Received: 16 March 2025 | Accepted: 10 August 2025

DOI: 10.1111/mice.70050

RESEARCH ARTICLE

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

A data augmentation method for pavement crack detection based on super-resolution and denoising diffusion probabilistic models

Hui Yao[1] | **Yanhao Liu**[1] | **Svetlana Besklubova**[2,3] | **Ioannis Brilakis**[2] | **Meng Guo**[4] | **Jin Wang**[1] | **Min Wang**[5]

[1]Beijing Key Laboratory of Traffic Engineering, College of Metropolitan Transportation, Beijing University of Technology, Beijing, China

[2]Department of Engineering, University of Cambridge, Cambridge, United Kingdom

[3]Academy of Silesia, Katowice, Poland

[4]The Key Laboratory of Urban Security and Disaster Engineering of the Ministry of Education, Beijing University of Technology, Beijing, China

[5]Department of Statistics and Data Science, The University of Texas, San Antonio, Texas, USA

Correspondence
Yanhao Liu, Beijing Key Laboratory of Traffic Engineering, College of Metropolitan Transportation, Beijing University of Technology, Beijing, China.
Email: liuyh0929@emails.bjut.edu.cn

Svetlana Besklubova, Department of Engineering, University of Cambridge, Civil Engineering Building, 7a JJ Thomson Ave, Cambridge, CB3 0FA, United Kingdom.
Email: sb2837@cam.ac.uk

Funding information
Hunan Expressway Group Co. Ltd.; Transportation Science and Technology Progress and Innovation Program of Hunan Province, Grant/Award Number: 202152; Beijing's high-level overseas talents and Key Scientific Research Projects of BBMG Corporation, Grant/Award Number: KYJC018; Engineering and Physical Sciences Research Council (EPSRC), Grant/Award Number: EP/V056441/1

Abstract

Automated detection of pavement cracks is a task of wide interest. With the improvement of industrialization, high-resolution (HR) images are increasingly favored by researchers due to their ability to provide rich information about pavements and diseases. However, the acquisition of effective training data is not easy, which affects the accuracy and robustness of the detection model. Although the recently emerged denoising diffusion probabilistic model (DDPM) overcomes the inherent pattern collapse problem of generative adversarial networks and is capable of generating more diverse and realistic pavement data, its high sampling cost hinders the generation of HR images with rich texture information. To overcome this limitation, this paper proposed a low-cost, two-step data augmentation method that combines DDPM with super-resolution. The method first generated small-sized pavement crack images using DDPM and then enhanced resolution and texture details using an improved SwinIR model. The resulting HR and diverse crack images were used to augment the dataset. The effectiveness of the proposed method was evaluated using four state-of-the-art object detection models. Experimental results showed that all models trained with the augmented training dataset exhibited better performance. Furthermore, when combined with geometric transformation techniques, the proposed method was able to improve the crack detection accuracy by up to approximately 12%.

1 | INTRODUCTION

Cracks are early indicators of pavement damage, and if not repaired on time, they can lead to further pavement structure deterioration, increased maintenance costs, and potential traffic accidents (Ai et al., 2023). As a result, the timely detection of pavement cracks is crucial. Traditionally, manual and visual inspection is labor-intensive, time-consuming, and susceptible to environmental conditions. In addition, with rapid urbanization, the traditional method struggles to meet the growing demand for large-scale, high-efficiency inspections. Therefore, exploring automated and intelligent pavement crack detection technologies is essential for extending road service life, ensuring driving safety, and optimizing resource allocation.

Traditional crack detection mainly uses image processing algorithms, such as threshold segmentation (Oliveira & Correia, 2009) and edge detection algorithms (Zhao et al., 2010). Despite the simplicity of the process, these methods are less robust. In recent years, the rise of deep learning technology not only promotes the development of medical treatment (Mammone et al., 2023; Nogay & Adeli, 2023, 2024; Perez-Sanchez et al., 2024), robot development (Nguyen & Cheahl, 2022), and agricultural detection (Dhanya et al., 2022) but also plays an important role in traffic engineering (Kashyap et al., 2022) and civil engineering (Z. Li & Adeli, 2018; Perez-Ramirez et al., 2019; Rafiei & Adeli, 2018). As a result, automatic crack detection based on deep learning has increasingly gained popularity. Deep learning methods tend to provide more robust and accurate detection results and show great advantages in representing complex pavement features. Existing studies primarily consist of image-level classification methods, pixel-level classification algorithms, and object detection methods. The image-level classification methods (L. Zhang et al., 2016) use deep learning models to categorize pavement images as either diseased or disease-free. Although such methods have high accuracy, they are unable to locate crack locations. In contrast, object detection methods (Yi et al., 2023) can classify and roughly locate pavement cracks with faster detection speeds. However, the inability of such methods to quantify pavement cracks hinders practical applications. Pixel-level classification algorithms (J. Chen & He, 2022; Huang et al., 2024; Zhu et al., 2024) can identify every crack pixel, providing stronger support for crack quantification and practical applications but at the expense of detection efficiency. To address this issue, a two-step approach (J. Liu et al., 2020) was proposed. It successfully achieves faster crack segmentation by feeding the output detection boxes of the object detection model into the pixel-level classification model. Therefore, it is especially critical to improve the performance of object detection models that provide candidate targets.

Sufficient data enable the model to better capture features and uncover patterns, significantly enhancing the accuracy of tasks like prediction and classification. However, obtaining pavement crack images can be challenging, making the use of data augmentation methods essential. Traditional data augmentation methods, such as rotation and flipping, have been adopted in several pavement crack detection studies (C. Wang et al., 2024; Yao et al., 2022, 2024) and led to better training results. However, the samples generated by such methods tend to be relatively homogeneous, which cannot provide more crack patterns and have a limited effect in improving the model's generalization ability. In 2014, a generative adversarial network (GAN) was pioneered (Goodfellow et al., 2014), and it can generate more diverse and realistic samples through adversarial training of generators and discriminators. Therefore, various variants of GAN have been developed and applied to the generation of pavement crack images. T. Zhang et al. (2024) used Wasserstein's GAN (wGAN) to generate realistic pavement images and proposed the class-specific image augmentation (CSIA) method to solve the category imbalance problem. Zhong et al. (2023) combined an improved wGAN with a Poisson image hybrid algorithm to alleviate the problem of dataset scarcity in trench cement concrete pavement crack detection. Lu et al. (2023) proposed a GAN-based framework for pavement texture regeneration, which effectively solves the problems of texture information loss and non-convergence training through improved data augmentation and M-sigmoid normalization methods. Shim (2024) utilized a trained conditional GAN to synthesize real crack images from labeled or predicted images and combined it with a self-training method to improve crack segmentation performance. Despite the excellent performance of GANs in image generation, the issue of pattern collapse (W. Li et al., 2021) remains a challenge, limiting their practical application. Additionally, GANs require significant expertise and experience from the trainers.

In 2020, Ho et al. (2020) proposed a denoising diffusion probabilistic model (DDPM), which is capable of generating realistic data through a step-by-step denoising process. Compared to GAN, DDPM has significant advantages in training stability, sample diversity, and generation quality (Yu et al., 2024). Shen et al. (2024) showed that DDPM was suitable for generating structural damage image data of varying complexity and could significantly improve the performance of detection models. However, this study also pointed out that the disadvantage of DDPM is the low sampling efficiency. To address the limited availability of pavement crack image data, Cano-Ortiz et al. (2024) first applied DDPM to the study of data augmentation for

pavement damage detection and demonstrated that the method was able to improve the mAP@0.5 of YOLOv5l from 0.604 to 0.622. In 2024, Hancheng Zhang et al. (2024) proposed a crack diffusion model. By adjusting the conditional inputs of cracks, the location and morphology of cracks can be better controlled. The generated images significantly improved the effectiveness of detection and segmentation. While DDPM demonstrates strong performance in generating pavement images, its excessive sampling cost—due to the gradual denoising sampling process—limits its ability to generate high-resolution (HR) images with rich texture information in practical applications. Although the latent diffusion model (Rombach et al., 2022) performs diffusion operations in latent space to address this problem, it has great difficulty in reconstructing the complex texture of the pavement using a variational autoencoder, and the training cost is large.

Super-resolution (SR) method is an emerging technique to improve image quality at a low computational cost. Yuan et al. (2022) proposed an SR reconstruction method based on an improved GAN and demonstrated that image reconstruction is effective for detection and segmentation tasks. Xiang et al. (2022) and Oh et al. (2025) used SR to solve the problem of insufficient resolution in concrete crack images, which improves the accuracy of semantic segmentation methods. Therefore, combining DDPM with SR presents a promising approach for low-cost HR image generation, especially in the field of automatic pavement crack detection. In this paper, a two-step data augmentation method was proposed based on this concept. As shown in Figure 1, a small-sized pavement crack image is first generated using DDPM. Then, ×2 SR is applied using the improved SwinIR model, supplementing pavement texture details. Finally, the generated diverse HR crack images are combined with real images to form an augmented dataset. The effectiveness of the proposed approach is evaluated with several state-of-the-art object detection models. The objective of this study is to propose a low-cost, HR pavement crack image generation method that enhances the precision and generalization ability of crack detection models from a data perspective.

2 | METHODOLOGY

2.1 | DDPM

2.1.1 | Network architecture

DDPM is a generative model based on the diffusion process, which starts with Gaussian noise during sampling and generates new samples after several denoising iterations. As in Figure 2, DDPM mainly includes the forward diffusion process and the reverse denoising process. The diffusion process transforms the original image x_0 into a purely noisy image x_T by successively adding Gaussian noise, and the image x_t at any time step can be represented by x_{t-1}, as in Equation (1).

$$x_t = \sqrt{1-\beta_t} * x_{t-1} + \sqrt{\beta_t} * \epsilon_{t-1} \tag{1}$$

where β_t is used to control the quantity of noise added at each diffusion step. Specifically, β_t determines how much the image changes relative to the previous time step at time step t, thus affecting the accumulation of noise; ϵ_{t-1} is the Gaussian noise at time step $t-1$; x_t can be obtained directly from x_0 through the use of the reparameterization trick as in Equation (2).

$$x_t = \sqrt{\bar{\alpha}} * x_0 + \sqrt{1-\bar{\alpha}} * \epsilon \tag{2}$$

where $\bar{\alpha}$ is a hyperparameter, which is calculated according to Equations (3) and (4); ϵ is a standard Gaussian noise.

$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i \tag{3}$$

$$\alpha_t = 1 - \beta_t \tag{4}$$

The inverse process is the process of removing noise, which generates new data samples by predicting and removing noise. The inverse process can be expressed as Equation (5).

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta (x_t, t) \right) + \sigma_t z \tag{5}$$

where $\epsilon_\theta(x_t, t)$ represents the noise predicted by the neural network at time step t; z denotes the noise from a standard Gaussian distribution; and σ_t is the predefined variance at time step t, which is calculated according to Equation (6).

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t \tag{6}$$

The neural network used for noise prediction is based on U-Net (Ronneberger et al., 2015). This network consists of a down-sampling path, an intermediate layer, an up-sampling path, and skip connections, and the architecture is shown in Figure 3. The down-sampling path is responsible for extracting image features and includes three down-sampling layers and eight ResBlocks (He et al., 2016), each containing two convolutional layers with 3 × 3 kernels. The intermediate layer integrates the extracted features and consists of two ResBlocks. The up-sampling path is used to increase the image size and includes three up-sampling layers with twelve ResBlocks. To improve

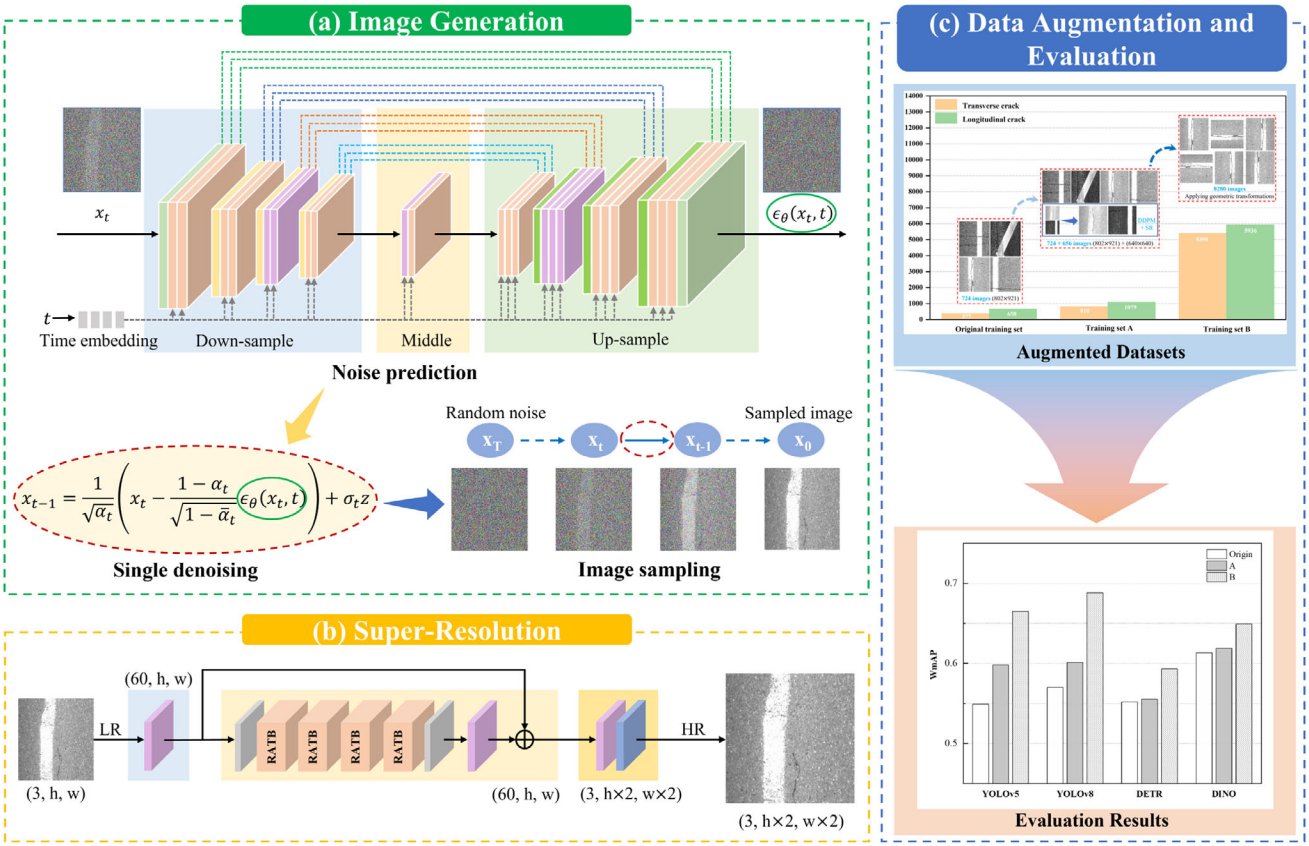

FIGURE 1 Flowchart of the proposed method.

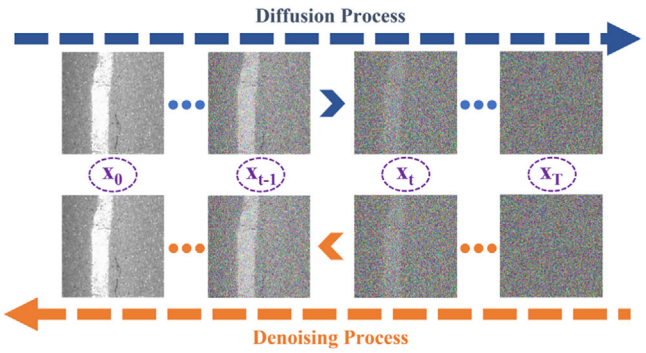

FIGURE 2 Diffusion and denoising process of denoising diffusion probabilistic model (DDPM).

model performance, the first ResBlock of the intermediate layer, as well as the ResBlocks at the third scale of both the down-sampling and up-sampling paths, is enhanced with self-attention (Vaswani et al., 2017). Additionally, the time embedding is introduced into each ResBlock by using a predefined embedding layer to process the time step and sum it with the output of the first convolutional layer of the ResBlock and feeding it into its second convolutional layer so as to realize the prediction of noise at different time steps. The model's loss function is defined as Equation (7).

$$\text{Loss} = \left\| \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t} * x_0 + \sqrt{1-\bar{\alpha}_t} * \epsilon, t\right) \right\|^2 \quad (7)$$

where $\epsilon_\theta(\sqrt{\bar{\alpha}_t} * x_0 + \sqrt{1-\bar{\alpha}_t} * \epsilon, t)$ denotes the noise predicted by the neural network at time step t; ϵ denotes the actual added noise. By reducing the error between the prediction and the actual added noise, the DDPM can reproduce the original image from the noise.

2.1.2 | Training settings

In this paper, the total number of time steps T is 1000, and β_t conforms to a linear growth starting from a small initial value of 0.0001 and gradually increasing to a final value of 0.02. In addition, AdamW (Loshchilov & Hutter, 2017) was employed as the optimizer, and a weight decay of 0.0001 was set. The learning rate lr decreases as the training proceeds as shown in Equation (8) during the training process.

$$lr = lr_{\min} + 1/2\,(lr_{\max} - lr_{\min})\left(1 + cos\left(\frac{e}{E}\pi\right)\right) \quad (8)$$

where $lr_{\min}$ represents the minimum learning rate, which is 0 in this paper; $lr_{\max}$ represents the maximum learning rate, which is 0.0001; e denotes the current training epoch; and E represents the total training epoch, which is 240.

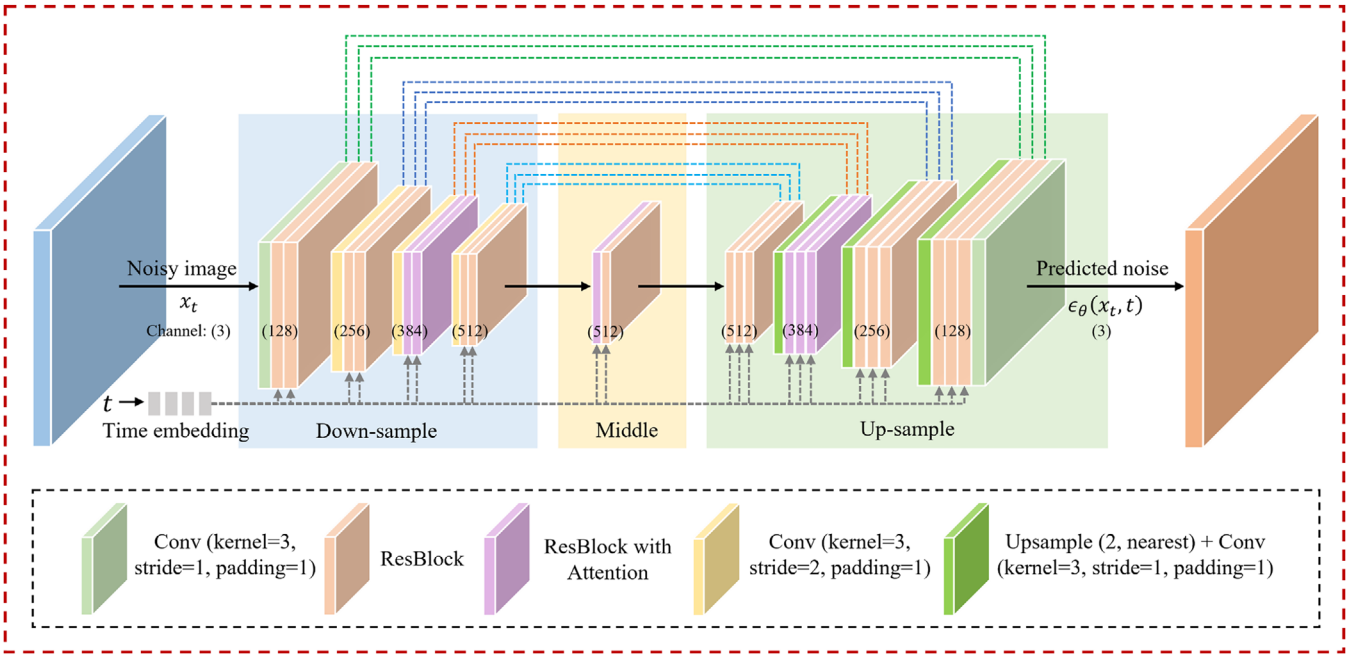

FIGURE 3 Architecture of U-Net for predicting noise.

2.2 | SR reconstruction

2.2.1 | Network architecture

The aim of SR is to increase the resolution of the low-resolution (LR) image to make it clearer while ensuring no degradation in image quality as much as possible. Since image super-resolution using convolutional neural network (SRCNN) (Dong et al., 2015) pioneered the use of deep learning for SR tasks, methods such as residual block (Lim et al., 2017), generative adversarial training (Ledig et al., 2017), dense block (Y. Zhang et al., 2018), and other methods have been successively introduced to enhance performance. However, limited by the local processing of convolutional kernels, these methods perform poorly in capturing long-range dependencies. Therefore, Transformer models have attracted attention. In 2021, Image restoration using swin transformer (SwinIR) (Liang et al., 2021), based on Swin Transformer (Z. Liu et al., 2021), was proposed, which not only refreshes the accuracy of image SR tasks but also possesses a relatively smaller number of parameters. In view of this, the improved SwinIR was used in this paper to process pavement crack images.

Figure 4 shows the architecture of the proposed SR model. First, the LR images are transformed into the feature space by a convolutional layer. Next, the image features are passed through four residual attention Transformer blocks (RATB) for deep feature processing. Finally, the processed features are passed through an up-sampling module to realize the final HR image transformation. As shown in Figure 4a, RATB contains three Swin Transformer attention layers (STAL) and a 3×3 convolutional layer. Compared to the original SwinIR, the squeeze and excitation (SE; Hu et al., 2018) module is added to the STAL for global information integration, which helps the model to understand and reconstruct complex details in HR images and better handle long-term patterns (X. Chen et al., 2023). Specifically, as shown in Figure 4b, the windows multi-head self-attention (W-MSA) module is serially connected to the shifted windows MSA (SW-MSA) module within STAL, with an SE block added after the SW-MSA module. In addition, residual connections are introduced to maintain stable training.

The model's loss function is defined as Equation (9).

$$\text{Loss} = \frac{1}{n}\sum_{i=1}^{n} |y_{\text{GT}} - y_{\text{HR}}| \tag{9}$$

where y_{GT} indicates the ground truth; y_{HR} denotes the model output; n is the total pixels in the image.

2.2.2 | Training settings

In this paper, the window size of W-MSA and SW-MSA is set to 8, and six attention heads are used. Furthermore, Adam is introduced as the optimizer, and the learning rate decreases as training progresses. Specifically, the initial learning rate is set to 0.0002 and changes to half of the previous learning rate at the 30,000th and 36,000th training epochs. In this paper, the total training epoch is 42,000. In addition, during the training process, the LR images are

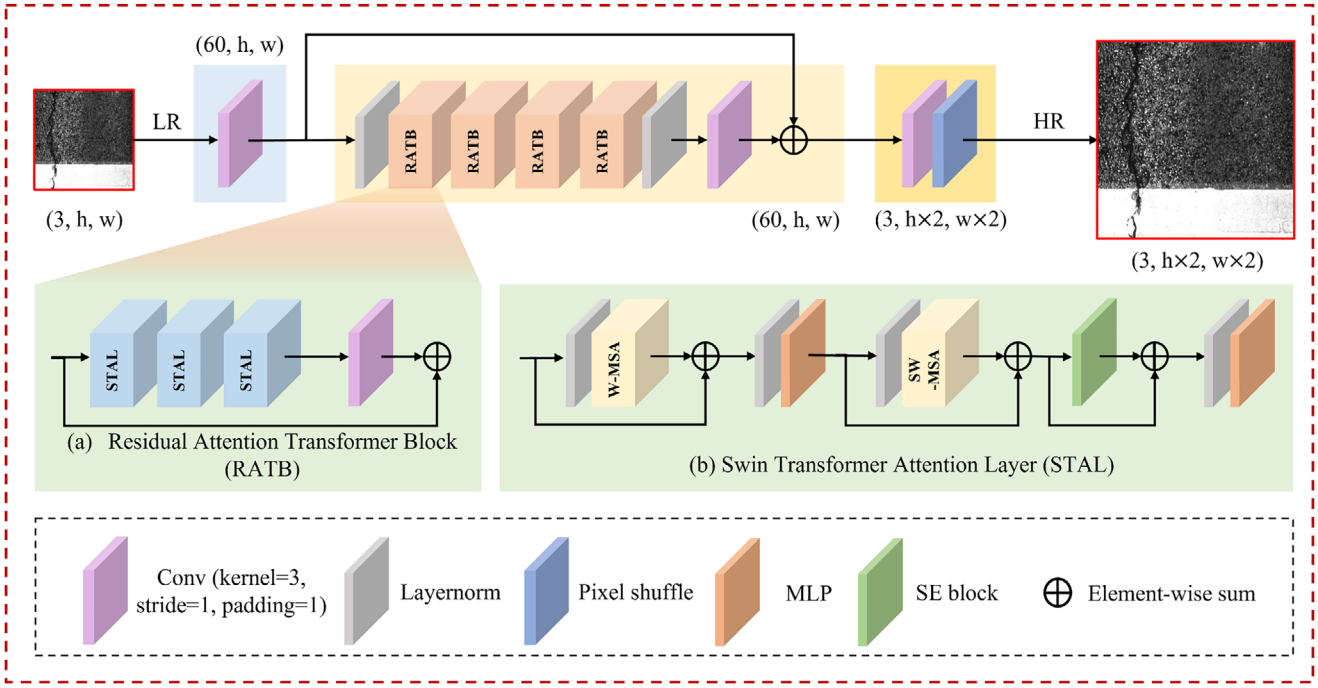

FIGURE 4 Architecture of the proposed super-resolution (SR) model.

randomly cropped to a 64 × 64 size, paired with the corresponding ground truth, and fed into the model, which is very useful to reduce the computational load.

3 | RESULTS AND EVALUATION

The data processing, network training, and evaluation were all performed on a single computer with NVIDIA GeForce RTX 3090 graphics processing unit (GPU) and an Intel Xeon Gold 6130 central processing unit (CPU).

3.1 | Datasets

The pavement crack images used were taken on a road inspection vehicle, and a total of 1206 images of size 802 × 921 were obtained. Some of the collected images are shown in Figure 5. All images were manually labeled and divided into two classes: transverse cracks and longitudinal cracks. To facilitate the subsequent object detection experiments, the images were divided into 724 training images, 241 validation images, and 241 test images.

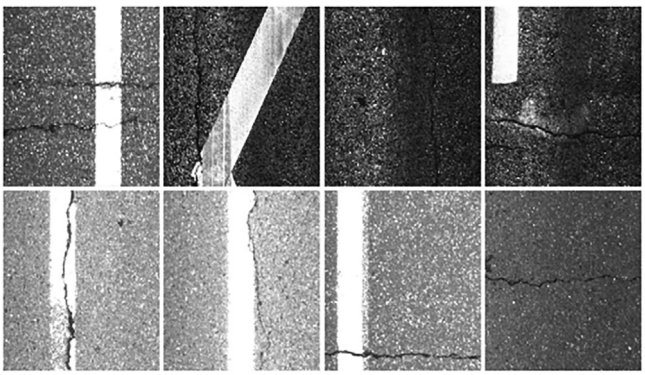

FIGURE 5 Collected pavement crack images.

3.1.1 | Datasets for DDPM

Only the training images were used for DDPM training. Since most existing object detection models scale the input image to 640 × 640 in equal proportion to their length and width, 640 × 640 (320 × 320 for DDPM) was chosen as the generated image size. In addition, the training images were first scaled to 640 × 734. Then the patches of size 640 × 640 were cropped from the top side and bottom side, and such operation can maximize the consistency of the texture information between the generated images and the scaled true images. Finally, the bicubic interpolation was applied to all patches to obtain 1448 training images of size 320 × 320.

3.1.2 | Datasets for SR

For SR training, the same scaling and cropping operations as described above were applied to all the dataset images. As a result, 1448, 482, and 482 images were used for train-

ing, validation, and testing, respectively. It is worth noting that the 640 × 640 patches are paired as ground truth with the 320 × 320 images.

3.2 | Evaluation metrics

3.2.1 | Evaluation metrics for image generation

Fréchet inception distance (FID; Heusel et al., 2017) is used to evaluate the quality of the images generated by DDPM in this paper. It measures the similarity between the generated images and the real images by calculating the difference between their distributions in the feature space. Specifically, the calculation of FID involves the average probabilistic distance between two image sets in the Inception network.

3.2.2 | Evaluation metrics for SR

Peak signal-to-noise ratio (PSNR) is used to accurately evaluate the SR performance and is calculated as Equation (10).

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{255^2}{\mathrm{MSE}}\right) \tag{10}$$

where mean squared error (MSE) is calculated as Equation (11).

$$\mathrm{MSE} = \frac{1}{3mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\sum_{l=0}^{2}[R(i,j,l) - T(i,j,l)]^2 \tag{11}$$

where R and T denote two three-channel images of size $m \times n$.

3.3 | Results of pavement crack image generation by DDPM

As the training proceeds, the images generated by DDPM gradually converge to real images. Figure 6 shows some of the images sampled by DDPM at different training stages, and it can be seen that the images generated in the early training stage are dark and unrealistic in color, with a high FID. After 180 training epochs, the generated image is almost identical in color to the real image, and the texture details have been improved. The generated image is only slightly deficient in the crack pattern, with an FID of 98.12. The final image generated by the DDPM nearly rivals the real image in both texture information and color, as well as the crack pattern, which further demonstrates the effectiveness of using the DDPM to generate images of pavement cracks.

TABLE 1 Performance of denoising diffusion probabilistic model (DDPM) and current state-of-the-art generative models.

Models	Fréchet inception distance	Parameters (params)	Floating point operations (FLOPs)
PG-GAN	75.01	18.35 M	5.71 G
StyleGANv3	42.43	23.32 M	105.00 G
DDPM	36.20	82.92 M	474.93 G

Table 1 shows the performance comparison results of the DDPM and current state-of-the-art generative models, including progressive growing of GANs (PG-GAN) (Karras et al., 2018) and StyleGANv3 (Karras et al., 2021) on the pavement crack image generation task. It should be noted that the DDPM used for comparison at this point uses 256 × 256 as the generated image size, limited by the image generation size of PG-GAN and StyleGANv3. The results show that PG-GAN generates the worst image quality with an FID of 75.01 but has the least number of parameters and computations. The DDPM achieves the highest FID of 36.20 but also has the largest computational cost. Overall, StyleGANv3 is more balanced in terms of performance and computational cost. However, as shown in Figure 7, there is a large difference between the different models in generating light-toned images. As observed from the red-labeled region in the figure, the generated images of PG-GAN and StyleGANv3 have ripple or grid-like artifacts, which lead to their lack of reality. Comparatively, the image textures generated by DDPM are more similar to real textures and thus more suitable for data augmentation for computer vision tasks.

3.4 | Results of pavement crack image SR

In this study, a comparison experiment between the proposed SR model and classical SR models including SRCNN, image super-resolution using a generative adversarial network (SRGAN), enhanced super-resolution generative adversarial networks (ESRGAN) (X. Wang et al., 2018), enhanced deep residual networks for single image super-resolution (EDSR), residual dense network for image super-resolution (RDN), and SwinIR was conducted on the same pavement crack dataset. As shown in Table 2, the PSNR of all methods except SRCNN is above 29.87 with a small gap. In addition, the convolutional neural network (CNN)-based model has a more obvious characteristic of trading computational cost for accuracy, and the improvement is limited. Comparatively, the Transformer-based SwinIR has better accuracy and

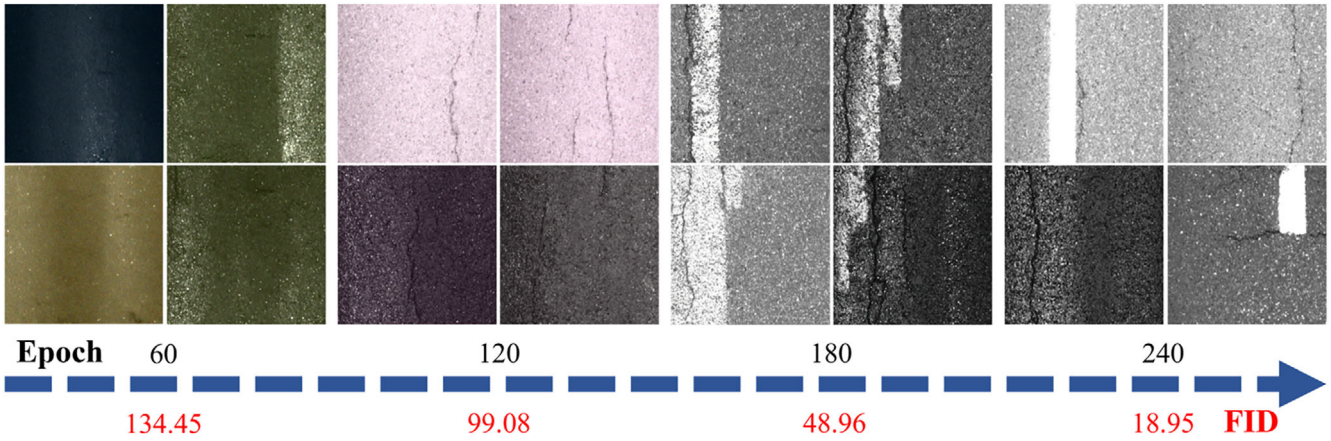

FIGURE 6 Images sampled using DDPM at different training stages. FID, Fréchet inception distance.

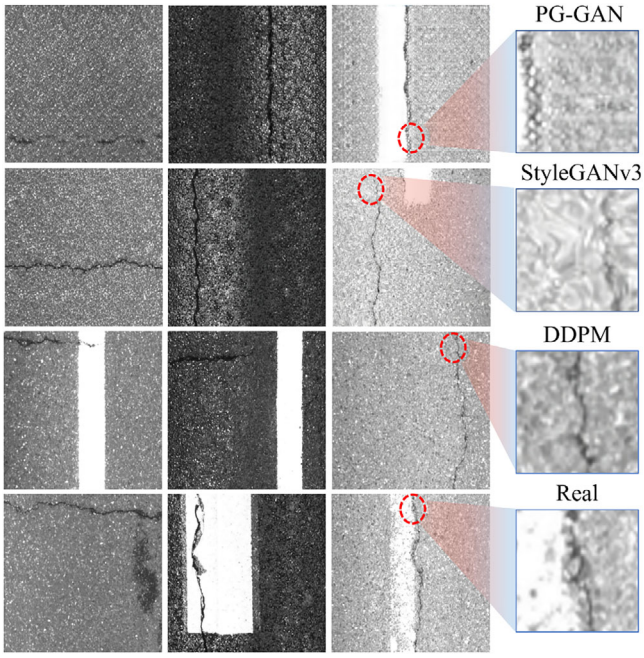

FIGURE 7 Comparison of the generation effect of DDPM with other generative models.

TABLE 2 Performance of the proposed super-resolution (SR) model and several classical SR models.

Models	SR rate	Peak signal-to-noise ratio	Params	FLOPs
SRCNN	×2	29.5555	0.02 M	8.51 G
ESRGAN	×2	29.8769	16.70 M	459.06 G
SRGAN	×2	29.9276	0.59 M	64.91 G
EDSR	×2	29.9601	1.37 M	140.56 G
RDN	×2	29.9718	22.12 M	2264.97 G
SwinIR	×2	29.9939	0.88 M	92.24 G
Ours	×2	29.9958	0.96 M	92.32 G

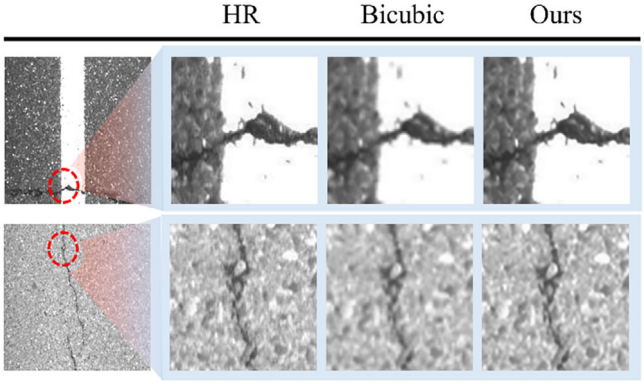

FIGURE 8 SR reconstruction effect of the proposed SR model. HR, high-resolution.

fewer parameters, and the SR model proposed in this study also has this advantage and achieves the highest PSNR of 29.9958 with almost no increase in computational cost. Although it is difficult for the human eye to perceive the subtle changes caused by the PSNR differences on an image, for computers, these differences represent different data features, which are more helpful for computers to capture the texture details in an image. Therefore, the proposed approach is more suitable for the pavement crack image SR task. As shown in Figure 8, the proposed method is effective for texture information supplementation compared to the bicubic method.

3.5 | Results of data augmentation for pavement crack detection

Six hundred and fifty-six pavement crack images of size 320 × 320 with varying backgrounds were first generated using the trained DDPM. These generated images were then upscaled using the proposed SR model to obtain 656 pavement crack images of size 640 × 640. By combining these

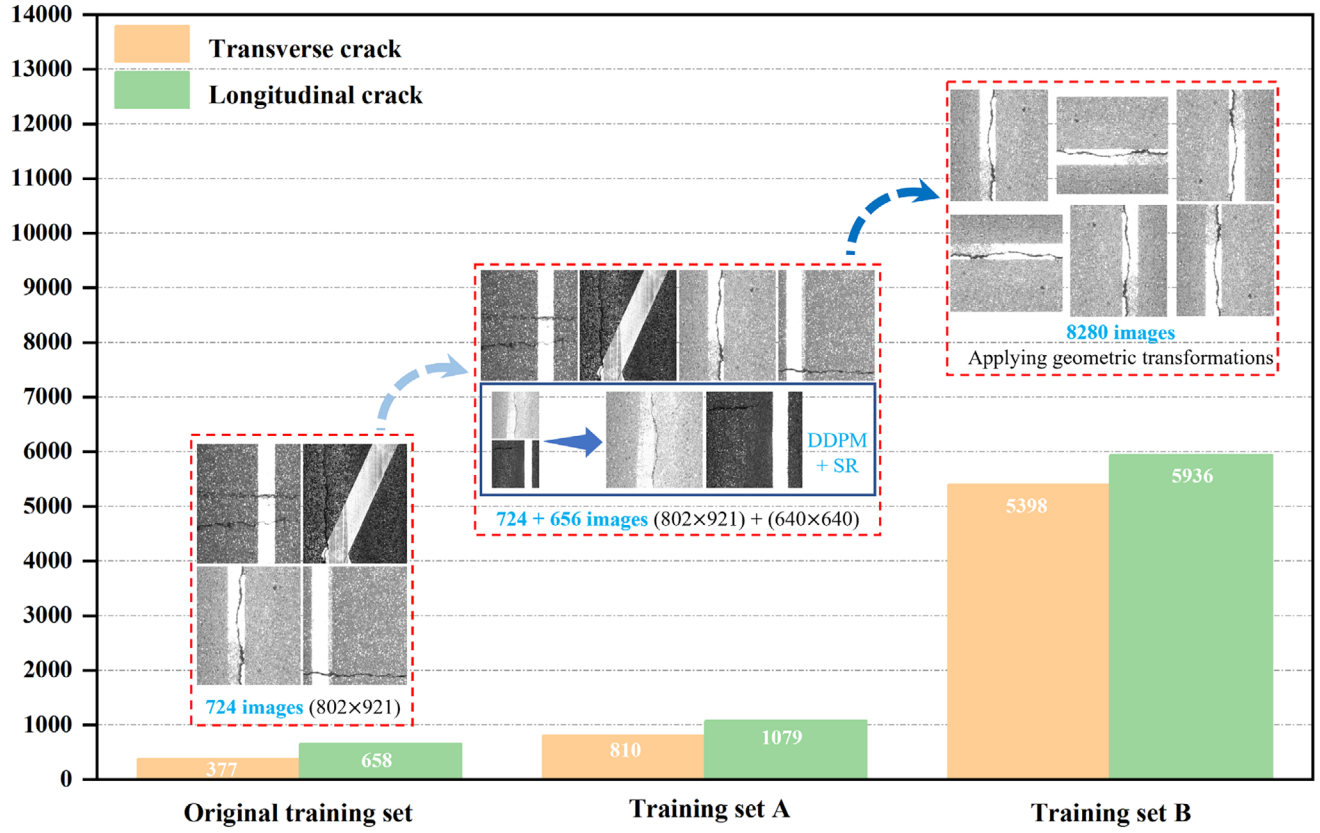

FIGURE 9 Category distribution of the original training set, augmented training set A, and augmented training set B.

generated images with the original training set, a new training set containing 1380 images was created, referred to as training set A. To further enhance the data, geometric augmentation methods—including rotations by 90°, 180°, and 270° as well as horizontal and vertical flips—were applied to training set A. This resulted in a new training set B containing 8280 images. The specific category distribution of these datasets is shown in Figure 9.

To validate the proposed data augmentation method, four current state-of-the-art object detection models, including CNN-based you only look once version 5 (YOLOv5) and you only look once version 8 (YOLOv8) (Jocher et al., 2023) and Transformer-based detection transformer (DETR) (Carion et al., 2020) and DETR with improved denoising anchor boxes (DINO) (Hao Zhang et al., 2022), were trained on the different training sets mentioned above and tested on the same test set using their best-performing weights on the validation set. During this process, each model was trained to fit different datasets. Meanwhile, each model was trained and tested three times, and the average was taken as the test result. The classical mAP@0.5 and mAP@0.5:0.95 (Yao et al., 2022) were used, and this study weighted these two metrics and used them as a composite metric, weighted mAP (WmAP), to evaluate the object detection algorithm. The WmAP is calculated as

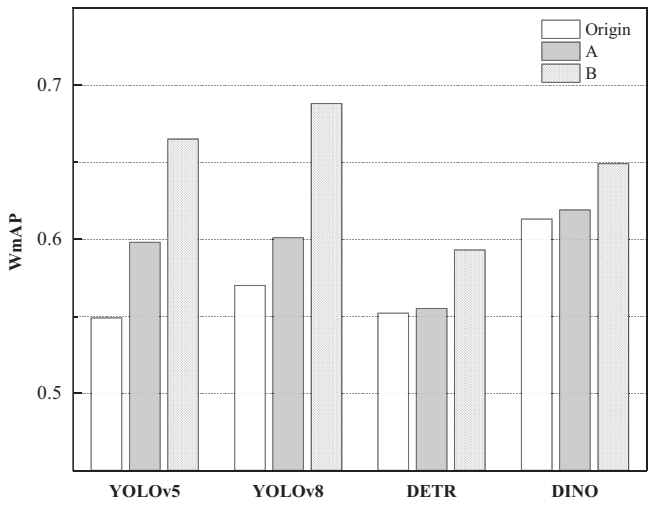

FIGURE 10 Performance of several object detection models trained using different training sets.

in Equation (12).

$$\mathrm{WmAP} = 0.1 * \mathrm{mAP@0.5} + 0.9 * \mathrm{mAP@0.5:0.95} \quad (12)$$

As shown in Figure 10, all models trained using the augmented training set outperform those trained with the

TABLE 3 Performance of several detection models trained using different training sets.

Models	Training set	mAP@0.5	mAP@0.5:0.95	WmAP
YOLOv5	Origin	0.827	0.518	0.549
YOLOv5	A	0.863	0.569	0.598
YOLOv5	B	0.878	0.641	0.665
YOLOv8	Origin	0.855	0.539	0.570
YOLOv8	A	0.860	0.572	0.601
YOLOv8	B	0.902	0.664	0.688
DETR	Origin	0.831	0.521	0.552
DETR	A	0.832	0.524	0.555
DETR	B	0.862	0.563	0.593
DINO	Origin	0.878	0.584	0.613
DINO	A	0.875	0.591	0.619
DINO	B	0.906	0.621	0.649

original training set, suggesting that the proposed method helps to improve the performance of the pavement crack detection model. Among the different augmented training sets, training set B leads to significantly greater performance improvements compared to training set A. This suggests that combining the proposed data augmentation method with traditional geometric transformations provides greater benefits. As shown in Table 3, when trained with training set B, both YOLOv5 and YOLOv8 show performance improvements of approximately 12%, while DETR and DINO only improve by 3% to 4%. This highlights that the proposed data augmentation method is more effective for CNN-based models than for Transformer-based models, which may be due to the different model architectures. In addition, on the one hand, YOLOv8 achieved the highest WmAP of 0.688, indicating that YOLOv8 has a performance advantage in pavement crack detection. On the other hand, DINO achieved a WmAP of 0.613 when trained on the original dataset, which may be due to its subtle contrast denoising training method, demonstrating its suitability for pavement crack detection in data-poor scenarios.

Figure 11 shows some of the YOLOv8 detection results with and without applying the data augmentation algorithm. By combining the gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017) visualization method, on the one hand, it can be observed that the model trained with the augmented training set B has an advantage in capturing tiny cracks (green circle) on dark images, being able to put more attention to the crack region, compared to the original model, thus avoiding false negative results. On the other hand, the original model incorrectly directs attention to interfering objects (yellow circles) that are similar to the cracks, leading to false positive results, while the improved model avoids this problem.

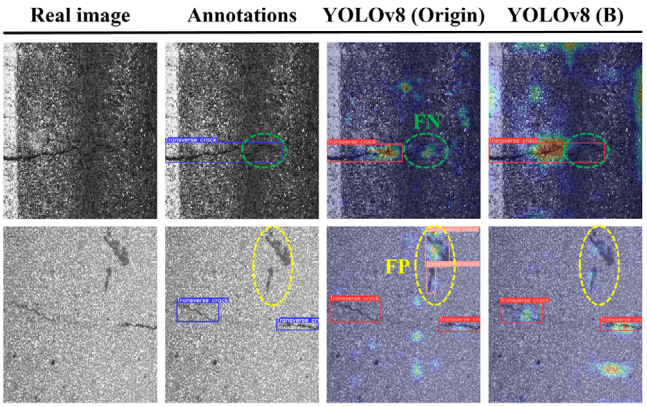

FIGURE 11 Detection results with and without applying the data augmentation algorithm.

TABLE 4 Ablation study for SR.

Models	Training set	mAP@0.5	mAP@0.5:0.95	WmAP
YOLOv8	Origin	0.855	0.539	0.570
YOLOv8	A (NSR)	0.855	0.566	0.595
YOLOv8	A	0.860	0.572	0.601
YOLOv8	B (NSR)	0.888	0.662	0.684
YOLOv8	B[a]	0.905	0.662	0.686
YOLOv8	B	0.902	0.664	0.688

Note: (NSR) represents not using SR.
[a]Represents using the SwinIR for SR.

TABLE 5 Computational cost of image generation with and without SR.

Final generated image size	Initial generated image size	SR rate	FLOPs
640 × 640	640 × 640	–	2968.313 G
640 × 640	320 × 320	×2	834.401 G

4 | DISCUSSIONS

4.1 | Ablation study on SR

In this paper, the SR model serves to improve the image resolution and complement the texture details to help the detection model learn sufficient pavement disease information. To further validate the necessity of SR, a series of ablation experiments was conducted in this study using YOLOv8. As shown in Tables 4 and 5, when SR is not applied, the performance of YOLOv8 trained using training set A and training set B decreases by 0.6% and 0.4%, respectively, which indicates that applying SR to small-sized generated images to enhance texture information is beneficial to the crack detection. When a 640 × 640 pavement crack image is generated, the use of the SR strategy can save 2133.912 G of computational cost, compared to direct generation. In summary, the use of

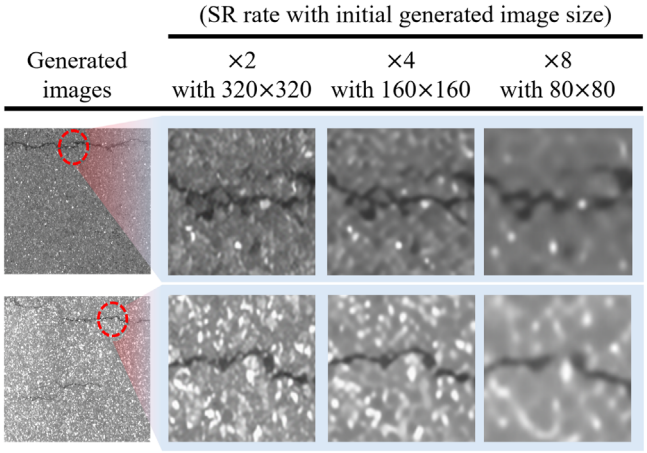

FIGURE 12 Texture of synthesized images with different generation strategies.

SR is necessary to enhance the effectiveness of data augmentation with limited computational cost. Furthermore, compared to using SwinIR, the proposed SR model is more conducive to improving the data augmentation effect, which is consistent with the inference in Section 3.4. However, the improvement is limited and further research is needed in the future to enhance it.

4.2 | Discussion on different SR rates for image generation

The above experimental results indicate that 320 × 320 image generation combined with ×2 SR rate is a low-cost and effective data augmentation method suitable for pavement crack detection tasks. However, different combinations of SR rates with corresponding initial generated image sizes have not yet been considered, and the determination of such combinations can help to further balance the computational cost and data augmentation effectiveness, which is particularly important for resource-constrained application scenes. Specifically, with the decrease of the initial generated image size, the computational cost and texture information also decrease together. As shown in Figure 12, although an SR model can complement a certain amount of texture information, it is still not able to restore the texture information of initially generated images of different sizes to the same level. Thus, whether these lost details are important for the object detection task becomes a critical issue.

Experiments were conducted using YOLOv8 to determine the optimal combination of the initial generated image size and the corresponding SR rate. Meanwhile, the entropy of the gray-level co-occurrence matrix of the final generated images with different strategies was calculated

TABLE 6 Performance of data augmentation strategies using different SR rates.

Models	Training set	Entropy	WmAP
YOLOv8	Origin	49.240[a]	0.570
YOLOv8	B (×8) (NSR)	34.415	0.674
YOLOv8	B (×8)	37.903	0.678
YOLOv8	B (×4) (NSR)	40.472	0.671
YOLOv8	B (×4)	43.230	0.680
YOLOv8	B (×2) (NSR)	44.804	0.684
YOLOv8	B (×2)	47.665	0.688

Note: (NSR) represents not using SR.
[a]Represents the entropy of real images.

TABLE 7 Performance of YOLOv8 using both B (×2) and B (×8) strategies.

Models	Training set	WmAP
YOLOv8	B (×8)	0.678
YOLOv8	B (×2)	0.688
YOLOv8	B (×2) and B (×8)	0.690

to characterize the richness of the texture information. As shown in Table 6, the richer the texture information of the generated image is, the higher the detection performance gain brought by the data augmentation strategy, which also verifies the necessity of the SR strategy. Second, by observing the performance of all models not using SR, it can be found that the model performance gain brought by SR in the B (×4) strategy is the largest, which is 0.9%, while the gain brought by SR in the B (×2) and B (×8) strategies is the same, which is 0.4%. This suggests that for the dataset used in this paper, the texture information supplemented by the SR step in the B (×4) strategy is more critical, and thus the identification of this critical texture information is the key to improving the detection performance, which needs to be further explored in the future. Furthermore, it is counterintuitive that the B (×8) strategy achieves comparable performance to the B (×4) strategy, which implies that the main factor for improving the detection performance is not the texture information in this scale of image. One possibility is that the images generated by the B (×8) strategy are able to provide information about the crack skeleton at another scale of the detection model, which enhances the learning performance of the model from a different scale. The results in Table 7 show that the combination of the B (×2) strategy and the B (×8) strategy can further improve the performance of the model, which further validates the above point, but the generation cost is too large. Overall, the B (×2) strategy has the best performance, while the B (×8) strategy strikes a better balance between training cost and performance, and is more suitable for scenarios with limited computational cost.

4.3 | Time cost issue of image generation and additional annotation

Although combining DDPM with an SR model helps reduce the high time cost of image generation, it remains inadequate for scenarios with strict computational constraints. Therefore, future research should focus on developing a lightweight DDPM for pavement crack detection. In addition, since DDPM cannot generate image annotations simultaneously, the generated images still require manual labeling, further increasing the time cost. Thus, another key focus of future research is designing a DDPM capable of generating both images and their corresponding annotations.

5 | CONCLUSION

A two-step data augmentation approach combining DDPM and SR for pavement crack detection was proposed in this paper. First, DDPM was used to generate 320 × 320 pavement crack images. Then, an improved SR model based on SwinIR upscaled the images by a factor of two while enhancing texture details. The obtained 640 × 640 HR pavement crack images were composed with real images to enhance the dataset, thus improving the accuracy and robustness of crack detection at a lower cost. The main conclusions are shown as follows:

1. The proposed two-step data augmentation method is effective on YOLOv5, YOLOv8, DETR, and DINO and can improve the performance of YOLOv8 by approximately 12% when combined with geometric transformation methods.
2. The proposed SR model performs better in restoring the texture information of pavement crack images, which achieves a PSNR of 29.9958.
3. The richer the texture information of the generated image is, the higher the detection performance gain brought by the data augmentation strategy, which verifies the necessity of the SR strategy.
4. For different combinations of initially generated image sizes and SR rates, the B (×2) strategy has the best performance, while the B (×8) strategy strikes a better balance between training cost and performance, and is more suitable for scenarios with limited computational costs.

Although the proposed method can reduce the cost of image generation by DDPM to some extent, it still remains impractical for scenarios with strict computational constraints. Future work will focus on developing a lightweight DDPM tailored for pavement crack detection. In addition, since the generated images still require manual labeling, further research will explore generation models capable of producing both images and corresponding labeled data simultaneously.

ACKNOWLEDGMENTS

The authors appreciate the financial support from Hunan Expressway Group Co. Ltd. and the Transportation Science and Technology Progress and Innovation Program of Hunan Province (No. 202152) in China. The authors also appreciate the funding support from Beijing's high-level overseas talents and Key Scientific Research Projects of BBMG Corporation (KYJC018) in China. The authors would like to thank Prof. Chengshun Xu from the Beijing University of Technology and Dr. Weiwei Lu from the Changsha University of Science and Technology for their generous help. Any opinion, finding, and conclusion expressed in this paper are those of the authors and do not necessarily represent the views of any organization.This research is supported by the Digital Roads Prosperity Partnership (DR). DR is supported by the Engineering and Physical Sciences Research Council (EPSRC) grant number [EP/V056441/1], Costain, National Highways, the University of Cambridge, Department for Transport, and Didimi. The authors gratefully acknowledge the collaboration of all academic and industrial project partners. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the institutes mentioned above.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

Ai, D., Jiang, G., Lam, S.-K., He, P., & Li, C. (2023). Computer vision framework for crack detection of civil infrastructure—A review. *Engineering Applications of Artificial Intelligence*, *117*, 105478.

Cano-Ortiz, S., Iglesias, L. L., del Árbol, P. M. R., & Castro-Fresno, D. (2024). Improving detection of asphalt distresses with deep learning-based diffusion model for intelligent road maintenance. *Developments in the Built Environment*, *17*, 100315.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, & J. M. Frahm (Eds.), *Lecture notes in computer science: Vol.* 12346. *European conference on computer vision* (pp. 213–229). Springer.

Chen, J., & He, Y. (2022). A novel U-shaped encoder–decoder network with attention mechanism for detection and evaluation of road cracks at pixel level. *Computer-Aided Civil and Infrastructure Engineering*, *37*(13), 1721–1736.

Chen, X., Wang, X., Zhang, W., Kong, X., Qiao, Y., Zhou, J., & Dong, C. (2023). *HAT: Hybrid attention transformer for image restoration*. arXiv. https://arxiv.org/pdf/2309.05239

Dhanya, V. G., Subeesh, A., Kushwaha, N. L., Kumar Vishwakarma, D., Kumar, T. N., Ritika, G., & Singh, A. N. (2022). Deep learning based computer vision approaches for smart agricultural applications. *Artificial Intelligence in Agriculture*, *6*, 211–229.

Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 295–307.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*, Montreal, Quebec, Canada.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV (pp. 770–778).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, *30*, Long Beach, CA.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, Virtual (pp. 6840–6851).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT (pp. 7132–7141).

Huang, Y., Liu, Y., Liu, F., & Liu, W. (2024). A lightweight feature attention fusion network for pavement crack segmentation. *Computer-Aided Civil and Infrastructure Engineering*, *39*(18), 2811–2825.

Jocher, G., Qiu, J., & Chaurasia, A. (2023). *Ultralytics YOLO (Version 8.0.0)*. https://github.com/ultralytics/ultralytics

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. arXiv. https://doi.org/10.48550/arXiv.1710.10196

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, *34*, Online (pp. 852–863).

Kashyap, A. A., Raviraj, S., Devarakonda, A., Nayak K, S. R., KV, S., & Bhat, S. (2022). Traffic flow prediction models–A review of deep learning techniques. *Cogent Engineering*, *9*(1), 2010510.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., & Wang, Z. (2017). Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI (pp. 4681–4690).

Li, W., Fan, L., Wang, Z., Ma, C., & Cui, X. (2021). Tackling mode collapse in multi-generator GANs with orthogonal vectors. *Pattern Recognition*, *110*, 107646.

Li, Z., & Adeli, H. (2018). Control methodologies for vibration control of smart civil and mechanical structures. *Expert Systems*, *35*(6), e12354.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). SwinIR: Image restoration using Swin Transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, BC, Canada (pp. 1833–1844).

Lim, B., Son, S., Kim, H., Nah, S., & Mu Lee, K. (2017). Enhanced deep residual networks for single image super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, HI (pp. 136–144).

Liu, J., Yang, X., Lau, S., Wang, X., Luo, S., Lee, V. C. S., & Ding, L. (2020). Automated pavement crack detection and segmentation based on two-step convolutional neural network. *Computer-Aided Civil and Infrastructure Engineering*, *35*(11), 1291–1305.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, BC, Canada (pp. 10012–10022).

Lu, J., Pan, B., Ren, W., Liu, Q., Liu, P., & Oeser, M. (2023). Regeneration of pavement surface textures using M-sigmoid-normalized generative adversarial networks. *Computer-Aided Civil and Infrastructure Engineering*, *38*(16), 2225–2241.

Loshchilov, I., & Hutter, F. (2017). *Fixing weight decay regularization in Adam*. arXiv. https://arxiv.org/pdf/1711.05101v2/1000

Mammone, N., Ieracitano, C., Adeli, H., & Morabito, F. C. (2023). AutoEncoder filter bank common spatial patterns to decode motor imagery from EEG. *IEEE Journal of Biomedical and Health Informatics*, *27*(5), 2365–2376.

Nguyen, H.-T., & Cheah, C. C. (2022). Analytic deep neural network-based robot control. *IEEE/ASME Transactions on Mechatronics*, *27*(4), 2176–2184.

Nogay, H. S., & Adeli, H. (2023). Diagnostic of autism spectrum disorder based on structural brain MRI images using, grid search optimization, and convolutional neural networks. *Biomedical Signal Processing and Control*, *79*, 104234.

Nogay, H. S., & Adeli, H. (2024). Multiple classification of brain MRI autism spectrum disorder by age and gender using deep learning. *Journal of Medical Systems*, *48*(1), 15.

Oh, D., Jeong, S., Bae, S.-K., Kim, B., & Cho, S. (2025). Training deep learning segmentation models using super-resolution crack images for detection of thin concrete cracks. *Journal of Computing in Civil Engineering*, *39*(4), 04025035.

Oliveira, H., & Correia, P. L. (2009). Automatic road crack segmentation using entropy and image dynamic thresholding. *2009 17th European Signal Processing Conference*, Glasgow, UK (pp. 622–626).

Perez-Ramirez, C. A., Amezquita-Sanchez, J. P., Valtierra-Rodriguez, M., Adeli, H., Dominguez-Gonzalez, A., & Romero-Troncoso, R. J. (2019). Recurrent neural network model with Bayesian training and mutual information for response prediction of large buildings. *Engineering Structures*, *178*, 603–615.

Perez-Sanchez, A. V., Amezquita-Sanchez, J. P., Valtierra-Rodriguez, M., & Adeli, H. (2024). A new epileptic seizure prediction model based on maximal overlap discrete wavelet packet transform, homogeneity index, and machine learning using ECG signals. *Biomedical Signal Processing and Control*, *88*, 105659.

Rafiei, M. H., & Adeli, H. (2018). Novel machine-learning model for estimating construction costs considering economic variables and indexes. *Journal of Construction Engineering and Management*, *144*(12), 04018106.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA (pp. 10684–10695).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab,

J. Hornegger, W. Wells, & A. Frangi (Eds.), *Lecture notes in computer science: Vol*. 935. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference* (pp. 234–241). Springer.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy (pp. 618–626).

Shen, W., Zeng, D., Zhang, Y., Tian, X., & Li, Z. (2024). Image augmentation for nondestructive testing in engineering structures based on denoising diffusion probabilistic model. *Journal of Building Engineering*, *89*, 109299.

Shim, S. (2024). Self-training approach for crack detection using synthesized crack images based on conditional generative adversarial network. *Computer-Aided Civil and Infrastructure Engineering*, *39*(7), 1019–1041.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, Long Beach, CA.

Wang, C., Liu, H., An, X., Gong, Z., & Deng, F. (2024). SwinCrack: Pavement crack detection using convolutional swin-transformer network. *Digital Signal Processing*, *145*, 104297.

Wang, X., Yu, K., Wu, S., Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., & Change Loy, C. (2018). ESRGAN: Enhanced super-resolution generative adversarial networks. In L. Leal-Taixé & S. Roth (Eds.), *Lecture notes in computer science: Vol*. 11133. *European conference on computer vision (ECCV) workshops* (pp. 63–79). Springer.

Xiang, C., Wang, W., Deng, L., Shi, P., & Kong, X. (2022). Crack detection algorithm for concrete structures based on super-resolution reconstruction and segmentation network. *Automation in Construction*, *140*, 104346.

Yao, H., Liu, Y., Li, X., You, Z., Feng, Y., & Lu, W. (2022). A detection method for pavement cracks combining object detection and attention mechanism. *IEEE Transactions on Intelligent Transportation Systems*, *23*(11), 22179–22189.

Yao, H., Liu, Y., Lv, H., Huyan, J., You, Z., & Hou, Y. (2024). Encoder–decoder with pyramid region attention for pixel-level pavement crack recognition. *Computer-Aided Civil and Infrastructure Engineering*, *39*(10), 1490–1506.

Yi, C., Liu, J., Huang, T., Xiao, H., & Guan, H. (2023). An efficient method of pavement distress detection based on improved YOLOv7. *Measurement Science and Technology*, *34*(11), 115402.

Yu, T., Li, C., Huang, J., Xiao, X., Zhang, X., Li, Y., & Fu, B. (2024). ReF-DDPM: A novel DDPM-based data augmentation method for imbalanced rolling bearing fault diagnosis. *Reliability Engineering & System Safety*, *251*, 110343.

Yuan, B., Sun, Z., Pei, L., Li, W., Ding, M., & Hao, X. (2022). Super-resolution reconstruction method of pavement crack images based on an improved generative adversarial network. *Sensors*, *22*(23), 9092.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., & Shum, H.-Y. (2022). DINO: DETR with improved denoising anchor boxes for end-to-end object detection. arXiv. https://arxiv.org/pdf/2203.03605

Zhang, H., Qian, Z., Zhou, W., Min, Y., & Liu, P. (2024). A controllable generative model for generating pavement crack images in complex scenes. *Computer-Aided Civil and Infrastructure Engineering*, *39*(12), 1795–1810.

Zhang, L., Yang, F., Daniel Zhang, Y., & Zhu, Y. J. (2016). Road crack detection using deep convolutional neural network. *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ (pp. 3708–3712).

Zhang, T., Wang, D., & Lu, Y. (2024). A data-centric strategy to improve performance of automatic pavement defects detection. *Automation in Construction*, *160*, 105334.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). Residual dense network for image super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT (pp. 2472–2481).

Zhao, H., Qin, G., & Wang, X. (2010). Improvement of canny algorithm based on pavement edge detection. *2010 3rd International Congress on Image and Signal Processing*, Yantai, China (pp. 964–967).

Zhong, J., Huyan, J., Zhang, W., Cheng, H., Zhang, J., Tong, Z., Jiang, X., & Huang, B. (2023). A deeper generative adversarial network for grooved cement concrete pavement crack detection. *Engineering Applications of Artificial Intelligence*, *119*, 105808.

Zhu, G., Liu, J., Fan, Z., Yuan, D., Ma, P., Wang, M., Sheng, W., & Wang, K. C. P. (2024). A lightweight encoder–decoder network for automatic pavement crack detection. *Computer-Aided Civil and Infrastructure Engineering*, *39*(12), 1743–1765.

How to cite this article: Yao, H., Liu, Y., Besklubova, S., Brilakis, I., Guo, M., Wang, J., & Wang, M. (2025). A data augmentation method for pavement crack detection based on super-resolution and denoising diffusion probabilistic models. *Computer-Aided Civil and Infrastructure Engineering*, 1–14. https://doi.org/10.1111/mice.70050