effectively enhance the detection accuracy of the model with only a slight increase in the number of parameters and computational costs, the FFM simultaneously realizes model lightweight and accuracy improvement. And the MSD-head increases the computational costs but greatly improves the detection accuracy. Through the comprehensive application of each module, this study proposed RSDD, whose model parametric quantity is only 16.5 M, computational complexity is 30.9G, and the average accuracy for road surface damage detection reaches 70.8%. Compared with the baseline YOLOv8s,

Backbone			Neck		MSD-head	mAP_{50}(%)	F1	Params(M)	GFLOPs
DSCD	MFE	SPPF	FFM	FSM					
		✓				65.3	65.2	21.4	28.4
✓		✓				66.4	66.5	20.0	26.8
	✓	✓				66.5	66.9	20.3	26.0
✓	✓					65.2	65.5	17.6	23.9
✓	✓	✓				66.9	67.0	18.8	24.4
		✓	✓			65.8	65.6	19.8	26.8
		✓	✓	✓		66.8	66.5	20.0	27.8
		✓			✓	69.0	68.7	20.6	36.4
✓	✓	✓	✓	✓		68.3	67.5	17.4	23.7
✓	✓	✓			✓	69.7	69.0	18.0	32.4
✓	✓	✓	✓	✓	✓	70.8	70.2	16.5	30.9

Table 8. The influence of each module on the performance evaluation indicators of the model in ablation experiments.

Models	P(%)	R(%)	mAP_{50}(%)	$mAP_{50:95}$(%)	F1	Params(M)	GFLOPs	Latency(ms)
YOLOv8n	63.1	53.9	57.5	29.4	58.1	5.97	8.1	4.4
YOLOv8s	69.7	61.3	65.3	34.9	65.2	21.4	28.4	5.1
YOLOv8m	72.4	63.7	68.2	37.1	67.8	49.6	78.7	7.0
YOLOv8l	72.1	63.9	67.9	36.9	67.8	83.6	164.8	10.4
YOLOv8x	73.3	65.3	69.0	37.4	69.1	130	257.4	14.7
YOLOv6n	54.8	45.8	47.3	24.2	49.9	8.3	11.8	3.8
YOLOv6s	58.6	52.6	54.6	28.3	55.4	31.3	44.0	4.6
YOLOv6m	63.9	52.0	56.0	30.0	57.3	99.5	161.1	8.9
YOLOv6l	57.7	44.0	48.0	24.8	49.9	211	391.2	16.1
YOLOv5n	60.6	52.9	56.0	28.3	56.5	5.04	7.1	3.8
YOLOv5s	70.1	59.9	64.7	34.2	64.6	17.6	23.8	4.3
YOLOv5m	71.2	63.6	68.4	36.4	67.2	48.1	64	6.7
YOLOv5l	72.3	64.5	68.3	37.4	68.2	101	134.7	9.3
YOLOv5x	73.8	64.7	69.5	38.0	69.0	185	246	15.8
YOLOv9	71.2	62.3	67.7	37.6	66.5	240	116.8	13.3
Ours	72.4	68.2	70.8	38.0	70.2	16.5	30.9	4.5

Table 9. Comparison of performance evaluation indicators of different object detection models on the test set of this study.

RSDD effectively achieves higher road surface damage detection accuracy while reducing the cost of model deployment.

Comparison experiments

To demonstrate the advantages of RSDD, it is analyzed in comparison with four state-of-the-art object detection models (different versions of YOLOv5, YOLOv6, YOLOv8, and YOLOv9). Table 9 shows a detailed comparison of the detection performances of each model on the test set of this study, compared to which RSDD achieved the best performance in terms of mAP_{50}, $mAP_{50:95}$, and F1-Score. Although it is not the smallest of all models, its detection accuracy is much higher than other smaller detection models.

In Fig. 15 a comparison between RSDD and each model in terms of the tradeoffs between speed and accuracy and scale and accuracy is shown, from which it is found that RSDD has a better ability to tradeoff between accuracy, scale, and speed. RSDD can achieve higher detection accuracy with fewer parameters and fast inference speed with higher detection accuracy. RSDD is more suitable to be deployed on devices with limited memory and computational power and is more suitable for some task scenarios that require fast detection. In summary, the detection model proposed in this study has a stronger trade-off between accuracy, speed, and scale, and performs best in road surface damage detection tasks.

Visualization of detection results

To more visually validate the advantages of RSDD, we collected different road surface damage images and detected them on some roads in Xi'an, China, and visualized the detection results. For the fairness of comparison, YOLO models with a similar scale as the RSDD model, i.e., YOLOv5m, YOLOv6m, YOLOv8m, and YOLOv9,

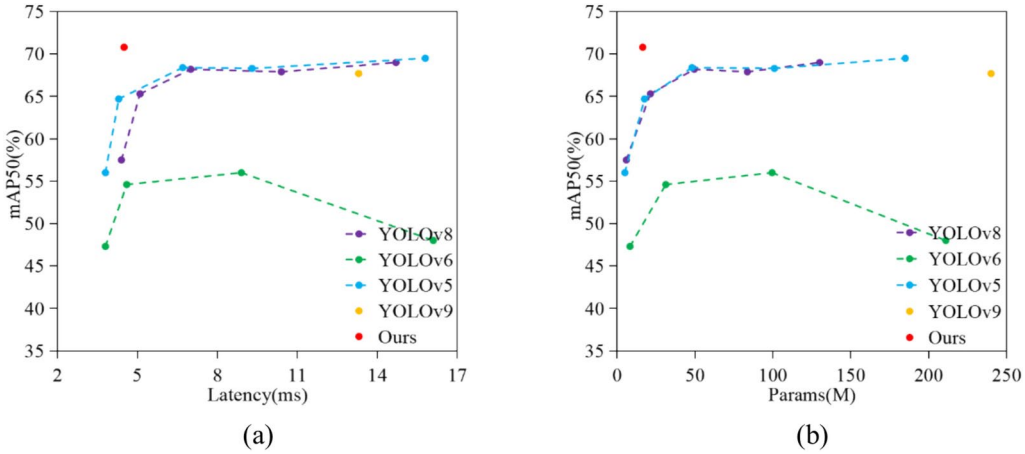


Fig. 15. Comparison of different models in terms of speed-accuracy and scale-accuracy trade-offs. (**a**) Speed-accuracy; (**b**) scale-accuracy.

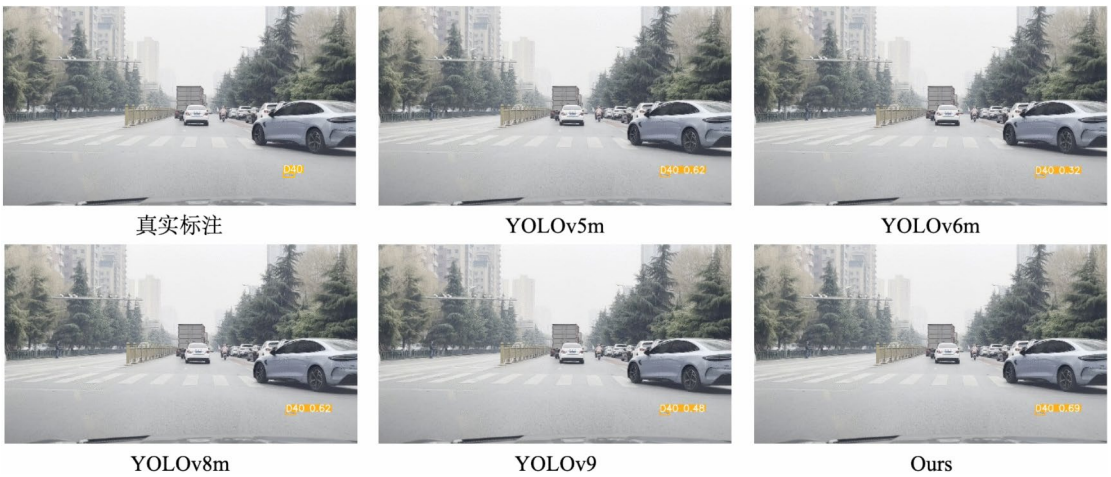


Fig. 16. Visualization of detection results of different detection models for a single category of damage in a scenario-simple environment.

which are the top-ranked versions of the corresponding models in terms of the detection results, were selected for the visualization of the results, respectively.

As shown in Figs. 16, 17, 18, 19, and 20, the visualization of the detection results of different models for a variety of damages in different scenarios is demonstrated. As can be seen in Fig. 16, each model can recognize the location of small-sized potholes in simple scenarios, but RSDD shows the optimal performance in terms of confidence. From Fig. 17a, it can be found that without too many interfering factors, only RSDD and YOLOv5m can effectively detect the location of small-sized pothole, and the confidence level of RSDD is 26% higher than that of YOLOv5m. Moreover, RSDD and YOLOv6m are more accurately localized in tiny longitudinal crack detection. While the sampling time point of Fig. 17b is a little earlier than Fig. 17a, it has a wider view than Fig. (a), which both contain the same pothole, but the location of the pothole is not marked in the labeling of Fig. 17b. However, RSDD can accurately predict the location of potholes and has excellent performance for tiny crack detection. Figure 18 shows the results of the different models on the detection of damage in complex scenarios. The other models have omitted or misdirected the damage, however, the RSDD can still accurately identify these small transverse cracks. In Fig. 19, only the RSDD can recognize the damage that is far away and has thin features. Figure 20 illustrates a complex scenario with the presence of environmental disturbances and at the same time serious road surface damage. From the individual detection results, it can be seen that each model is effective in recognizing road damage for obvious features, but for the small and not obvious features, only the RSDD can detect them effectively. Comparing the visualization of each model in the detection results further proves the advantages of the model proposed in this study in road surface damage detection, which greatly avoids misdetection and leakage detection. In particular, the detection of tiny damage has an obvious improvement effect, which better meets the detection needs in the actual complex background environment.

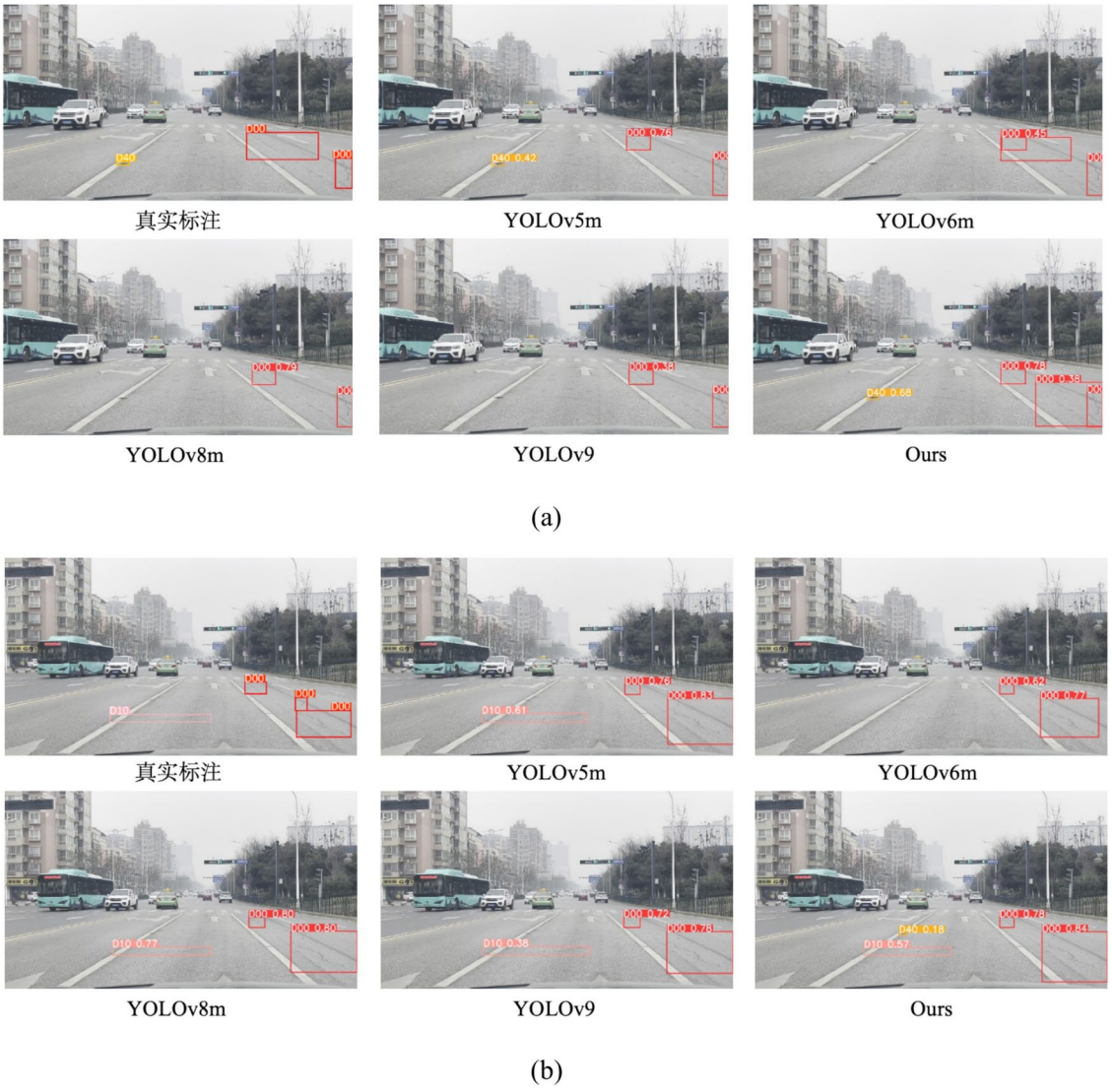


Fig. 17. Visualization of detection results of different detection models for multiple categories of damage in a scenario-simple environment. (**a**) Images after the sampling time; (**b**) images before the sampling time.

Generalizability experiments

To evaluate the generalization ability of RSDD, this study used the publicly available dataset RDD2022 to train each model, which includes road images from six countries, i.e., China, Japan, the Czech Republic, Norway, the United States, and India. In this dataset, we randomly selected 10,000 road images, similar to the previous experimental design, where road damages were categorized as longitudinal cracks, transverse cracks, alligator cracks, and potholes, and the image data were divided into training, validation, and test sets in the ratio of 7:2:1. Table 10 demonstrates the performance evaluation indicators of each model on the test set, and it can be found that RSDD still achieves the best detection accuracy on the RDD2022 test set, with a mAP_{50} of 61.2%. Moreover, the detection accuracy of alligator cracks and potholes is also the highest among all models. By comparing the detection performance on different datasets, the generalization and effectiveness of RSDD for the road surface damage detection task are further validated.

Discussion

Road surface damage detection plays a crucial role in highway maintenance, which can reduce maintenance costs and ensure road safety by timely detecting and repairing the damage. The RSDD network proposed in this study realizes more accurate detection of tiny damages on road surfaces at an early stage, which effectively improves the problem of leakage and misdetection of traditional models in the detection of tiny damage. Specifically, the feature extraction backbone in this study innovatively combined the advantages of small sensory field convolution and large sensory field convolution, so that the model can simultaneously capture detail features, multi-scale contextual information, and global semantic information to comprehensively understand the road surface environment and damage features, and to avoid the problems of insufficient feature extraction

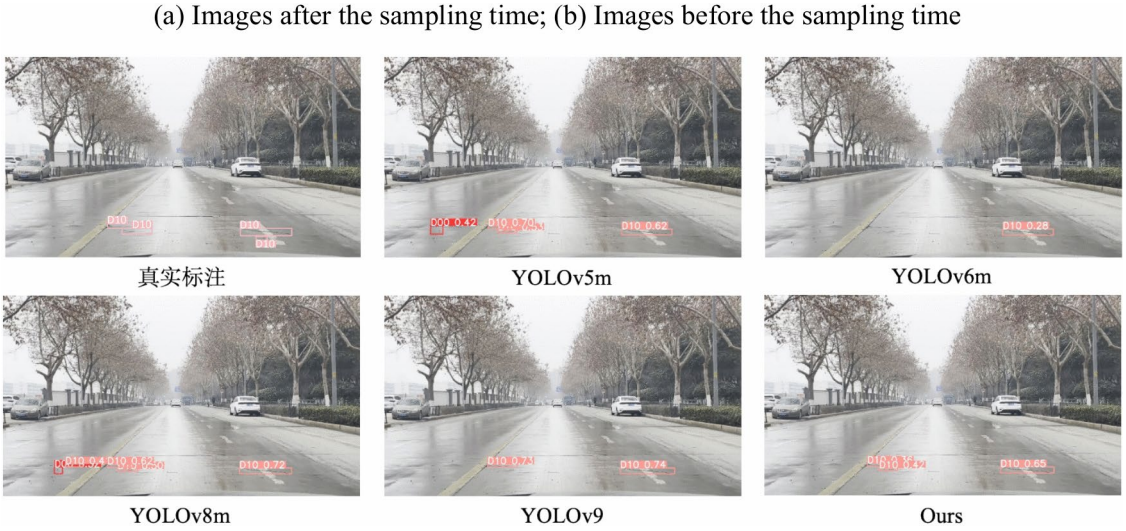


Fig. 18. Visualization of detection results of different detection models for a single category of damage in a complex environment.

Fig. 19. Visualization of detection results of different detection models for multiple categories of damage in a complex environment.

and feature loss that are commonly found in the traditional methods. In addition, this study adopts different attention mechanisms for feature screening for the multilevel features extracted from the backbone and enhances the characterization ability of feature maps at different scales through the two-way feature fusion strategy of top-down and bottom-up. Finally, this study designed a multi-scale detection head, which effectively solves the problem of large size differences in road surface damage and scale changes due to scene perspective by detecting different sizes of damage on each of the four output layers, and significantly improves the detection performance.

The RSDD network was subjected to comprehensive performance evaluation experiments on both self-constructed and publicly available datasets. The experiments show that RSDD exhibits significant performance advantages over state-of-the-art object detection models in the road surface damage detection task, especially in the detection of tiny damages. The model achieves a better trade-off between accuracy, speed, and scale, and can significantly reduce the number of model parameters and improve the computational efficiency while ensuring higher detection accuracy. Specifically, the RSDD network achieves 70.8% of mAP_{50} value with only 16.5 M model parameters and a low latency of 4.5 ms. In addition, RSDD shows superior performance in critical evaluation indicators such as Precision, Recall, mAP, and FLOPs, fully verifying its effectiveness and practicality in road surface damage detection tasks.

Although the RSDD network achieves a better balance between accuracy, speed, and scale, in practical application scenarios, some of the devices are limited by computational resources, and is difficult to meet the demand for high computational power, which poses a challenge to its deployment and application. Therefore, to further enhance the practicality of RSDD, future research needs to focus on the lightweight design of the model for a wider range of practical engineering scenarios.

Fig. 20. Visualization of detection results of different detection models for multiple categories of damage in a more complex environment.

Models	AP_{50}(%)				P(%)	R(%)	mAP_{50}(%)
	Longitudinal cracks	Transverse cracks	Alligator cracks	Potholes			
YOLOv8n	55.3	55.9	66.8	42.3	62.6	50.8	55.1
YOLOv8s	57.2	59.0	68.5	46.8	62.7	54.0	57.9
YOLOv8m	59.0	60.0	68.6	47.8	64.7	54.3	58.9
YOLOv8l	59.5	61.0	68.4	48.5	64.9	54.6	59.4
YOLOv8x	59.8	61.2	68.5	49.5	64.2	56.2	59.7
YOLOv6n	54.1	55.3	65.5	38.5	60.8	49.8	53.3
YOLOv6s	57.4	58.1	67.3	45.2	64.6	51.6	57.0
YOLOv6m	55.5	56.9	65.7	45.2	63.1	51.9	55.8
YOLOv6l	54.6	55.6	63.6	39.2	61.1	49.4	53.2
YOLOv5n	54.3	53.9	64.3	40.4	60.7	49.9	53.2
YOLOv5s	57.0	57.0	67.5	43.7	63.0	51.2	56.3
YOLOv5m	59.0	53.7	65.3	48.5	61.7	54.6	56.6
YOLOv5l	59.6	56.5	65.1	50.3	63.9	54.7	57.9
YOLOv5x	60.1	59.9	69.9	50.4	64.4	55.7	60.1
YOLOv9 (gelan)	62.5	60.8	69.1	49.7	65.5	55.6	60.5
YOLOv9 (gelan-c)	62.0	60.9	68.5	49.8	65.4	55.2	60.3
Ours	60.3	59.2	71.5	53.6	66.7	56.0	61.2

Table 10. Comparison of performance evaluation indicators of different object detection models on the RDD2022 test set.

Conclusion

In this study, a new road surface damage detection model is proposed to address the problems of leakage and misdetection that tend to occur in the detection of tiny road surface damage. The model effectively solves the problems of insufficient extraction and loss of tiny damage features caused by downsampling and feature extraction in the traditional method by introducing DSCD and MFE in the backbone. At the same time, FSM and FFM are utilized to enhance the characterization ability of the feature maps at each stage in the feature pyramid and combined with the multi-scale detection head to achieve accurate detection of different sizes of damage, which significantly improves the adaptability of the model to scale changes. The experimental comparison reveals that compared with the existing object detection models, RSDD has obvious advantages in detection accuracy, inference speed, and model scale, especially making breakthroughs in the tiny damage detection task. The model is more suitable to be deployed in resource-constrained devices and provides important technical support for road maintenance. Although RSDD has made significant progress in terms of accuracy, generalization, and real-time performance, further light-weighting of the model is still a focused direction for future research. Follow-up work will focus on optimizing the computational efficiency of the model to enhance its deployment performance on mobile devices.

Data availability

The open-source datasets used in this study are publicly available, it can be accessed at https://github.com/sekilab/RoadDamageDetector. If you would like to request a copy of the datasets collected in this study and the designed network code, please email the corresponding author.

Received: 16 January 2025; Accepted: 21 March 2025
Published online: 01 April 2025

References

1. Hosseini, S. & Smadi, O. How prediction accuracy can affect the decision-making process in pavement management system. *Infrastructures* **6**(2), 28 (2021).
2. Han, C. et al. Long-term maintenance planning method of rural roads under limited budget: A case study of road network. *Appl. Sci.-Basel* **13**(23), 12261 (2023).
3. Wang, Y. et al. Diagnosis of cervical lymphoma using a YOLO-v7-based model with transfer learning. *Sci. Rep.* **14**(1), 11073 (2024).
4. Wu, D. et al. YOLO-Claw: A fast and accurate method for chicken claw detection. *Eng. Appl. Artif. Intell.* **136**(1), 108919 (2024).
5. Wang, P. et al. Multicategory fire damage detection of post-fire reinforced concrete structural components. *Comput.-aided Civ. Infrastruct. Eng.* **40**(1), 91–112 (2024).
6. Antonio, B., Moroni, D. & Martinelli, M. Efficient adaptive ensembling for image classification. *Expert. Syst.* **42**(1), 13424 (2023).
7. Zhang, S. et al. An enhanced YOLOv8n object detector for synthetic diamond quality evaluation. *Sci. Rep.* **14**(1), 28035 (2024).
8. Kaushal, A., Gupta, A. & Sehgal, V. A semantic segmentation framework with UNet-pyramid for landslide prediction using remote sensing data. *Sci. Rep.* **14**(1), 30071 (2024).
9. Girshick, R., Donahue, J., & Darrell, T., et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 580–587 (2014).
10. Girshick, R. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 1440–1448 (2015).
11. Ren, S. et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2015).
12. Cha, Y. et al. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Comput.-Aided Civ. Infrastruct. Eng.* **33**(9), 731–747 (2018).
13. Ju, H. et al. Detection of sealed and unsealed cracks with complex backgrounds using deep convolutional neural network. *Autom. Constr.* **107**, 102946 (2019).
14. Li, J., Zhao, X. & Li, H. Method for detecting road pavement damage based on deep learning. SPIE Smart Structures + Nondestructive Evaluation, 10972 (2019).
15. Sun, Z. et al. Pavement sealed crack detection method based on improved Faster R-CNN. *J. South China Univ. Technol. (Nat. Sci. Ed.)* **48**(02), 84–93 (2020).
16. Karen, S. & Andrew, Z. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2015).
17. Zeiler, M. & Fergus, R. Visualizing and understanding convolutional neural networks. *Eur. Conf. Comput. Vis.* **8689**, 818–833 (2014).
18. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015).
19. Hascoet, T., Zhang, Y. & Persch, A. et al. FasterRCNN monitoring of road damages: Competition and deployment. In *IEEE International Conference on Big Data*, 5545–5552 (2020).
20. Li, C., Li, L. & Jiang, H. et al. YOLOv6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 (2022).
21. Wang, C., Bochkovskiy, A. & Liao, H. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696 (2023).
22. Jocher, G., Chaurasia, A., Milne, A., Qiu, J. & Ingham, F. YOLO by ultralytics (2023).
23. Wang, C., Yeh, J., & Liao, H. YOLOv9: Learning what you want to learn using programmable gradient information. In *European Conference on Computer Vision*, 15089, 1–21 (2024).
24. Zhang, Y. et al. Road damage detection using UAV images based on multi-level attention mechanism. *Autom. Constr.* **144**, 104613 (2022).
25. Xiang, X., Wang, Z. & Qiao, Y. An improved YOLOv5 crack detection method combined with transformer. *IEEE Sens. J.* **22**(14), 14328–14335 (2022).
26. Diao, Z., Huang, X., Liu, H. & Liu, Z. LE-YOLOv5: A lightweight and efficient road damage detection algorithm based on improved YOLOv5. *Int. J. Intell. Syst.* **2023**, 8879622 (2023).
27. Guo, G. & Zhang, Z. Road damage detection algorithm for improved YOLOv5. *Sci. Rep.* **12**(1), 15523 (2022).
28. Wang, J. et al. Road defect detection based on improved YOLOv8s model. *Sci. Rep.* **14**(1), 16758 (2024).
29. Doshi, K., & Yilmaz, Y. Road damage detection using deep ensemble learning. In *IEEE International Conference on Big Data*, 5540–5544 (2020).
30. Jeong, D. Road damage detection using YOLO with smartphone images. In *IEEE International Conference on Big Data*, 5559–5562 (2020).
31. Mandal, V., Mussah, A., & Adu-Gyamf, Y. Deep learning frameworks for pavement distress classifcation: A comparative analysis. In *IEEE International Conference on Big Data*, 5577–5583 (2020).
32. Wang, S., Tang, Y. & Liao, X. et al. An ensemble learning approach with multi-depth attention mechanism for road damage detection. In *IEEE International Conference on Big Data*, 6439–6444 (2022).
33. He, K. et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1904–1916 (2015).
34. Wang, Q., Wu, B. & Zhu, P. et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11531–11539 (2020).
35. Misra, D., Nalamada, T. & Arasanipalai, A. et al. Rotate to attend: convolutional triplet attention module. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3138–3147 (2021).
36. Liu, M. et al. LSKANet: Long strip kernel attention network for robotic surgical scene segmentation. *IEEE Trans. Med. Imaging* **43**(4), 1308–1322 (2024).
37. Liu, Y., Shao, Z. & Germany, G. Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv preprint arXiv:2112.05561 (2021).
38. Woo, S., Park, J. & Lee, J. et al. CBAM: Convolutional block attention module. In *European Conference on Computer Vision*, Vol. 11211, 3–19 (2018).
39. Li, Y. et al. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(2), 1489–1500 (2023).
40. Zhang, Y., Li, K. & Zhong, B. et al. Residual non-local attention networks for image restoration. arXiv preprint arXiv:1903.10082 (2019).

41. Guo, M., Lu, C. & Hou, Q. et al. SegNeXt: Rethinking convolutional attention design for semantic segmentation. In *Advances in Neural Information Processing System 35*, Neurips (2022).
42. Wan, Q., Huang, Z. & Lu, J. et al. SeaFormer: Squeeze-enhanced axial transformer for mobile semantic segmentation. arXiv preprint arXiv:2301.13156 (2023).
43. Lin, L., Dollar, P. & Girshick, R. et al. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 936–944 (2017).
44. Tan, M., Pang, R. & Le, Q. EfficientDet: Scalable and efficient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10778–10787 (2020).
45. Zhao, G., Ge, W. & Yu, Y. GraphFPN: Graph feature pyramid network for object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2743–2752 (2021).
46. Quan, Y. et al. Centralized feature pyramid for object detection. *IEEE Trans. Image Process.* **32**, 4341–4354 (2023).
47. Wang, C., He, W. & Nie, Y. et al. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. In *Advances in Neural Information Processing Systems* (2023).
48. Chen, Y. et al. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. *Comput. Biol. Med.* **170**, 107917 (2024).
49. Wang, W., Xie, E. & Song, X. et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 8439–8448 (2019).

Acknowledgements

The authors would like to express their thanks for the support of the Key Research and Development Program of Shaanxi Province (2023-YBSF-104) and the Natural Science Basic Research Program in Shaanxi, China (2022JM-172).

Author contributions

C.W.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing—original draft, and Writing—review and editing. M.Y.: Formal analysis, Supervision, and Writing—review and editing. H.L.: Data curation and Investigation. J.Z.: Investigation and Validation.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025