ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Road Defect Detection Using Deformable Convolution and Context Enhancement Algorithms

Siyu Dong[1,*] *and Zhuozhen Xu*[1]

[1] School of Computer Science, Hangzhou Dianzi University, Hangzhou, Zhejiang, 310018, China

SUMMARY: *We proposed a road defect detection method that applies deformable convolution and context enhancement algo- rithms. Our method focuses on texture feature extraction and fusion to enhance the defect detection capability of the model. Our method uses deformable convolution to effectively adapt the considerable variations in shape and size among various types of road defects. By obtaining more efficent features, our model avoids the region of interest deviating from the ground to appear in the sky. An enhanced contextual module is introduced to facilitate the more efficient fusion of multi-scale texture features. This adaptation tackles the challenge of varying defect sample scales arising from the fluctuating distance between the camera and the ground. Additionally, Our Method also incorporates the CBAM (Convolutional Block Attention Module) to obtain supe- rior feature representation and higher level of critical information perception by considering both spatial positional information and channel-related details. The experimental results show that the mAP is increased to 64.7%, and the number of parameters is reduced to 13.5M. This method not only successfully obtains stronger defect feature expression ability, but also improves the fusion ability of multi-scale features. The proposed model is a high-performance and low-cost road defect detection model.*

KEYWORDS: *Road defect detection, Deformable Convolution, Context Enhancement, Attention Mechanism, Multi-scale fea- tures, YOLOv5s, Deep Learning*

1 Introduction

Roads infrastructure play a indispensable role in modern so- ciety. It support for people movement and good transportation [1, 2]. by using automation technology, the cost of regulaly road inspection tasks is reduced, the efficiency is improved, when the incidence of traffic accidents is reduced[3]. There- fore, an efficient and accurate road defect detection algorithms is a crucial part in road evaluation-maintenance system, or enabling automated driving system[4, 5].

With the continuous development of deep learning, DC- NNs(Deep convolutional Neural Networks) based method has become a popular method, because of its excellent logical abstraction ability and feature extraction ability.

DCNN-based road defect detection methods are categorized into three types[1]: SSD (Single Shot Multibox Detector), R- CNN, and YOLO (You Only Look Once). SSD, as a single- stage target detection network, contains two main components: the SSD head part and the backbone model part[6]. Gupta et al[7]. proposed a SSD model for all-weather pothole

*dongsy3790@163.com
https://doi.org/10.65102/is20261141 Published: 10 June, 2026

detection in road thermal images which uses ResNet-50 as the backbone network. R-CNN is a two-stage target detection network. Ko- rtmann et al[8]. proposed the use of an additional network of regional experts to determine the location of defect. And each independent RCNN is used to identify cracks and potholes in the specified area. However, due to Faster R-CNN‘s two-stage architecture, this model possesses a larger number of model parameters and higher computational complexity compared to single-stage detectors. In contrast, YOLO is a single-stage target detection network with faster speed[9]. The YOLO- based detection algorithm segments the road image into a set of fixed-size grids, determines the position of the bounding box by calculating the class probability of each grid. The model structure of the algorithm is more intensive and more efficient. Baek et al[10]. focused on detecting potholes by construct-ing a model with smaller parameter scales based on YOLOv1. Due to the additional edge detection model excluding defects other than potholes, the model still achieves acceptable pot- hole detection performance. Soung et al[11]. constructed a pothole detection model based on YOLOv2[12], and due to the inclusion of five new anchor frames provided by the K-means algorithm, the model is more fitting to the boundary boxes obtained from medium to large-sized pits. Shim et al[13]. designed a lightweight semantic segmentation network.

Optimize the parameters of the model, but do not consider whether the detection speed of the model will be affected.

In addition to model optimization, in-depth research has been conducted to address the issue of variable defect sample scales caused by the variability of the distance between the camera and the road surface. khwah et al[14]. applied dif- ferent variations of the YOLO_ v3[15] architecture, including YOLO_v3, YOLO_v3 Tiny, and YOLO_v3 SPP, to train a pothole detection model. They focused on grey-scale road im- ages and aimed to address the challenge of detecting potholes at various scales in asphalt pavement images. Jeong et al[16]. proposed three integrated learning strategies to enhance the de- tection performance of the YOLO_v5x model. These strategies aim to address the challenge of multi-scale object detection while considering computational efficiency. Sheta et al[17]. proposed a lightweight convolutional neural network model to detect pavement cracks.The architecture performed well in detecting cracks, but it cannot solve the multi-scale problem due to the varying size of defects.Li et al[18]., proposed a method for detecting multi-scale crack features in pavement images using densely connected and deeply supervised net- works. The approach aimed to leverage the complementary information from features at different levels and fuse feature maps at each scale to improve crack extraction. However, this method has a ability to extract fine cracks in pavement images with many interfering factors.Wu X et al[19]. generate pixel-centred blocks on several different scales and input the blocks into different convolution operations.However, this multi-scale and multi-convolution approach is at the expense of higher computational cost to acquire more effective crack features and improve detection accuracy.Liu et al[20]. designed a pixel- level classification network that integrates local and global information to enhance crack detection accuracy. The network architecture is designed to capture richer multi-scale feature information, allowing for a more comprehensive understanding of cracks in images.

In summary, the effectiveness of road defect features needs to continue to be improved, and the model’s feature extraction as well as multiscale feature fusion capabilities are low when dealing with interference tasks with similar textures.To address the above problems, based on the YOLOv 5s model, we proposes a method that applies deformable convolution and context enhancement algorithms to road defect detection. This method reduces the interference of similar features, improves the feature expression ability, enhances the effect of multi- scale feature fusion, and improves the accuracy and robustness of road defect detection. Overall, the main contributions of this study to road crack defect detection are summarized in

the following three parts:

1) Deformable Convolution Net(DCN) is used to improve the feature extraction ability of the model when dealing with interference tasks with similar textures, and the detection attention concerns are concentrated on the ground to solve the problem that the interesting region is occasionally shifted to the sky in traditional methods.

2) Context Augmentation Module (CAM) is used to en- hance the contextual information and solve the problem of multi-scale feature fusion in the process of road defect sample acquisition where the distance between the camera and the road surface is not fixed.

3) CBAM is used can enhance the perception and selection of features of different channels and spatial locations, to further improve the performance of road defect detec- tion.

This paper is structured as follows:

Section II describes the improvement of the road defect detection model network. Section III describes the experimen- tal configurations, datasets, preparatory experiments, and the results of the experiments, and their visualization and analysis. Finally, in Section IV, the paper concludes and offers insights into prospects.

2 Proposed Method

2.1 Overview

Deformable convolution is introduced to improve the model's ability to process multi-scale images under different fields of view, and by enhancing the model's ability to process contextual features, so as to improve the performance of road defect detection in real scenes. Our method is an improved model based on YOLO v5. The overall structure of the model is shown in Fig. 1. It mainly consists of three parts: Backbone, Neck and Head. Backbone serves as the entrance of the neural network, and its main function is to extract the road defect features and generate rich feature map information for use by the later modules; Neck is responsible for multi-scale feature fusion of the feature maps and delivers these features to the prediction layer; and Head performs the final prediction.

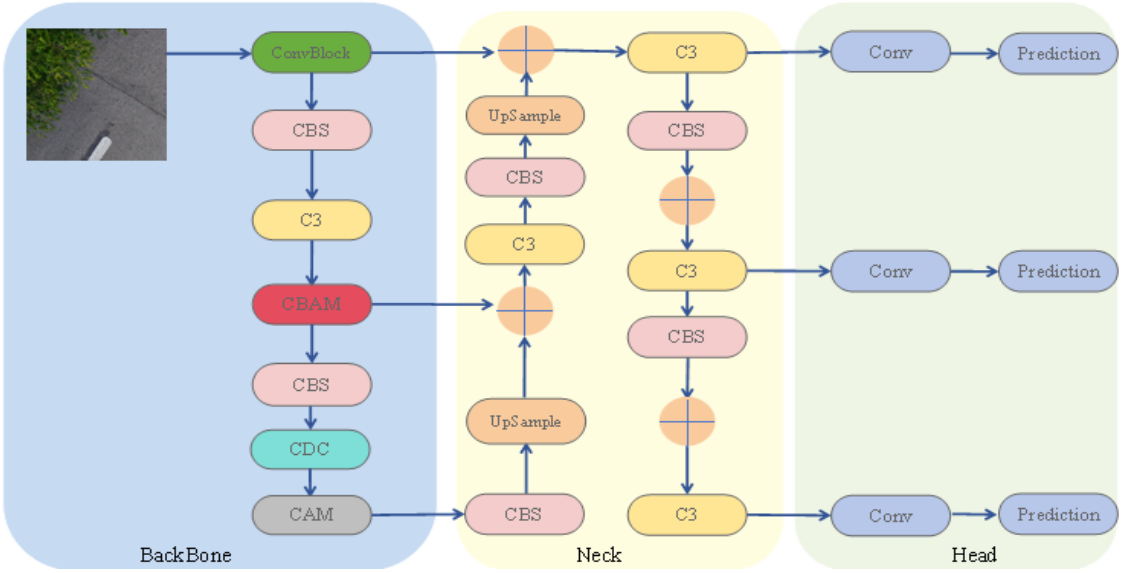

Figure 1: Road Defect Detection Model Based on Deformable Convolution and Context Enhancement

In the Backbone section, CDC module is designed and CBAM and CAM context enhancement module are introduced. It contains several internal modules, the details diagram of which are shown in Fig. 2. The internal module mainly consists of CBS module, which is composed of Conv2D, BatchNorm and SiLu modules, and the CBS module is the key component for extracting the features.Bottleneck module is composed of two CBS modules in series, and then the residual structure is summed up with the initial input to get the output value at last.The C3 module does two kinds of processing for the inputs respectively, one is to obtain the output value output1 through the residual network, the other is to obtain the output value output2 directly through the Conv convolutional network, and finally the two outputs values are added together and then the final output value is obtained through a CBS module.The Block module and the ConvBolock module are composed of one CBS and two CBSs connected in series with the C3 module, respectively.

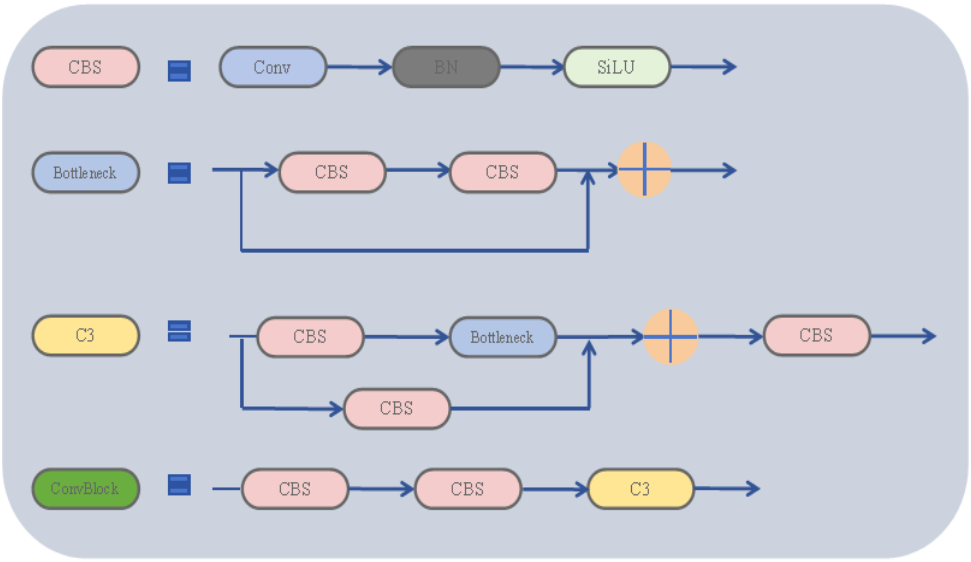

Figure 2: detailed map

2.2 Deformable Convolution Module

Existing road detect. As shown in Fig. 3, 3x3 standard convolution versus deformable convolution sampling locations:

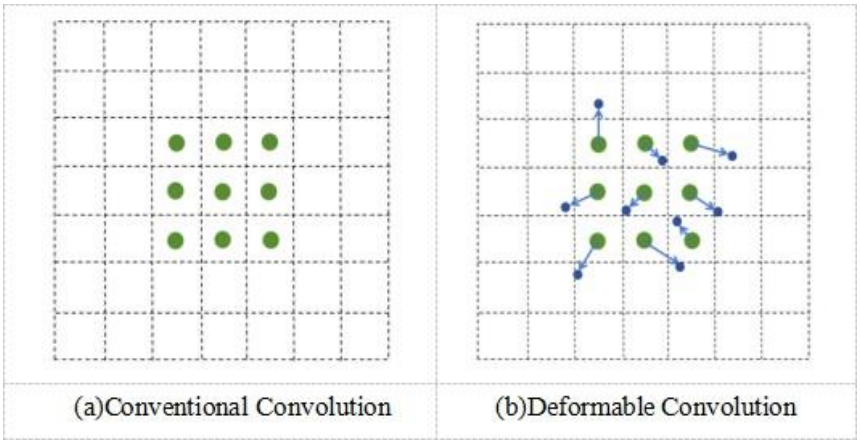

Figure 3: 3x3 standard convolution versus deformable convolution sampling locations

The deformable convolution structure is shown in Fig. 4. The deformable convolution extracts feature maps based on the input image using traditional convolutional kernels, and then takes the resulting feature maps as inputs and passes through one more convolutional layer in order to obtain the deformation offsets for deformable convolution, which is similar to traditional convolutional neural networks. During the training process, the convolution kernel used to generate the output features and the convolution kernel used to generate the offsets are learned simultaneously. Our method uses DCNv3 as the core operator to improve the network.

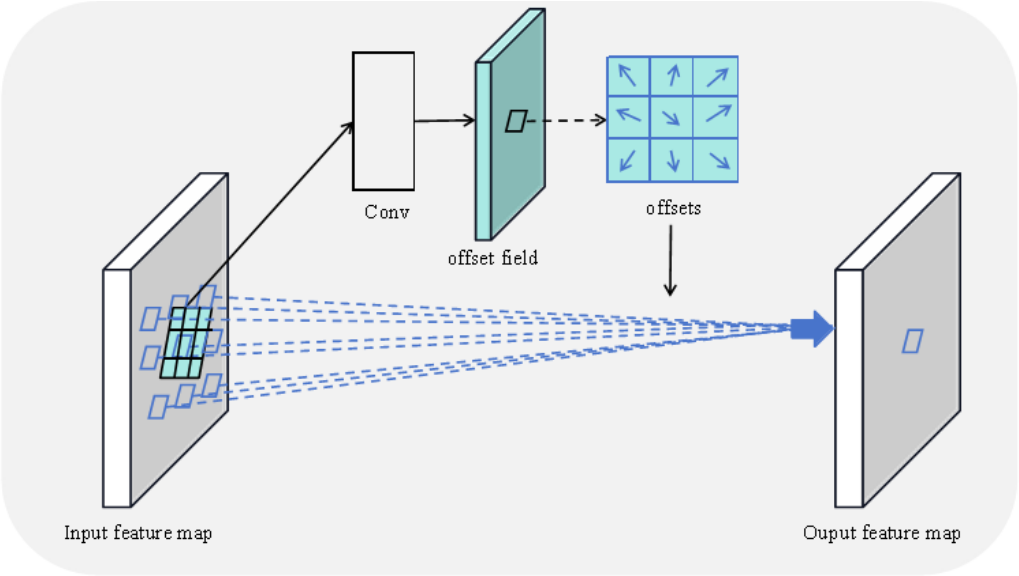

Figure 4: Deformable Convolutional Structures

Traditional convolution consists of two steps:

1) sampling a regular grid R on the input feature map X $X \in RC\times H\times W$, and weighting and summing the sampled values with weights denoted as w. The size of the grid R is the same as the size of the sensory field, and for each position p0, on the output feature map y, there are:

$$y(p_0) = \sum_{p_n \in R} W(p_n) * X(p_0 + p_n) \quad (1)$$

where n denotes the total number of sampling points, p_n denotes the position in traversing R.

Traditional convolution operates on regular grids. In DCNv1[21],the valid regions by adding an offsets term Δp_n on the regular grid R ,It allows the sampling grid to be freely deformed and sampled at irregular and offset positions $p_n + \Delta p_n$ to extract irregular features. The offsets are learnt from the previous feature maps by an additional convolutional layer and are adjusted by the input x. For each position p_0 on the output feature map y, there are:

$$y(p_0) = \sum_{p_n \in R} W(p_n) * X(p_0 + p_n + \Delta p_n) \quad (2)$$

In DCNv2[22], each sample not only learns the offsets in DCNv1, but also modulates them by the learned feature magnitudes, so that the module is able to change the spatial distribution of the samples and their influence on each other. For each position p_0 on the output feature map y, there are:

$$y(p_0) = \sum_{k=1}^{K} W_k \cdot X(p_0 + p_k + \Delta p_k) \cdot \Delta m_k \quad (3)$$

where Δm_k is the modulation scalar at the kth position, $\Delta m_k \in [0, 1]$. Δp_k and Δm_k can be obtained by a separate convolution on the same input feature map x, which has the same spatial resolution and expansion as the current convolution layer. The number of output channels is 3K, where the first 2K channels correspond to the learned offsets Δp_k and the remaining K channels are then sent to the sigmoid layer to obtain the modulation scalar Δm_k. For long-range dependencies, the sampling offsets Δp_k are flexible enough to allow interaction with long-range features. For adaptive aggregation, both the sampling offsets Δp_k and the modulation factor Δm_k are learned and related to the input x.

In DCNv2, each modulation scalar is normalised by sigmoid and the modulation scalars are in the range [0, 1], with the modulation scalar sums of all sampled points varying in [0, k]. As a result, the gradient of the output in the DCNv2 layer is unstable when training with parameters and data that take a wide range of values. Therefore, the sigmoid normalisation for element pairs is correspondingly changed to softmax normalisation for multiple sample points, which achieves constraining the sum of modulation scalars to 1, making the model training process more stable at different scales. Combined with the above modifications, the extended DCNv2 is labeled as DCNv3[23]. For each position Δp_0 on the output feature map y, there are:

$$y(p_0) = \sum_{g=1}^{G} \sum_{k=1}^{K} W_g \cdot X_g(p_0 + p_k + \Delta p_{gk}) \cdot \Delta m_{gk} \quad (4)$$

where G denotes the total number of aggregated groups. For the gth group, W_g denotes the position-independent projection weight of the group, Δm_{gk} denotes the modulation scalar of the kth sampling point in the gth group, which is normalised by a softmax function along dimension K. X_g denotes the input feature map of the slice. Δp_{gk} is the offset corresponding to the grid sampling position Δp_k in the gth group.

DCNv3 is introduced to reconstruct the CDC mod- ule.DCNv3 improves the feature processing capability of the model by predicting the sampling offsets and modulation scales of the input features x through separable convolution. The introduction location of DCNv3 was studied by exper- iments.The combination is shown in Fig.5, the experimental results are shown in Table 1, and the visualisation results are shown in Fig.6.

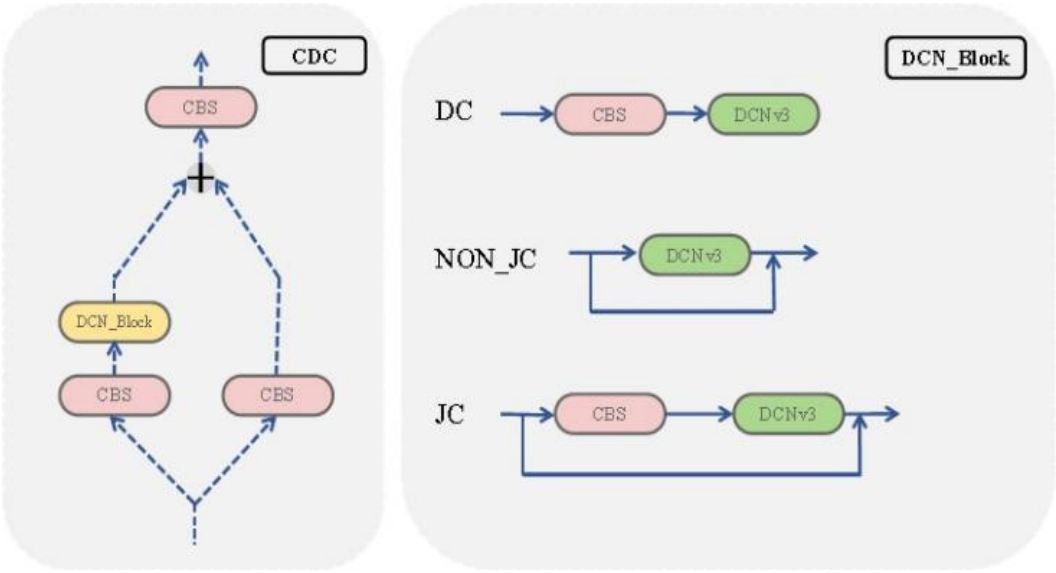

Figure 5: CDC module

Table 1: Experimental results of DCNv3 combination mode

Module	mAP@.5	Precision	Recall	Parameters	GFLOPs
YOLOv5s	56.9%	62.2%	54.9%	7.03M	16.0
DC	59.1%	46.3%	51.6%	6.31M	14.4
NON-JC	59.4%	64.6%	51.7%	6.31M	14.4
JC	62.0%	63.5%	56.7%	6.31M	14.4

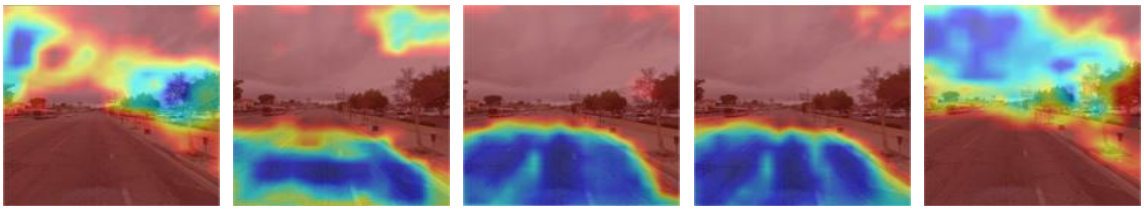

Figure 6: DCNv3 combination approach visualisation results

According to the experimental results, the performance of the models is improved by adding DCN_Block module to the existing models. The experiments are done by designing different DCN connection structures for DCN_Block and it is found that the best experimental results are obtained by combining the (JC) combination approach. Therefore, our method used the (JC) combination to reconstruct the CDC model.

DCNv3 is able to better adapt to complex textures and shapes, and better extract defective texture features when dealing with interference tasks with similar textures by learn- ing deformable sampling and modulation scales. Therefore, deformable convolution can be uesed to effectively extract the texture features of ground defects when facing the interference of sky texture, thus correcting the problem that the focus of the established methods appears in the sky, locating the focus to the ground, and effectively improving the model's capability to detect road defective textures.

2.3 Context Augmentation Module

The impact of the multi-scale problem on the model per- formance is more significant, reducing the accuracy of road defect detection, due to the uncertainty of the distance between the shooting camera and the ground. Multi-scale features have different information densities among different scale features, which cannot be directly fused, and directly fusing multi-scale features without additional processing leads to semantic con- flicts, resulting in limited expression of multi-scale features. Therefore, our method used context enhancement technology to realize the fusion of multi-scale features with different information densities to obtain more effective context features.

1 In YOLO_v5s model, SPPF is the key feature fusion module. In contrast to the SSD model, which utilizes a series of convolutional filters to detect objects across various levels of feature maps, achieving a similar effect to the feature pyramid. CAM (Context Augmentation Module) module[24] employs an adaptive fusion method to capture the features of defects at different scales by using dilation convolution with different dilation rates on the feature maps after different degrees of downsampling to obtain contextual information from different receptive fields, thus enhancing the ability of multiscale defect detection. Our method experimentally compared the use of multiple convolution rates to obtain the best results. The CAM structure is shown in Fig. 7.

The experiment employs various convolutional opeartions and diverse dilation rates to capture feature information of dis- tinct scales for the feature graph C5 subsequent to extracting features from the preceding layer. The convolution operation after one convolution

kernel of size 3x3 with a dilation convolution rate of 1 is denoted to as scheme 2; and after three convolution kernels of size 3x3 with dilation convolution There are three strategies for fusion, as shown in Fig. 8. Strategy (a) adopts weighted fusion and strategy (b) adopts adaptive fusion, that is, the size of the input can be expressed as (bs, C, H, W), and the spatial adaptive weight connection of (bs, 3, H, W) was obtained through convolution operation and calculated using Softmax. Three channels correspond to the three inputs one by one, and contextual information is aggregated to the output by calculating the weighted sum. which can improve the defect detection ability of the model at different scales and enhance the multi-scale feature extraction ability of the model. Strategy (c) is to do concatenation, which operates the input feature maps in series.

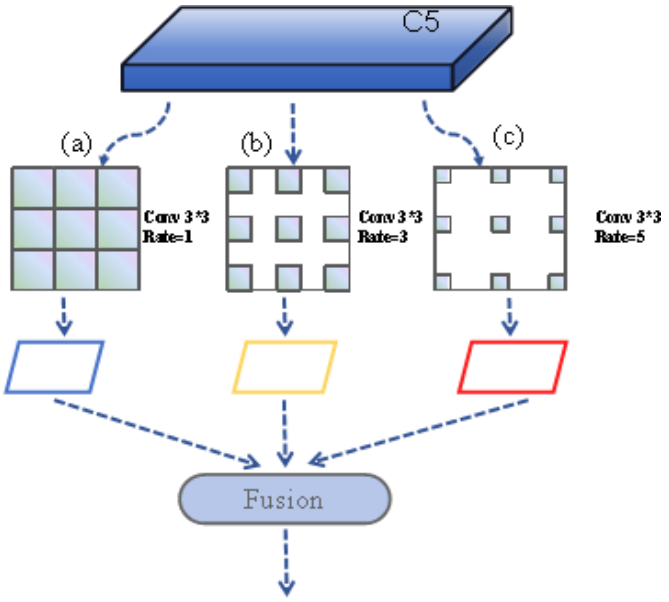

Figure 7: Context Augmentation Module Structure

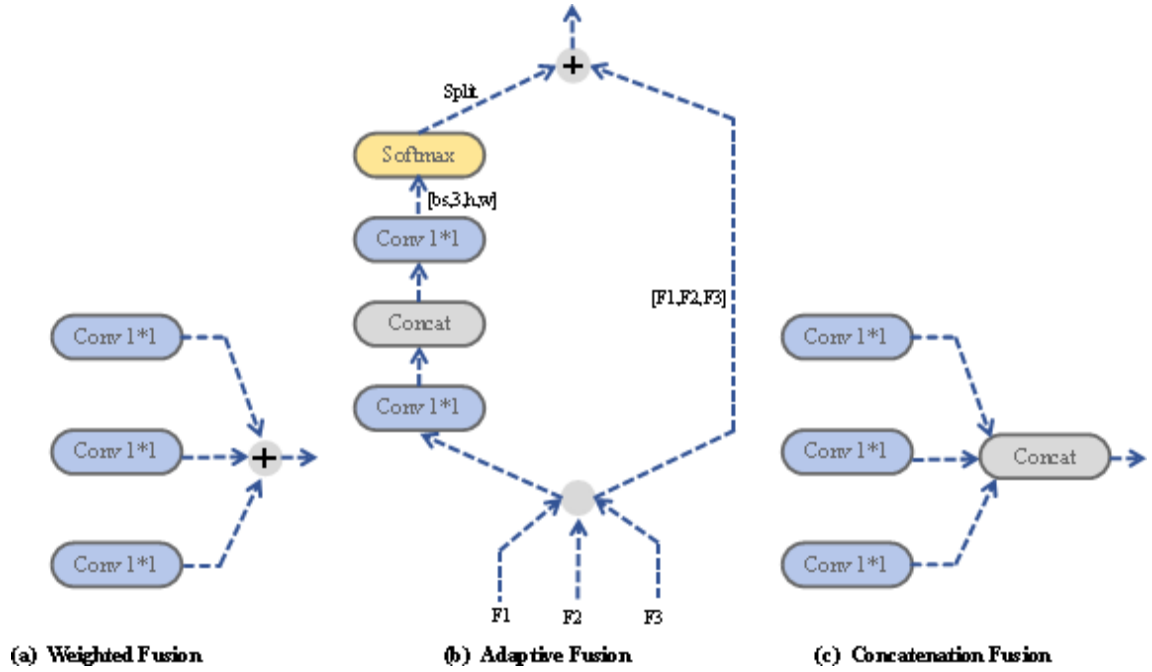

Figure 8: fusion method (a), (b), (c)

The experimental results of processing C5 feature maps through different convolution operations and expansion convo- lution rates are shown in Table 2, and the visualization results are shown in Fig. 9.

Table 2: Experimental results of different convolution operations and expansion convolution rates to process C5 feature maps

Module	mAP@.5	Precision	Recall	Parameters	GFLOPs
YOLOv5s	56.9%	62.2%	54.9%	7.03M	16.0
Scheme 1	56.6%	64.6%	48.2%	8.66M	16.3
Scheme 2	59.4%	65.6%	51.6%	11.3M	18.4
Scheme 3	60.0%	62.1%	54.3%	13.9M	20.5

According to the experimental results, it can be seen that the best experimental results are obtained for scheme 3. This is due to the fact that features with different sensory fields are obtained by dilation convolution operations with different dilation convolution rates, and the contextual information can be enriched by fusing the results of these convolution operations to obtain better feature extraction effects.

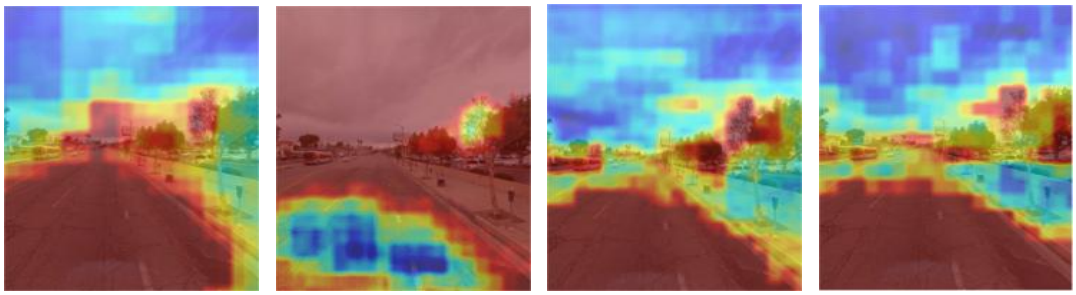

Figure 9: the Visualisation Results

2.4 Channel and Spatial Attention Module

Feature extraction networks extract more accurate target fea- tures by continuously adjusting the parameters of the convo- lution kernel during the training process. However, traditional convolution has the characteristic of focusing only on local information and ignoring global information during the train- ing process of neural networks. For the road defect detection task, it is necessary to rely on the defective texture features extracted by the neural network to improve the detection of cloud-like textures and multi-scale features. Therefore, our method introduced the CBAM attention module[25] into the model to improve the detection of some difficult-to-localise road defects by improving the spatial interaction between the defect texture features.

CBAM can enhance the ability of information interaction between different locations in the feature map, and through the attention mechanism enhance the attention to different spatial scopes of the defect texture features, which is con- ducive to better distinguishing the global information from the detailed information of the defect area, and thus improve the model recognition ability of defects. The Convolutional Block Attention Module (CBAM) combines the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). The structure of CBAM is shown in Figure 10. is shown in Fig. 10.

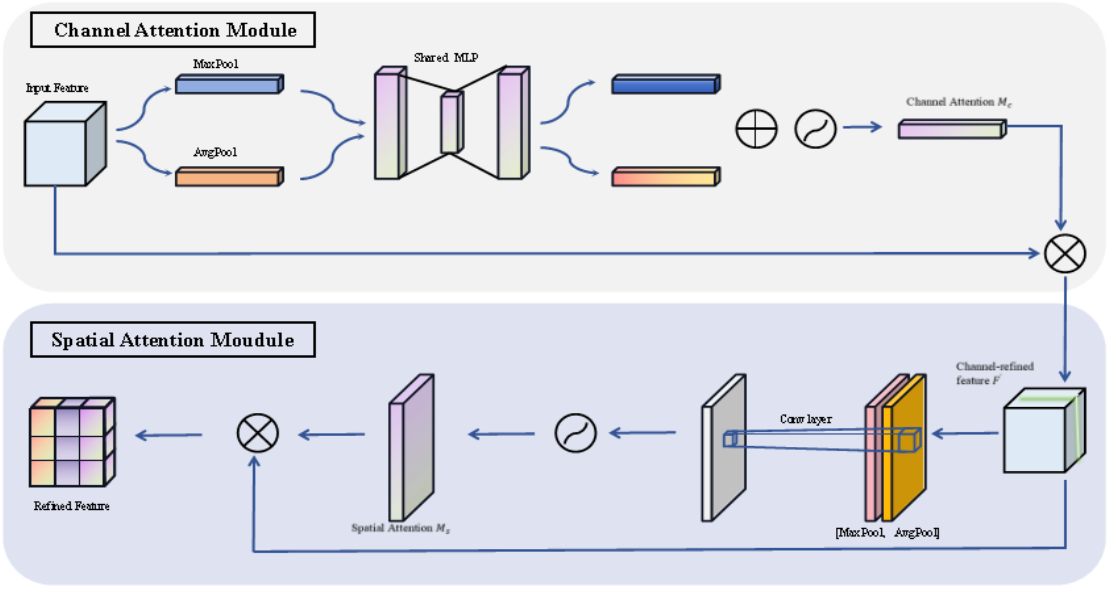

Figure 10: CBAM Attention Mechanism Structure

CBAM input features $F \epsilon R^{C*H*W}$, multiply the convolu- tion result by the one-dimensional convolution $M_c \in R^{c*1*1}$ of CAM by the original feature map, take the CAM output as input, perform two-dimensional convolution $M_s \in R^{1*H*W}$ by SAM, and then multiply the output result by the original feature map.

$$F' = M_c(F) \otimes F \tag{5}$$

$$F'' = M_c(F) \otimes F' \tag{6}$$

CAM controls the importance of different channels by learning the relationship between them. First, the features of each channel are compressed using global average pooling and maximum pooling, then the weights between different channels are adjusted through a fully connected layer, and finally the weights are normalized using an activation function. This allows the network to focus more on the channels that are important for a particular task and thus extract more useful features.

$$M_c(F) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathrm{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\mathrm{F}))) \tag{7}$$

SAM controls the importance of different positions by learning their relationships in the feature map. It first uses the results of the CAM as input, then learns the weights between different locations through a convolutional layer and finally normalizes the weights using an activation function. This allows the network to focus more on the spatial locations that are important for a particular task and thus extract more accurate features.

$$M_s(F) = \sigma(\mathrm{f}^{7\times7}([\mathrm{AvgPool}(\mathrm{F}) ; \mathrm{MaxPool}(\mathrm{F})])) \tag{8}$$

By introducing the CBAM module, the feature extraction network can automatically learn the importance of different channels and locations to extract more useful and accurate features.

Through the above analysis, compared with the original feature extraction network, the introduction of CBAM atten- tion module can make the network in the feature extraction

process to strengthen the spatial interaction between defective features, which can effectively improve the detection effect of road defects.

3 Experiments

3.1 Datasets

Our method is based on a dataset of road images provided by the IEEE Big Data Global Road Damage Detection Chal- lenge 2022 (RDD2022)[26, 27]. These images were captured by in-vehicle smartphones or drones and contained 38,385 road images from six countries: China, Japan, India, Czech Republic, Norway, and the United States. The data set contains four types of road damage, namely longitudinal cracks, trans- verse cracks, alligator cracks and potholes. Fig. 11 illustrates the four road crack patterns.

Figure 11: Road defects

Our method divided the data set into training set, verification set and test set according to the ratio of 8:1:1. The distribution of the data set is shown in Table 3. The training set is used for the learning of road defect characteristics by the neural network. The validation set is used to evaluate the model's learning of the defect features during the training process to better validate the model effect. The test set is then used to evaluate the generalization ability of the model.

Table 3: Road defects dataset

Train	Test	Val	Total
Labels 17194	3060	3513	23767
Images 27734	4894	5757	38385

3.2 Preparatory Experiment

By observing the existing methods in road defect detection, the problem of sensing areas focusing on the sky occurs occasionally. In order to explore the characteristics of the sky and the ground at the feature level, Our method first conducts three sets of preparatory experiments based on the grey-level co-occurrence matrix. Gray-level co-occurrence ma- trix (GLCM)[28] is obtained by calculating the gray-level image, and then some eigenvalues of the matrix are obtained by calculating this co-occurrence matrix to represent some texture features of the image respectively. Our experiments are compared in horizontal, vertical and tilted directions re- spectively. ASM, Contrast, Entropy and IDM metrics are used for validation.

(1) ASM: Angular Second Moment also called energy. The energy transform reacts to the uniformity of gray scale distribution and texture coarseness. the smaller the ASM, the more detailed the texture features.

$$\mathrm{ASM} = \sum_i \sum_j p(i,j)^2 \tag{9}$$

(2) Contrast: react to the depth of the grooves.the larger the contrast, the clearer the texture.

$$\mathrm{Con} = \sum_i \sum_j (i-j)^2 p(i,j) \tag{10}$$

(3) Entropy: react to the complexity of different regions.the larger the Entropy, the more complex the image.

$$\mathrm{Ent} = -\sum_i \sum_j p(i,j) \log p(i,j) \tag{11}$$

(4) IDM: Inverse Differential Moment. React to the clarity and regularity of the texture, which is clear and regular. the larger the IDM, the easier it is to uncover texture features.

$$\mathrm{IDM} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{P(i,j \mid d,\theta)}{1+(i-j)^2} \tag{12}$$

The results obtained through the gray scale covariance matrix are shown in Fig. 12.

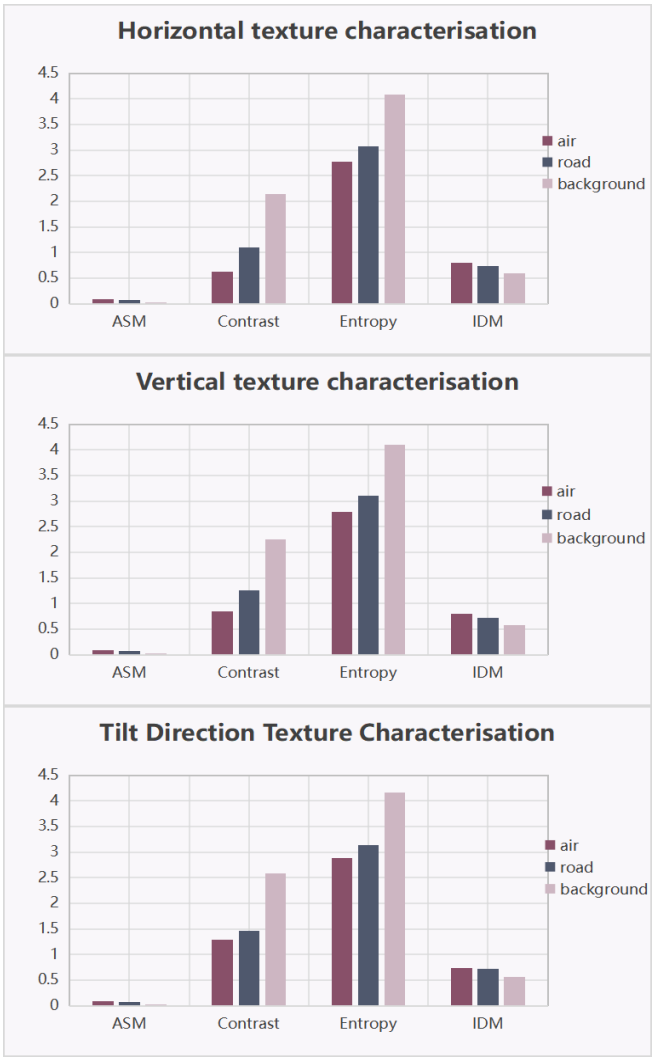

Figure 12: Gray-level co-occurrence matrix results

According to the IDM values, it is known that the sky, ground and background have similar degree of clarity and regularity, so their texture features are similar. According to the ASM value it is known that the sky and ground texture features are similar in degree of detail, therefore feature points may be present on both sky and ground. The Contrast and Entropy values show that the background is closer to the sky and ground compared to the sky and ground, and the background feature differences are more different from the sky and ground. The results of the preparatory experiments show that the complexity and information entropy of the textures of the sky and the ground are somewhat similar, which causes the model to have difficulty in distinguishing between the sky and the ground during the processing of some images.

3.3 Metrics of Evaluation

In order to evaluate the effectiveness of the model in detecting road defects and to validate the performance of the proposed method, the experiments were conducted using metrics such as mAP, GFLOPs, and the number of parameters to evaluate the proposed model.

(1) mAP: Mean Average Precision, is the main evaluation index of the target detection algorithm. It is the average value of AP (Average Precision), the higher the mAP value, the better the effect of the model detection.

(a) The accuracy rate: The accuracy rate, also known as the detection rate, is used to assess the accuracy of the model for classification, that is, the probability of detecting the target correctly, which can be expressed by the following equation:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{13}$$

where P denotes the precision rate, TP (True Positive) denotes the number of samples that are actually positive and predicted to be positive, when there are instances of damage in the ground truth, the number of instances that the model predicts to be of the correct type of defect during the detection process; FP (False Positive) denotes the number of samples that are actually negative but predicted to be positive, that is, the number of instances that the model predicts to be a particular instance of a defect during the detection process, but that are actually a non-defective or other defective type.

(b) The recall rate: also known as the check all rate, is used to assess the probability that the model recognizes a positive sample correctly, and it can be expressed in the following equation:

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{14}$$

where R denotes the recall rate and FN(False Negative) denotes the number of actual positive samples but pre- dicted negative samples, that is, the number of predicted non-defective objects but actual defective objects in the detection process.

For a category, the AP of the category can be calculated by plotting its P-R curve and calculating the area enclosed by the curve and the axes. mAP@.5 is used as the evaluation standard, which indicates the comprehensive detection capability of the model when the IOU threshold is set at 0.5.

(2) GFLOPs: Giga Floating-point Operations Per Second, that is, one billion floating-point operations per second, which is usually used to measure the complexity of the model.

The number of parameters: the sum of the number of parameters in the model, which does not directly affect the inference speed of the model, but the number of parameters affects the memory occupation of the model.

3.4 Implementation Details

In the experiments, our model algorithm is implemented through the PyTorch deep learning framework, PyTorch ver- sion is 1.12.1, and the training and testing are performed on GPU Nvidia RTX A5000. During training, the model training period is 140, the image size is 640 × 640, the batch size is 64, the initial learning rate is 0.01, and the optimizer is SGD.

3.5 Ablation Experiment

Our method performed a series of ablation experiments on the test set to verify the effectiveness and superiority of each improvement point in the algorithm. ✓indicates the use of this modular approach. The results of the experiments are shown in Table 4.

The experimental results show that each of the improvement points proposed by our method can improve the performance of the algorithm. In Scheme 1, our method adds deformable convolutional structures to the network. The experimental results show that mAP@.5 increased by 5.1 percentage points, accuracy increased by 1.3 percentage points, and recall rate increased by 1.8 percentage points. In Scheme 2, our method uses the context enhancement module as the feature fusion module. The experimental results show that mAP@.5 in- creased by 3.1 percentage points. CBAM In Scheme 3, our method adds CBAM attention mechanism to improve the spatial interaction between defective texture features and the experimental results show that mAP@.5 increased by 2.5 percentage points. Therefore, the use of DCNv3, CAM and CBAM attention mechanism are all able to effectively improve the detection of defects by the model. The improvement effect of scheme 1 is more obvious compared to scheme 2 and scheme 3, which indicates that the interference of similar texture features is the main reason for the poor detection of road defects. Our method improves 7.8 percentage points and the accuracy improves 0.2 percentage points compared to YOLOv5s. the experimental results show the effectiveness of the algorithm of this study in the target detection of road defects, and the algorithmic model effectively improves the accuracy of road defects detection.

3.6 Comparison Experiment

Our method compared the model with DiffusionDet[29], DINO[30], YoLoX[31] and u-YoLo[32] to demonstrate that our proposed model can effectively solve the problem of detecting attention away from the ground as well as multi- scale characteristics, The DiffusionDet diffusion model is to add noise to the input gradually, and learns how to recover the input out of the noise; DINO((DETR with Improved denoising anchor boxes) is an improved version of the DETR model using ResNet - 50 as the backbone; YOLOX uses a model similar to the YOLOv5s model parametric number of small models of similar size; and u-YoLo is the winning solution of the 2020 IEEE Global Road Defect Detection Competition, which combines the integrated model (EM) and the integrated prediction (EP) methods, however, the method is the worst in terms of speed of detection due to the long testing time for each image. According to the original paper the above method is reproduced using our dataset and the same hyperparameters are used during the experiments, the experimental results are shown in Table 5.

Based on the experimental results, it can be seen that our proposed road defect detection model performs better than other models.Compared with the encoder-decoder Diffusion- Det and DINO models, the mAP of our model is increased to 64.7%, while the number of parameters and GFLOPs are reduced to 13.5M and 20.2,respectively. Compared with the YOLOX model, the mAP of our model is improved by 13.2 percentage points, which is attributed to the fact that YOLOX adopts the anchor- free mechanism, compared to the

anchor- based mechanism of YOLOv5, whose detection mechanism is less effective for the road defect detection task. Compared with the u-YOLO model, our model improves the mAP by 11.4 percentage points. In summary, our proposed road defect detection model has a good performance in detection.

3.7 Visualisation

Our method is compared by thermal map[33] and confusion matrices to further prove the effectiveness of our algorithm. Fig. 13 shows the confusion matrix obtained by YOLOv5s network model and our proposed network model on the test set, and the results show that our proposed algorithm has higher accuracy and lower error rate and missed detection rate.

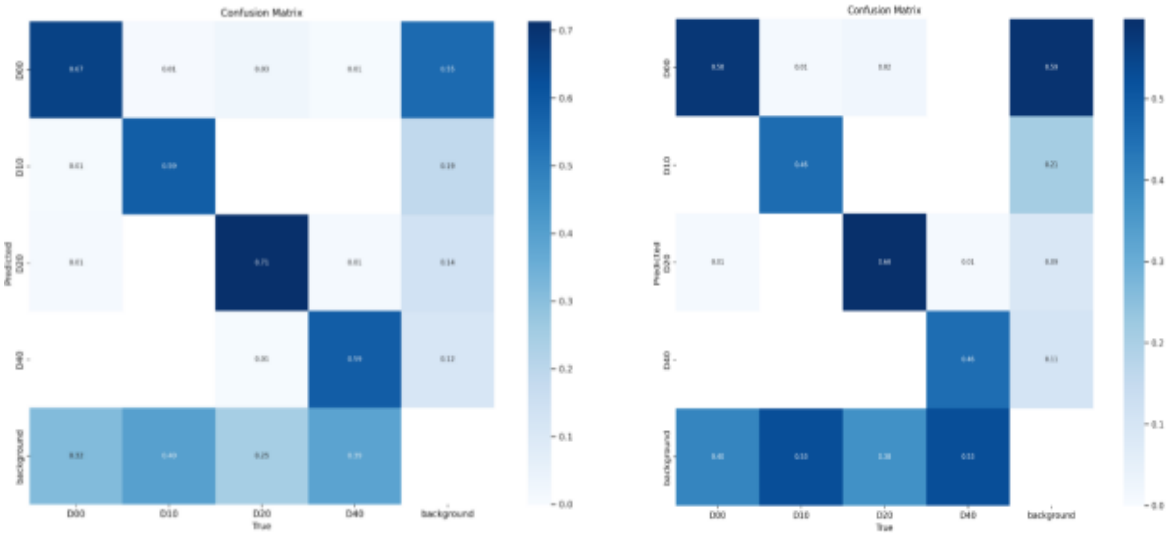

Figure 13: Confusion Matrix Comparison

Fig. 14 shows the thermal map comparison between YOLO 5s algorithm and our proposed algorithm. It can be observed that Our proposed algorithm is able to focus the road defect information better, has higher sensitivity to target detection and better performance.

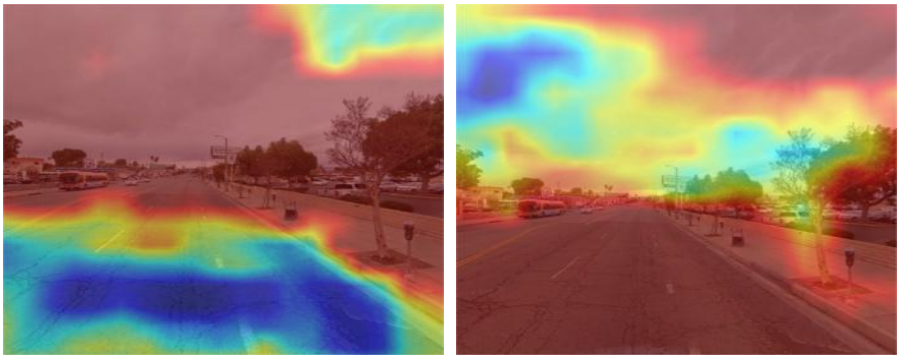

Figure 14: Comparison of heat maps

4 Conclusion

We proposes a road defect detection method based on deformable convolution and contextual enhancement to solve the problem of difficult to distinguish the similarity between texture features and multi-scale characteristics in road defect detection. In our method, the

deformable convolution module

Table 4: DCNv3 does not pass the experimental results of the combinatorial approach

Module	DCN	CAM	CBAM	mAP@.5	Precision	Recall	Parameters	GFLOPs
YOLOv5s				56.9%	62.2%	54.9%	7.03M	16.0
Scheme 1	✓			62.0%	63.5%	56.7%	6.31M	14.4
Scheme 2		✓		60.0%	62.1%	54.3%	13.9M	20.5
Scheme 3			✓	59.4%	60.3%	54.3%	6.7M	14.9
Scheme 4		✓	✓	61.5%	63.3%	53.0%	13.9M	20.5
Scheme 5	✓		✓	62.8%	61.8%	57.8%	6.32M	14.4
Scheme 5	✓	✓		62.8%	61.8%	57.8%	6.32M	14.4
Ours	✓	✓	✓	64.7%	62.4%	53.6%	13.5M	20.2

Table 5: COMPARISON EXPERIMENT

Module	mAP@.5	Parameters	GFLOPs
DiffusionDet	40.2%	40M	225
DINO	40.6%	47M	279
YOLOv5X	51.55%	9.0M	20.8
u-YOLO	53.3%	7.2M	20.64
Ours	64.7%	13.5M	20.2

DCNv3 is used in the feature extraction stage, which can effectively extract more abundant and more effective road defect features, so as to improve the detection effect of defect features of multiple shapes; CAM context enhancement module is used in the feature fusion stage, which improves the model's ability to extract defect features of different scales, and further strengthens the detection effect of defects of different scales; and finally, the CBAM attention mechanism is introduced to strengthen the spatial information interaction in the texture feature extraction process, which effectively improves the detection effect of multiple defect features. The experimental results based on the image dataset released by IEEE 2022 Global Road Damage Challenge show that our method achieves superior results and has higher detection performance compared to current cutting-edge methods.

In road defect detection, the main applications are road evaluation and maintenance and autonomous driving scenarios. In the automatic driving scenario, real-time detection becomes especially important. How to solve the detection model with many parameters and large size, which is difficult to meet the real-time requirements of defect detection, will be the main content of future research this In addition, due to the characteristics of tiny targets with low resolution and small size, it is difficult to carry out accurate detection, which will also become another research focus.

Acknowledgements

This work was supported by the National Scholarship of China, No.202308330336.

References

[1] Ma N, Fan J, Wang W, et al. Computer Vision for Road Imaging and Pothole Detection:

A State-of-the-Art Review of Systems and Algorithms. arXiv e-prints. 2022 Apr;:arXiv: 2204.13590.

[2] Kim YM, Kim YG, Son SY, et al. Review of recent automated pothole-detection methods. Applied Sciences. 2022;12(11):5320.

[3] Fan R, Liu M. Road damage detection based on unsuper- vised disparity map segmentation. IEEE Transactions on Intelligent Transportation Systems. 2019;21(11): 4906– 4911.

[4] Devine R. City of san diego asking residents to report potholes. URL: shorturl at/gnLPV. 2017.

[5] O'Donnell N, McConomy K. Jaquar land rover an- nounces technology research project to detect, predict and share data on potholes'. URL: shorturl at/btKS2. 2015;.

[6] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Nether- lands, October 11–14, 2016, Proceedings, Part I 14; Springer; 2016. p. 21–37.

[7] Gupta S, Sharma P, Sharma D, et al. Detection and lo- calization of potholes in thermal images using deep neu- ral networks. Multimedia tools and applications. 2020; 79:26265–26284.

[8] Kortmann F, Talits K, Fassmeyer P, et al. Detecting various road damage types in global countries utilizing faster r-cnn. In: IEEE International Conference on Big Data 2020; 2020.

[9] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. In: Computer Vision Pattern Recognition; 2016.

[10] Baek JW, Chung K. Pothole classification model using edge detection in road image. Applied Sciences. 2020; 10(19):6662.

[11] Suong LK, Jangwoo K. Detection of potholes using a deep convolutional neural network. Journal of Universal Computer Science. 2018;24(9):1244–1257.

[12] Redmon J, Farhadi A. Yolo9000: Better, faster, stronger. In: IEEE Conference on Computer Vision Pattern Recognition; 2017. p. 6517–6525.

[13] Shim S, Kim J, Lee SW, et al. Road surface dam- age detection based on hierarchical architecture using lightweight auto-encoder network. Automation in Con- struction. 2021;130:103833.

[14] Ukhwah EN, Yuniarno EM, Suprapto YK. Asphalt pave- ment pothole detection using deep learning method based on yolo neural network. In: 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA); 2019.

[15] Redmon J, Farhadi A. Yolov3: An incremental improve- ment. arXiv e-prints. 2018.

[16] Hegde V, Trivedi D, Alfarrarjeh A, et al. Yet another deep learning approach for road

damage detection using en- semble learning. In: 2020 IEEE International Conference on Big Data (Big Data); 2020.

[17] Sheta AF, Turabieh H, Aljahdali S, et al. Pavement crack detection using convolutional neural network. In: Computers and Their Applications; 2020.

[18] Li H, Zong J, Nie J, et al. Pavement crack detection algo- rithm based on densely connected and deeply supervised network. IEEE Access. 2021;PP(99):1–1.

[19] Wu X, Ma J, Sun Y, et al. Multi-scale deep pixel distribu- tion learning for concrete crack detection. In: 2020 25th International Conference on Pattern Recognition (ICPR); 2021.

[20] Liu Z, Cao Y, Wang Y, et al. Computer vision- based concrete crack detection using u-net fully con- volutional networks. Automation in Construction. 2019; 104(AUG.):129–139.

[21] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks. In: Proceedings of the IEEE international con- ference on computer vision; 2017. p. 764–773.

[22] Zhu X, Hu H, Lin S, et al. Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 9308–9316.

[23] Wang W, Dai J, Chen Z, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Confer- ence on Computer Vision and Pattern Recognition; 2023. p. 14408–14419.

[24] Xiao J, Zhao T, Yao Y, et al. Context augmentation and feature refinement network for tiny object detection. 2021;.

[25] Woo S, Park J, Lee JY, et al. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 3–19.

[26] Arya D, Maeda H, Ghosh SK, et al. Rdd2022: A multi- national image dataset for automatic road damage detec- tion. arXiv preprint arXiv:220908538. 2022.

[27] Arya D, Maeda H, Ghosh SK, et al. Crowdsensing-based road damage detection challenge (crddc'2022). In: 2022 IEEE International Conference on Big Data (Big Data); IEEE; 2022. p. 6378–6386.

[28] Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. IEEE Transactions on systems, man, and cybernetics. 1973;(6):610–621.

[29] Chen S, Sun P, Song Y, et al. Diffusiondet: Diffu- sion model for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vi- sion; 2023. p. 19830–19843.

[30] Zhang H, Li F, Liu S, et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:220303605. 2022.

[31] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:210708430. 2021.

[32] Hegde V, Trivedi D, Alfarrarjeh A, et al. Yet another deep learning approach for road damage detection using en- semble learning. In: 2020 IEEE International Conference on Big Data (Big Data); IEEE; 2020. p. 5553–5558.

[33] Selvaraju RR, Cogswell M, Das A, et al. Grad- CAM: Visual Explanations from Deep Networks via Gradient-based Localization. arXiv e-prints. 2016 Oct;:arXiv:1610.02391.