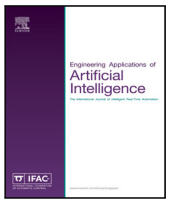
Engineering Applications of Artificial Intelligence 178 (2026) 115113

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Class-controllable diffusion with mask generation for pavement distress detection

Yuanyuan Hu [a,b,1], Hancheng Zhang [a,b,1], Meng Guo [c,d], Pengfei Liu [b,*]

[a] School of Information Engineering, Wenzhou Business College, Wenzhou, 325035, China
[b] Institute of Highway Engineering, RWTH Aachen University, Aachen, 52072, Germany
[c] State Key Laboratory of Bridge Engineering Safety and Resilience, Beijing University of Technology, Beijing, 100124, China
[d] The Key Laboratory of Urban Security and Disaster Engineering of Ministry of Education, Beijing University of Technology, Beijing, 100124, China

ARTICLE INFO

Keywords:
Pavement distress detection
Diffusion models
Data augmentation
Segmentation mask generation

ABSTRACT

Accurate and efficient detection of pavement distress is essential for maintaining road infrastructure safety and functionality. However, artificial intelligence-based detection methods remain constrained by limited labeled data, class imbalance, and high pixel-level annotation cost. To address these challenges, CrackMaskNet is proposed as a diffusion-based generative framework that models pavement appearance and pixel-level annotation as coupled stochastic variables within a unified diffusion process. By jointly embedding distress images and their corresponding segmentation masks into a shared noise trajectory, the framework enables coordinated evolution of visual texture and structural boundaries, supporting class-controllable synthesis of aligned image–mask pairs without additional annotation stages. Experimental results show that the generated pavement distress images achieve a Fréchet Inception Distance of 39.6 and a Kernel Inception Distance of 0.046, while the corresponding segmentation masks reach a mean Intersection over Union of 0.81 and a Dice score of 0.89, indicating coherent image–mask generation. Incorporating the synthesized samples improves downstream classification, detection, and segmentation performance, particularly for visually ambiguous and underrepresented crack categories. Evaluation on CRACK500 and related datasets further demonstrates stable image–mask generation under moderate domain shifts. CrackMaskNet provides a practical approach for constructing annotated pavement distress datasets while reducing manual labeling effort, supporting AI-enabled infrastructure asset management and big-data-driven maintenance decision-making.

1. Introduction

The durability and performance of pavement infrastructure are vital for maintaining the safety and efficiency of transportation networks (Ayman and Fakhr, 2023). However, pavements continuously experience deterioration due to environmental factors and mechanical loads, resulting in various types of distress such as cracks, potholes, and surface deformations (J. Li et al., 2022). Regular inspection and maintenance of pavements are essential for extending their lifespan and ensuring road safety. However, developing reliable and automated pavement distress detection methods remains challenging, particularly when applied to complex and diverse real-world conditions.

Traditional approaches for pavement condition assessment have primarily relied on conventional image processing techniques, including edge detection (Wang et al., 2007; Abbas and Ismael, 2021), texture analysis (Nair et al., 2018), and other classical image processing methods (Jing and Aiqin, 2010; Radopoulou et al., 2013). While these methods offer baseline capabilities for identifying distress, they often lack robustness when applied to diverse pavement conditions with varying textures, lighting, and environmental factors. The advancement of machine learning and deep learning has substantially improved image recognition capabilities for pavement distress. Classification models such as Residual Network with 50 layers (ResNet50) (Li et al., 2023) enable coarse identification of distress types. Object detection algorithms including enhanced You Only Look Once version 12 (YOLOv12) networks (Wang et al., 2024) can effectively localize damage regions while suppressing interference from complex environmental noise. Semantic segmentation techniques, exemplified by Pavement Distress Segmentation Network (PDSNet) (Wen et al., 2023) with its parallel feature extraction branches, can quantify the extent of damage by precisely labeling each pixel in an image.

Despite substantial progress in pavement distress detection, developing robust and generalizable models remains challenging due to

* Corresponding author.
E-mail addresses: hu@isac.rwth-aachen.de (Y. Hu), zhang@isac.rwth-aachen.de (H. Zhang), gm@bjut.edu.cn (M. Guo), liu@isac.rwth-aachen.de (P. Liu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.engappai.2026.115113
Received 14 April 2025; Received in revised form 7 March 2026; Accepted 10 May 2026
Available online 18 May 2026

several fundamental factors. First, the scarcity of large-scale, high-quality annotated datasets significantly constrains the performance of data-driven methods. Existing pavement distress datasets are often limited in size and exhibit severe class imbalance, with certain distress categories being markedly underrepresented (Maeda et al., 2018; Shi et al., 2016; Cui et al., 2015; Yang et al., 2019). Second, obtaining pixel-level annotations for pavement distress images requires specialized domain expertise and extensive manual effort, resulting in high annotation costs that restrict dataset scalability and diversity (Schreiner et al., 2006). Third, the visual similarity between different distress types — particularly between untreated cracks and repaired areas such as sealed or patched sections — introduces substantial ambiguity. Such morphological and color-level similarities frequently lead to misclassification, even when advanced deep learning models are employed (Gong et al., 2023). These challenges collectively limit the deployment of existing detection systems in real-world engineering scenarios, where robustness across diverse pavement conditions is essential.

Several strategies have been proposed to address data-related challenges. Traditional data augmentation techniques (Wu et al., 2020; Fawzi et al., 2016) enhance dataset diversity through transformations such as rotation (Xi et al., 2018), scaling (Chen et al., 2020), and flipping (Kim and Lee, 2022), but remain limited to the distribution boundaries of existing data. Transfer learning approaches (Gopalakrishnan et al., 2017; Li et al., 2021) employ models pre-trained on large-scale datasets, yet frequently struggle with the substantial domain gap between general imagery and specialized pavement textures. Semi-supervised and unsupervised learning methods (Ren et al., 2023; Li et al., 2020) reduce annotation requirements but typically underperform compared to fully supervised models for complex segmentation tasks. More advanced generative approaches have emerged to synthesize additional training data. Variational Autoencoders (VAEs) (Hu et al., 2024) provide probabilistic modeling capabilities but often produce blurry outputs lacking the fine details crucial for distress detection. Generative Adversarial Networks (GANs) (Ashwini et al., 2024) create synthetic images with sharper details but frequently exhibit artifacts including unrealistic textures and structural distortions. Although diffusion models (Zhang et al., 2024a) substantially improve generation fidelity, existing implementations provide limited control over multiple distress categories and rarely support joint generation of images and pixel-level annotations.

To address these limitations, this paper proposes CrackMaskNet, a novel diffusion-based generative model for pavement distress detection that systematically addresses the challenges. The primary contributions of this work are:

- CrackMaskNet formulates pavement appearance and pixel-level annotation as coupled variables within a unified diffusion framework, enabling class-controllable generation of aligned image–mask pairs under limited and imbalanced labeling conditions.
- A joint image–mask generation strategy treats segmentation masks as intrinsic generative variables rather than external conditioning signals, reducing reliance on manual pixel-level annotation.
- The proposed joint image–mask generation mechanism, together with a noise-aware adaptive feature scaling strategy embedded in the denoising network, enhances structural consistency and stabilizes feature representations across diffusion states, which helps alleviate misclassification caused by morphological and color similarity among different distress patterns.
- Extensive experiments demonstrate improved generation fidelity, stronger image–mask alignment, and stable performance under class imbalance and moderate domain shifts.

2. Literature review

Generative models have emerged as powerful tools for addressing data scarcity in pavement distress detection. This section explores three primary generative approaches: Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models, analyzing their capabilities and limitations in the context of pavement distress data generation.

2.1. VAEs

Variational Autoencoders implement a probabilistic framework consisting of an encoder network transforming input images into a latent distribution and a decoder network that reconstructs images by sampling from this distribution.

VAE-ResNet (Wang and Liu, 2024) integrates VAE, ResNet, and Residual Feature Distillation Block (RFDB) to suppress stochastic clutter in data and enables end-to-end processing for subsurface anomaly imaging. VAE-Diffusion (Chen et al., 2025) reduces zero observations in frequency data and handles multi-type tabular variables to enhance synthetic data quality and improve predictive accuracy. VD-CrackGAN (Yu et al., 2024) combines VAE and Deep Convolutional Generative Adversarial Network (DCGAN) for pavement crack data augmentation, integrating spectral normalization and the Swin Transformer to enhance crack generation quality. Structural Variational Autoencoder (SVAE) (Bacsa et al., 2025) learns structural dynamics from dynamic response data for classification in structural health monitoring, enabling damage detection, characterization, and transfer learning across different property clusters.

Despite their probabilistic interpretability, VAEs tend to oversmooth high-frequency textures, which are essential for distinguishing fine-scale pavement distresses. This degradation stems from pixel-wise reconstruction loss functions that optimize for average correctness rather than structural fidelity. Additionally, controlling specific geometric properties of generated distresses often requires complex architecture modifications, limiting practical applications (Bredell et al., 2023).

2.2. GANs

Generative Adversarial Networks employ a competitive framework between generator and discriminator networks, where the generator produces synthetic images while the discriminator evaluates their authenticity against real samples. This adversarial optimization drives continuous improvement in generation quality, enabling GANs to produce sharper, more detailed outputs compared to VAEs.

Conditional adversarial translation frameworks such as pix2pix (Isola et al., 2017) further extend GANs to paired image-to-image generation by learning mappings between structured inputs and target images. AsphaltGAN (Cano-Ortiz et al., 2024b) is a class-conditional GAN with attention mechanisms designed for dataset augmentation, generating diverse rare road defects to enhance object detection in pavement distress detection systems. Chen and Ren (2024) proposed a method that combines GAN-generated damage with texture synthesis to control severity levels and uses Poisson blending for realistic integration, automating sample selection to enhance road damage detection. Furthermore, Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) is utilized to facilitate the generation of high-quality pavement crack images, ensuring enhanced realism and structural fidelity in the synthesized data (Han et al., 2024; Zhang and Zhang, 2024).

Despite notable improvements in realism, GAN-based approaches remain vulnerable to training instability and mode collapse, which can distort distress morphology and reduce data reliability. For pavement applications, GANs lack mechanisms for generating corresponding segmentation masks alongside images, requiring additional annotation processes.

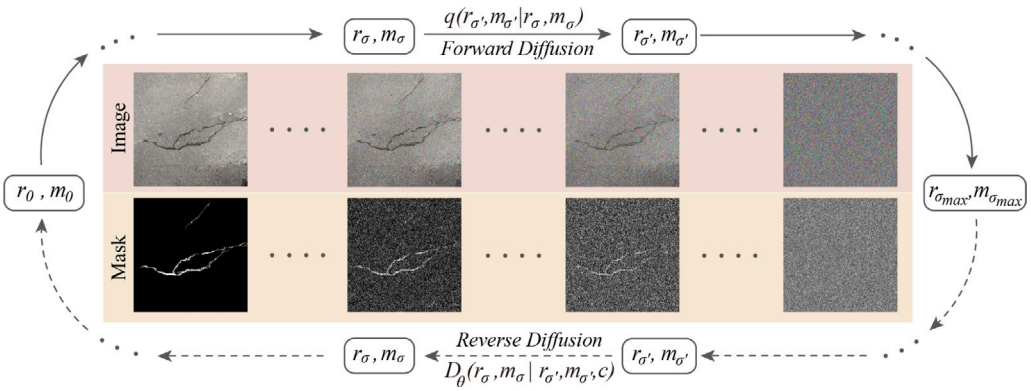


Fig. 1. CrackMaskNet framework. Overview of the joint diffusion pipeline for crack image and segmentation mask generation, illustrating the forward noise injection process and the noise-conditioned reverse denoising process guided by class labels.

2.3. Diffusion models

Diffusion models implement a principled probabilistic framework based on thermodynamics principles. These models operate through a forward diffusion phase that gradually adds noise to images and a reverse denoising phase that learns to recover the original image from noise (Croitoru et al., 2023). This decomposition into multiple denoising steps enables more stable training and higher quality outputs.

The iterative denoising framework provides fine-grained control at different noise levels and a well-defined learning target at each step (Delbracio and Milanfar, 2023). Recent studies have explored efficiency and conditioning strategies in diffusion models, including timestep caching mechanisms for video diffusion (Liu et al., 2025), context-aware diffusion-guided refinement for hyperspectral classification (Chatterjee et al., 2025), and event-guided diffusion for physically realistic video interpolation (Zhang et al., 2025). Conditional control mechanisms (ControlNet)have also been introduced to incorporate external structural signals into diffusion models (Zhang et al., 2023). In the pavement domain, recent research has primarily focused on conditional diffusion approaches to control crack position and morphology, as well as the generation of synthetic images for rare defect types (Zhang et al., 2024b; Cano-Ortiz et al., 2024a).

Despite producing superior quality images with fewer artifacts, diffusion models face significant limitations including limited fine-grained category control across multiple distress types and inability to generate corresponding pixel-level annotations alongside images. Although applications of diffusion models in pavement distress detection remain limited, their capacity for stable training and high-fidelity synthesis highlights a promising yet underexplored direction.

Building upon these insights from generative models, a significant research gap remains in developing integrated frameworks that simultaneously address the three critical challenges in pavement distress detection: data scarcity with class imbalance, high annotation costs, and visual similarity between distress types. The limitations of existing methods — VAEs producing oversmoothed textures, GANs suffering from instability, and diffusion models lacking integrated control and annotation capabilities — indicate the motivation for exploring a unified generative framework. The proposed CrackMaskNet methodology addresses these limitations through a specialized diffusion-based model with enhanced control mechanisms and dual-output capabilities, as detailed in the following section.

3. Methodology

The proposed CrackMaskNet is designed to generate high-quality, class-controllable pavement distress images together with their corresponding segmentation masks within a unified diffusion framework. By jointly modeling crack appearance and structural annotation through a coupled denoising process, the method aims to improve both the visual realism and structural consistency of synthesized crack patterns.

Unlike conditional diffusion frameworks, where structural annotations are introduced as external control signals or guidance inputs, CrackMaskNet models the segmentation mask as a stochastic variable that participates in both the forward noise injection and reverse denoising processes. Specifically, the image and its corresponding mask share the same diffusion trajectory and noise schedule, enabling their coordinated evolution under identical stochastic states rather than treating annotation as a post-hoc prediction target or conditioning cue.

As illustrated in Fig. 1, CrackMaskNet follows a diffusion-based generative pipeline consisting of forward and reverse processes. During the forward diffusion stage, Gaussian noise is progressively injected into paired crack images and masks, encoding them into a latent noisy space. In the reverse sampling stage, the model employs the Dual-Output Guided Synthesis Module (DOGS), together with class conditioning, to iteratively refine the latent representations and reconstruct semantically consistent crack images and segmentation masks. The overall architecture integrates multi-scale skip connections, attention mechanisms, and adaptive noise-aware scaling to facilitate stable and robust denoising across different noise levels.

3.1. Forward diffusion process

The forward diffusion process jointly encodes crack images and their corresponding segmentation masks into a shared stochastic latent space by progressively corrupting both modalities with identical Gaussian noise levels. Given a clean crack image r_0 and its corresponding segmentation mask m_0, the forward process generates noisy representations (r_σ, m_σ) under increasing noise levels, thereby transforming the data into a noise-dominated latent space.

Under this formulation, the noisy crack image and mask at noise level σ are sampled according to the following distributions:

$$\begin{aligned} q(r_\sigma \mid r_0) &= \mathcal{N}\left(r_\sigma; \frac{r_0}{\sqrt{1+\sigma^2}}, \frac{\sigma^2}{1+\sigma^2}\mathbf{I}\right), \\ q(m_\sigma \mid m_0) &= \mathcal{N}\left(m_\sigma; \frac{m_0}{\sqrt{1+\sigma^2}}, \frac{\sigma^2}{1+\sigma^2}\mathbf{I}\right), \end{aligned} \tag{1}$$

where σ controls the strength of the injected Gaussian noise and $\mathbf{I}$ denotes the identity matrix. The scaling factor $1/\sqrt{1+\sigma^2}$ attenuates the clean signal, while the variance term $\sigma^2/(1+\sigma^2)$ determines the magnitude of the added noise. As σ increases, the noisy samples progressively deviate from the original image and mask, enabling a gradual transition from clean data to latent noisy representations.

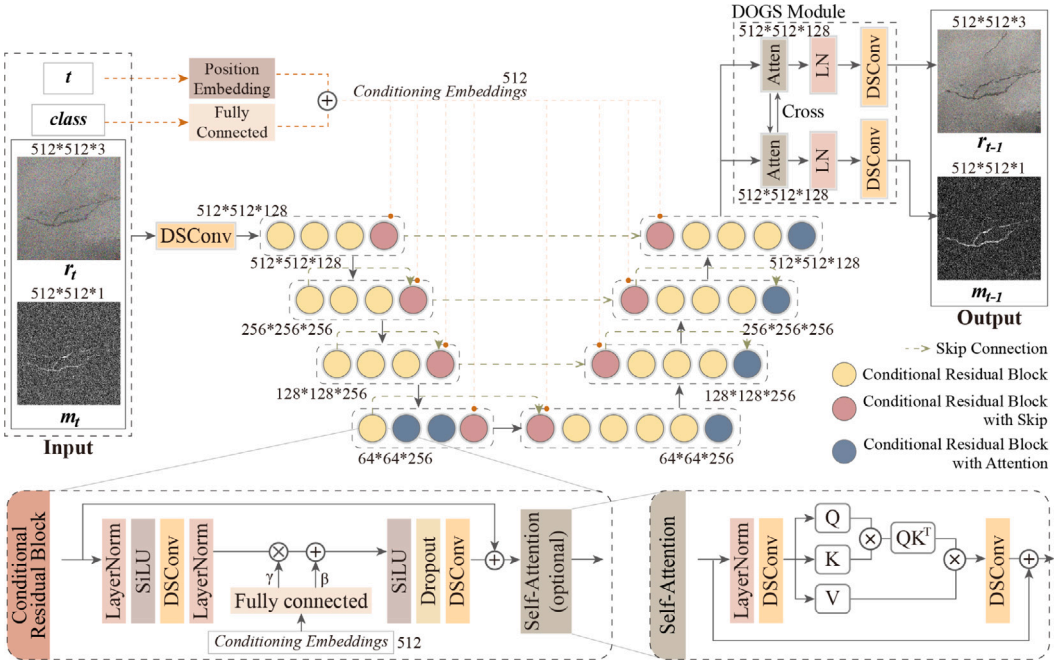


Fig. 2. Structure of CrackMaskNet. An overview of the network architecture, highlighting the Conditional Residual Blocks and the Dual-Output Guided Synthesis Module for joint image–mask denoising.

3.2. Reverse diffusion process

The reverse diffusion process progressively removes noise from corrupted crack images and their corresponding segmentation masks in order to recover clean and structurally consistent representations. Following the general principle of implicit diffusion sampling (Song et al., 2020), the reverse process allows flexible transitions between noise levels and supports efficient refinement of the joint image–mask representation.

Starting from a noisy crack image r_σ and mask m_σ at a given noise level σ, the model iteratively refines the inputs by estimating a denoised representation conditioned on the current noise level, the class label c, and the learned model parameters. Class conditioning is incorporated through embedding layers and injected into the network to guide the denoising process toward specific distress categories, enabling controllable joint synthesis of images and masks.

Under the continuous noise-level parameterization, the reverse update from noise level σ to a lower noise level $\sigma' \in (0, \sigma)$ is expressed as:

$$(r_{\sigma'}, m_{\sigma'}) = \frac{(r_\sigma, m_\sigma) + \sigma'^2 D_\theta(r_\sigma, m_\sigma, \sigma, c)}{1+\sigma'^2} + \sigma' \sqrt{\frac{1+\sigma^2}{1+\sigma'^2} - 1}\,\epsilon, \tag{2}$$

where $D_\theta(\cdot)$ denotes the denoising function implemented by CrackMaskNet, which outputs a noise-conditioned estimate of the underlying crack image and segmentation mask. The variable $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ represents standard Gaussian noise, and σ' specifies the target noise level of the current reverse step.

The first term in the update equation refines the current samples using the denoised estimate, while the second term introduces controlled stochasticity to ensure smooth transitions between successive noise levels. The reverse process is applied iteratively until a sufficiently low noise level is reached, yielding the final crack image and segmentation mask estimates.

3.3. Dual-output guided synthesis module

The Dual-Output Guided Synthesis Module (DOGS) is designed to facilitate joint denoising of crack images and their corresponding segmentation masks within a unified diffusion framework. Rather than producing the two outputs in isolation, DOGS introduces a lightweight interaction mechanism that allows information exchange between the image and mask prediction pathways under the same diffusion state.

Given a noisy image–mask pair (r_σ, m_σ) at a specific noise level, the shared backbone features are first processed to enable cross-modal interaction between appearance and structural representations. Based on these interacted features, DOGS then generates two task-specific outputs through separate prediction heads, yielding the denoised crack image and the associated segmentation mask, respectively. This design maintains task-specific flexibility while encouraging consistency between the synthesized visual content and its pixel-level annotation.

By allowing controlled interaction prior to output prediction, DOGS implicitly enforces structural alignment between crack appearance and geometry throughout the reverse diffusion process. This coupling is particularly important for pavement crack patterns, which are typically thin, elongated, and spatially sparse, and therefore sensitive to misalignment between visual texture and structural boundaries. Through joint conditioning and moderated feature interaction, DOGS enables stable and semantically consistent synthesis of crack images and segmentation masks without introducing additional supervision stages or complex architectural overhead.

3.4. CrackMaskNet architecture

CrackMaskNet is implemented as a conditional denoising architecture that realizes the joint image–mask refinement strategy introduced by the Dual-Output Guided Synthesis Module. As illustrated in Fig. 2, the network operates on noisy crack images and segmentation masks under a given diffusion state and produces corresponding denoised estimates through a shared backbone with task-specific outputs.

At each denoising step, the network takes four inputs: a noisy crack image r_σ, a noisy segmentation mask m_σ, the class label c, and the associated noise level σ. The noise level and class label are encoded via positional embeddings and lightweight fully connected layers to form conditioning signals, which are injected into multiple stages of the network. This noise-aware and class-conditioned modulation allows the denoising behavior to adapt across diffusion states while maintaining controllable generation of different distress categories.

The backbone of CrackMaskNet is composed of a hierarchy of conditional residual blocks, within which three complementary architectural components are integrated to support stable and effective denoising. First, multi-scale skip connections are employed to facilitate information propagation across diffusion stages and to preserve fine-grained structural details during progressive refinement. Second, attention-enhanced feature refinement is incorporated to improve spatial selectivity, enabling the network to emphasize crack-relevant regions that are typically thin, elongated, and sparsely distributed in pavement images.

Third, to improve robustness under varying noise intensities, CrackMaskNet introduces an adaptive noise-aware scaling mechanism within the conditional residual blocks. Given an intermediate feature map $\mathbf{h}$ and the corresponding noise level σ, a pair of noise-dependent modulation parameters — a scaling factor $\gamma(\sigma)$ and a shift term $\beta(\sigma)$ — are predicted from the noise embedding and applied as:

$$\mathbf{h}' = \gamma(\sigma) \odot \mathbf{h} + \beta(\sigma), \tag{3}$$

where $\odot$ denotes element-wise multiplication with broadcasting over spatial dimensions. This affine modulation adjusts feature magnitudes and offsets according to the current diffusion state, providing a simple yet effective mechanism to stabilize feature distributions from high-noise to low-noise regimes.

Built upon this shared noise-conditioned backbone, the network produces denoised estimates of the crack image and its corresponding segmentation mask through parallel output heads. By iteratively applying this architecture across reverse diffusion steps, CrackMaskNet reconstructs high-quality distress images together with pixel-accurate segmentation masks from latent noisy representations.

3.5. Training objective and noise schedule

CrackMaskNet is trained to jointly denoise crack images and segmentation masks under randomly sampled noise levels. Given a noisy image–mask pair (r_σ, m_σ), the network outputs denoised estimates for both modalities. The training objective consists of two complementary loss terms. For the crack image branch, a mean squared error loss is employed:

$$\mathcal{L}_r = \mathbb{E}_\sigma \left[w(\sigma) \left\| D_{\theta,r}(r_\sigma, m_\sigma, \sigma, c) - r_0 \right\|_2^2 \right]. \tag{4}$$

For the segmentation mask branch, a binary cross-entropy loss is adopted:

$$\mathcal{L}_m = \mathbb{E}_\sigma \left[w(\sigma)\, \mathrm{BCE}\big(D_{\theta,m}(r_\sigma, m_\sigma, \sigma, c), m_0 \big) \right]. \tag{5}$$

The overall objective is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_r \mathcal{L}_r + \lambda_m \mathcal{L}_m, \tag{6}$$

where $\lambda_r = \lambda_m = 1.0$. Since image reconstruction and mask prediction are complementary and non-adversarial, equal weighting provides stable joint optimization.

The noise-dependent weighting function $w(\sigma)$ is defined as:

$$w(\sigma) = \frac{\sigma^2 + \sigma_{\text{data}}^2}{(\sigma \cdot \sigma_{\text{data}})^2}, \quad \sigma_{\text{data}} = 0.5, \tag{7}$$

which balances gradient contributions across different noise levels and follows established diffusion training practices (Karras et al., 2022).

4. Experiment

4.1. Dataset settings

The experiments employ a dataset consisting of 1200 pavement distress and repair images distributed across four categories, including transverse cracks, longitudinal cracks, alligator cracks, and repaired cracks. The category-wise distribution is reported in Table 1. Consistent with practical pavement inspection scenarios, the dataset presents a pronounced class imbalance, where repaired cracks constitute the smallest subset. This imbalance introduces substantial challenges for model learning and evaluation.

Table 1
Distribution of the dataset categories.

Category	Number of images
Transverse cracks	312
Longitudinal cracks	356
Alligator cracks	362
Repaired cracks	170
Total	**1200**

All images were captured using a Hikvision MV-CE060-10UM industrial camera equipped with a WL1224-4MP 12 mm lens in Aachen, Germany, under varying natural lighting conditions to ensure robustness to real-world environmental variations. For standardization and computational efficiency, all images were preprocessed to a uniform resolution of 512 × 512 pixels prior to model training and evaluation.

4.2. Computational environment and implementation details

The experimental framework was implemented within a high-performance computing environment running Ubuntu 22.04. The hardware configuration consisted of dual Intel 5318Y processors supported by eight NVIDIA RTX A6000 Graphics Processing Units (GPUs), each equipped with 48 GB of Video Random Access Memory (VRAM), providing sufficient computational capacity for training the diffusion-based model. The software implementation utilized Python 3.8 and PyTorch 2.1 for deep learning operations.

Model training utilized the Adaptive Moment Estimation (Adam) optimizer with an initial learning rate of 1e–4 and a batch size of 16. The CrackMaskNet model was trained for 800,000 iterations, ensuring convergence of the diffusion process parameters and stability of the generated outputs. All quantitative results are reported as mean ± standard deviation over five independent runs with different random seeds. To support reproducibility and further research, the codebase and dataset are publicly available at: https://github.com/SEU-zhc/CrackMaskNet

4.3. Evaluation of indicators

To evaluate the performance of the proposed CrackMaskNet model, both the quality of the generated pavement images and the accuracy of the corresponding segmentation masks are quantitatively assessed. The generated images are compared with real images from the test set, and the generated masks are evaluated against their ground-truth annotations. The evaluation is conducted using four complementary metrics, including Fréchet Inception Distance (FID) (Szegedy et al., 2016), Kernel Inception Distance (KID) (Bińkowski et al., 2018), mean Intersection over Union (mIoU) (Rezatofighi et al., 2019), and Dice coefficient (Setiawan, 2020).

Fréchet Inception Distance (FID) is employed to measure the similarity between the distributions of generated and real pavement images. The metric computes the Fréchet distance between feature representations extracted from a pretrained Inception network. Lower FID values indicate closer alignment between the generated and real image distributions and therefore reflect higher visual fidelity. The FID is defined as

$$\mathrm{FID} = \left\| \mu_r - \mu_g \right\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}} \right), \tag{8}$$

where μ_r and μ_g denote the mean feature vectors, and Σ_r and Σ_g denote the covariance matrices of real and generated images, respectively.

Kernel Inception Distance (KID) is further adopted as a complementary metric for evaluating image generation quality. KID estimates the squared Maximum Mean Discrepancy between feature distributions

using polynomial kernels and provides an unbiased measure that is less sensitive to sample size. Lower KID values indicate better agreement between the generated and real image distributions. The KID is defined as

$$\mathrm{KID} = \mathbb{E}_{x,x'\sim P_r}\left[k(x,x')\right] + \mathbb{E}_{y,y'\sim P_g}\left[k(y,y')\right] - 2\mathbb{E}_{x\sim P_r,y\sim P_g}\left[k(x,y)\right], \tag{9}$$

where P_r and P_g represent the feature distributions of real and generated images, respectively, and $k(\cdot,\cdot)$ denotes a polynomial kernel function.

To evaluate the accuracy of the generated segmentation masks, mean Intersection over Union (mIoU) and Dice coefficient are employed. mIoU measures the spatial overlap between the generated mask and the ground-truth annotation across all categories and reflects the overall segmentation accuracy. It is defined as

$$\mathrm{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{|M_c^g \cap M_c^r|}{|M_c^g \cup M_c^r|}, \tag{10}$$

where C denotes the number of categories, and M_c^g and M_c^r represent the generated and reference masks for class c, respectively.

The Dice coefficient further evaluates mask similarity by emphasizing pixel-level agreement, which is particularly relevant for thin or underrepresented structures such as pavement cracks. It is defined as

$$\mathrm{Dice} = \frac{2|M^g \cap M^r|}{|M^g| + |M^r|}, \tag{11}$$

where M^g and M^r denote the generated and ground-truth masks, respectively. Higher mIoU and Dice values indicate more accurate and consistent segmentation results.

4.4. Generation results and quantitative evaluation

The generation capabilities of CrackMaskNet were evaluated through extensive experimentation focused on producing class-controllable pavement distress images and their corresponding segmentation masks. Fig. 3 presents representative results demonstrating the model's ability to synthesize four distinct pavement crack types: transverse cracks, longitudinal cracks, alligator cracks, and repaired cracks. Each generated image is accompanied by its precisely aligned segmentation mask, highlighting the pixel-level accuracy achieved by the dual-output architecture.

Quantitative evaluation of image generation quality yielded an FID score of 39.6 ± 0.8 and a KID value of 0.046 ± 0.008, indicating a strong alignment between the distributions of generated and real pavement images. These results suggest that CrackMaskNet is capable of capturing both the fine-grained textural patterns of pavement surfaces and the structural characteristics of different crack morphologies, while preserving background consistency across categories. The accuracy of the generated segmentation masks was assessed using mean Intersection over Union and Dice coefficient. The reported mIoU of 0.81 ± 0.02 and Dice score of 0.89 ± 0.01 were computed by directly comparing the generated masks with manually annotated ground-truth labels corresponding to the generated images. These results demonstrate a high degree of spatial consistency between the synthesized images and their masks, confirming the reliability of the joint image–mask generation process.

Across the four distress categories, CrackMaskNet produces stable and well-structured results for transverse, longitudinal, and alligator cracks, with clear morphological patterns and consistent spatial characteristics. In contrast, the generation of repaired cracks exhibits higher variability and reduced stability. This limitation primarily arises from the strong visual and geometric similarity between repaired and untreated cracks, as well as the smaller number of training samples available for the repaired crack category.

Despite these challenges, the explicit generation of repaired cracks remains a valuable capability, as repaired regions are a frequent source of misclassification in automated pavement inspection systems. By jointly generating pavement distress images and their corresponding pixel-level segmentation masks, the proposed approach eliminates the need for separate manual annotation and reduces the cost of constructing labeled datasets. This advantage is particularly relevant for underrepresented and visually ambiguous distress types, where additional annotated data can lead to meaningful improvements in downstream detection and segmentation performance.

Table 2
Ablation study of key components in CrackMaskNet.

Configuration	Skip	Atten.	Adaptive scale	FID ↓	KID ↓
Baseline	✗	✗	✗	50.3 ± 1.1	0.069 ± 0.011
Exp1	✓	✗	✗	48.7 ± 1.0	0.060 ± 0.011
Exp2	✗	✓	✗	44.7 ± 0.9	0.057 ± 0.011
Exp3	✗	✗	✓	41.9 ± 0.8	0.051 ± 0.008
Exp4	✓	✓	✗	44.5 ± 1.0	0.053 ± 0.011
Exp5	✓	✗	✓	40.4 ± 0.8	0.048 ± 0.009
Exp6	✗	✓	✓	39.9 ± 0.7	0.047 ± 0.010
Full	✓	✓	✓	**39.6 ± 0.8**	**0.046 ± 0.008**

5. Discussion

5.1. Ablation study

An ablation study was conducted to quantitatively assess the contribution of the three key components in CrackMaskNet: multi-scale skip connections, attention mechanisms, and adaptive noise-aware scaling. All configurations were trained under identical settings, and results are reported as mean ± standard deviation over five independent runs. Generative quality was evaluated using FID and KID.

As summarized in Table 2, enabling skip connections alone leads to a modest improvement over the baseline, reducing the FID by approximately 3%. This indicates that enhanced feature propagation helps stabilize the denoising process, although the benefit remains limited when the component is used in isolation. Introducing the attention mechanism produces a more noticeable gain, lowering the FID by about 11% relative to the baseline, which highlights the importance of spatially selective feature modeling for thin and sparse crack structures.

Among the three components, adaptive noise-aware scaling provides the most significant single-module improvement. When applied individually, it reduces the FID by roughly 17% compared with the baseline while also exhibiting lower performance variance across runs. A consistent downward trend can also be observed in the KID metric, suggesting improved generation stability.

Further improvements are observed when adaptive scaling is combined with other components. Integrating it with skip connections or attention yields additional performance gains of approximately 20% and 21% relative to the baseline, respectively. When all three components are jointly employed, the full model achieves the best performance, corresponding to an overall reduction of about 21% in FID compared with the baseline. These results demonstrate that the proposed modules contribute complementary benefits rather than redundant functionality, with adaptive noise-aware scaling playing a particularly important role in stabilizing diffusion-based crack synthesis.

5.2. Validation of joint appearance–annotation diffusion modeling

To validate the core novelty of CrackMaskNet, namely the explicit joint modeling of appearance and annotation under a shared diffusion trajectory, we compare the proposed framework with representative conditional and translation-based paradigms. Unlike approaches that treat annotation as an external condition or generate it in a separate stage, CrackMaskNet jointly denoises images and masks under identical noise levels. This comparison evaluates whether such shared stochastic evolution leads to improved image fidelity and stronger image–mask consistency.

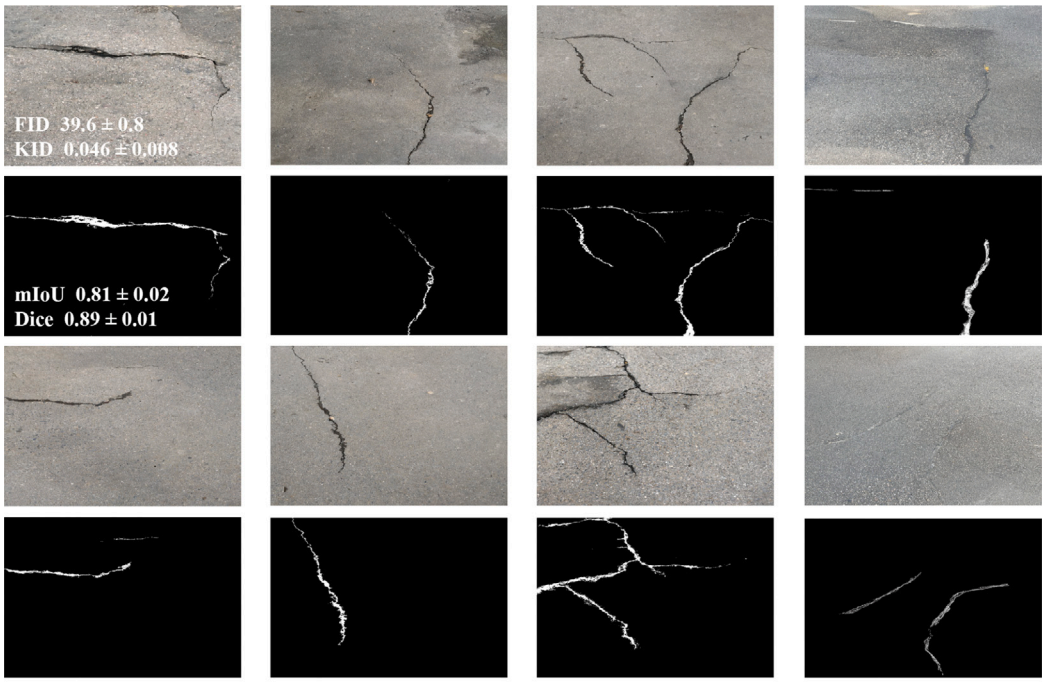


Fig. 3. Experimental results. Generated pavement distress images and corresponding segmentation masks.

The first alternative strategy adopts a sequential generation pipeline. An image-only diffusion model was constructed by removing the mask prediction branch from CrackMaskNet while retaining the same backbone architecture, noise schedule, and training configuration. This model generates pavement images from noise. Subsequently, a state-of-the-art segmentation model (FastSAM X. Zhao et al., 2023) trained on the real dataset was applied to predict masks for the generated images. The quality of generated images was evaluated using FID and KID, while mask accuracy was measured using mIoU by manually annotating a subset of generated images as reference ground truth.

The second alternative strategy follows conditional and translation-based paradigms. ControlNet (Zhang et al., 2023) and pix2pix (Isola et al., 2017) were trained using paired mask–image data from the dataset to perform mask-to-image generation. In this setting, the mask is treated as an externally provided structural condition rather than being jointly synthesized. Generated images were evaluated using FID and KID, and structural consistency was assessed by comparing the generated image against the input mask using manually verified mask annotations.

Table 3 summarizes the quantitative results. Compared with CrackMaskNet, the sequential diffusion–segmentation pipeline achieves comparable image generation quality, with only marginally higher FID and KID (approximately 3% and 6%, respectively). However, a clear degradation is observed in structural accuracy. The resulting mask quality is about 25% lower in mIoU, indicating that the predicted crack structures are often inconsistent with the generated image content. This limitation mainly stems from the segmentation stage. Since the segmentation model is trained on a relatively limited amount of real annotated data, its generalization to synthesized images is constrained. Consequently, thin crack branches and low-contrast regions are frequently missed or fragmented during mask prediction.

For conditional generation models, both ControlNet and pix2pix exhibit inferior performance compared with the proposed joint diffusion framework. Compared with ControlNet, CrackMaskNet achieves lower FID and KID values, decreasing from 43.6 to 39.6 and from 0.053 to 0.046, respectively. A more substantial difference is observed in structural consistency, where the mIoU increases from 0.70 to 0.81.

Table 3
Comparison with conditional and sequential generation frameworks.

Model	FID ↓	KID ↓	mIoU ↑
Sequential Diffusion + FastSAM	40.8 ± 0.9	0.049 ± 0.009	0.61 ± 0.04
ControlNet (Mask-to-Image)	43.6 ± 0.8	0.053 ± 0.010	0.70 ± 0.03
pix2pix (Mask-to-Image)	48.2 ± 1.1	0.081 ± 0.012	0.64 ± 0.02
Ours (CrackMaskNet)	**39.6 ± 0.8**	**0.046 ± 0.008**	**0.81 ± 0.02**

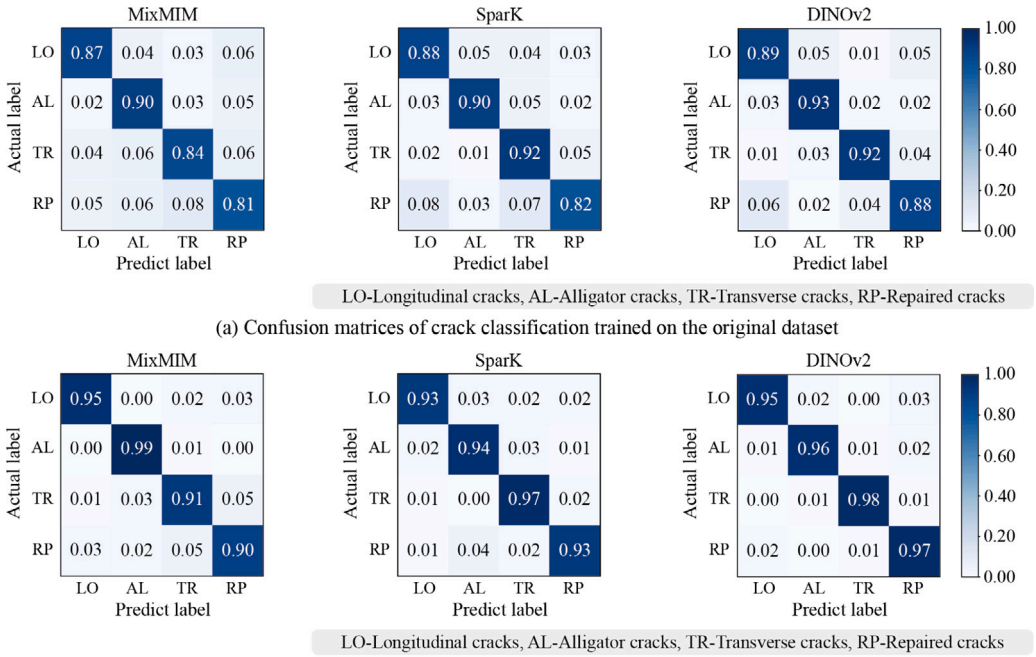
The performance gap becomes more pronounced when compared with pix2pix, which produces noticeably higher FID and KID values together with lower structural accuracy, with the mIoU decreasing to 0.64.

Although conditional methods enforce structural guidance through an input mask, the generated crack patterns do not always align with the intended structure. For ControlNet, the overall crack layout is roughly preserved, but noticeable spatial deviations and weakened fine branches can be observed. In the case of pix2pix, the mismatch becomes more evident, where crack shapes are distorted and thin crack segments are partially missing.

These observations suggest that treating annotation purely as an external condition is insufficient to guarantee spatially consistent coupling between crack appearance and structural representation. In contrast, CrackMaskNet jointly models appearance and annotation under a shared diffusion trajectory, allowing both modalities to evolve coherently during denoising. This joint formulation helps preserve thin crack continuity and maintain consistent boundary structures between generated images and masks.

5.3. Application in downstream pavement analysis tasks

The effectiveness of CrackMaskNet-generated images was evaluated across three pavement analysis tasks, including crack classification, detection, and segmentation. The real-world pavement dataset was divided into training and testing subsets following a 5:1 ratio, resulting in 1000 images used for model training and data augmentation and 200 images reserved exclusively for testing. To investigate the impact of synthetic data volume on downstream task performance, generated



(a) Confusion matrices of crack classification trained on the original dataset

(b) Confusion matrices of crack classification trained with CrackMaskNet-augmented data

Fig. 4. Confusion matrices of pavement crack classification under different data settings. (a) Confusion matrices of crack classification trained on the original dataset. (b) Confusion matrices of crack classification trained with CrackMaskNet-augmented data.

images were incrementally incorporated into the training set. The generation process followed a progressive augmentation strategy, where the number of synthetic samples was gradually increased relative to the original training set.

5.3.1. Crack classification

The performance of crack classification was evaluated using three representative models, including Mixed and Masked Image Modeling (MixMIM) (Liu et al., 2022), Sparse masKed modeling (SparK) (Tian et al., 2023), and DETR with Improved Denoising Anchor Boxes version 2 (DINOv2) (Oquab et al., 2023). Fig. 4(a) presents the confusion matrices obtained when the models were trained on the original dataset. Due to the pronounced class imbalance and the morphological similarity between repaired cracks and untreated crack patterns, the classification accuracy for repaired cracks is consistently lower than that of the other categories across all models. Misclassifications are primarily observed between repaired cracks and longitudinal or transverse cracks, indicating the difficulty of learning discriminative representations for this visually ambiguous and underrepresented class.

Fig. 4(b) shows the classification results after augmenting the training set with 6000 CrackMaskNet-generated images. A clear improvement in overall classification performance is observed for all three models. More importantly, the accuracy of the repaired crack category increases substantially, with a marked reduction in confusion with other crack types. This improvement demonstrates that the generated data effectively compensates for the limited availability of repaired crack samples and enhances the model's ability to distinguish visually similar distress categories.

These results confirm that the proposed generative augmentation strategy improves overall classification accuracy while delivering pronounced gains for underrepresented and challenging categories, leading to more robust pavement crack classification performance.

5.3.2. Crack detection

The effect of CrackMaskNet-generated images on crack detection performance was evaluated using three representative detection frameworks, namely YOLOv12 (Tian et al., 2025), RF-DETR (Sapkota et al., 2025), and GLIP (L.H. Li et al., 2022). Fig. 5(a) illustrates the variation of detection performance in terms of F1 score and mean Average Precision (mAP) as a function of both the number and proportion of generated images added to the training set.

Overall, all three detection models benefit consistently from the progressive incorporation of synthetic images. For YOLOv12, the F1 score increases from 0.77 on the original dataset to 0.96 when the number of generated images reaches 6000, while mAP improves from 0.46 to 0.72. A clear performance inflection point is observed at approximately 2000 generated images, corresponding to a synthetic-to-real ratio of about 2:1. Beyond this point, further gains become marginal, indicating diminishing returns from additional augmentation.

RF-DETR exhibits a similar trend, with F1 scores improving from 0.69 to 0.90 and mAP values increasing from 0.38 to 0.62 as the proportion of generated data increases. The most pronounced improvements occur when the synthetic data ratio increases up to approximately 200–300 percent of the real training set. GLIP also demonstrates steady performance gains. Compared with YOLOv12 and RF-DETR, GLIP shows a slightly delayed saturation behavior, reaching its performance plateau at a higher augmentation level of around 4000 generated images. These observations indicate that, for crack detection tasks, an effective balance between real and synthetic data is achieved when the amount of generated images is approximately two to four times that of the real training samples.

5.3.3. Crack segmentation

The impact of synthetic data augmentation on crack segmentation was further examined using Fast segment anything (FastSAM) (X. Zhao et al., 2023), Visual Perception with pre-trained Diffusion models (VPD) (W. Zhao et al., 2023), and OneFormer (Jain et al., 2023). Fig. 5(b) reports the F1 score and mIoU as functions of the number and proportion of generated images included during training.

All segmentation models show clear performance improvements as the proportion of generated images increases. FastSAM achieves an increase in F1 score from 0.64 to 0.82 and an improvement in mIoU from 0.59 to 0.79. The most substantial gains occur when the number of generated images reaches approximately 2000, corresponding to a

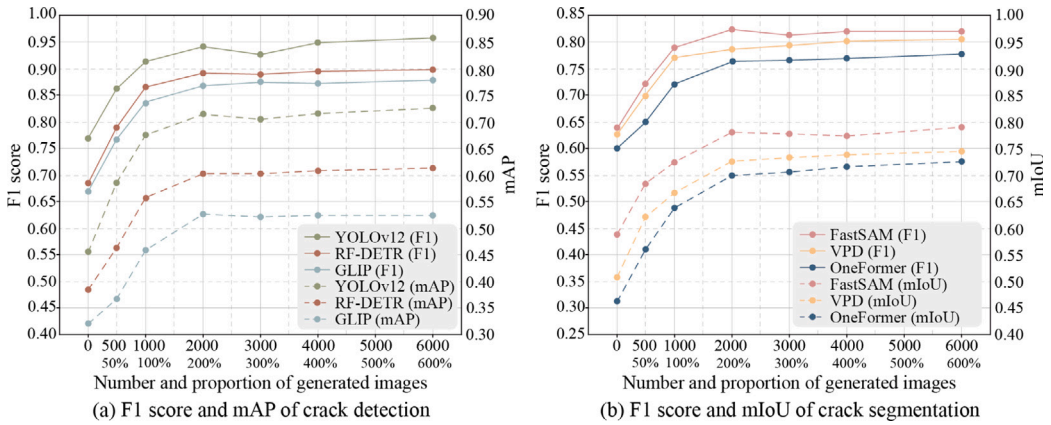


Fig. 5. Effect of progressive synthetic data augmentation on crack detection and segmentation performance. (a) F1 score and mAP of crack detection. (b) F1 score and mIoU of crack segmentation.

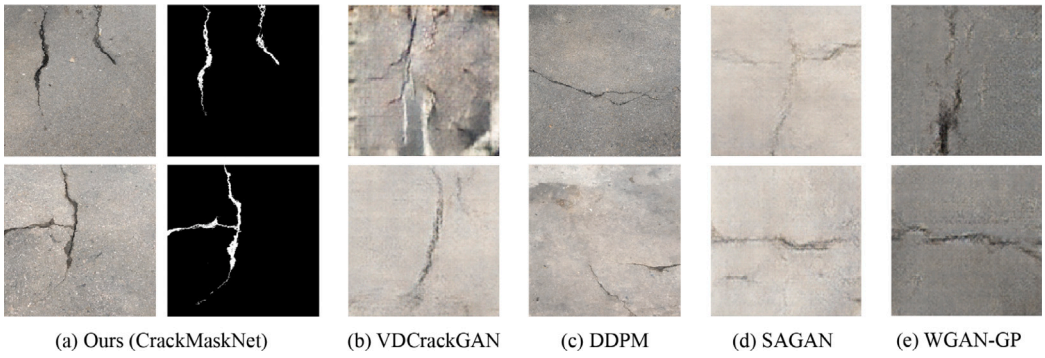


Fig. 6. Comparison with state-of-the-art generative methods. Qualitative comparison of pavement crack images generated by CrackMaskNet and representative state-of-the-art models, including VDCrackGAN, DDPM, SAGAN, and WGAN-GP. All methods are evaluated under identical dataset and training settings.

synthetic-to-real ratio of about 2:1. Beyond this point, performance improvements gradually stabilize.

VPD follows a similar trend, with F1 scores increasing from 0.62 to 0.81 and mIoU values rising from 0.51 to 0.74. OneFormer, which starts from a lower baseline, benefits noticeably from synthetic augmentation, achieving F1 and mIoU improvements from 0.60 to 0.78 and from 0.46 to 0.72, respectively. Compared with FastSAM and VPD, OneFormer exhibits a slower convergence pattern, with its performance plateau occurring at a higher augmentation level of approximately 4000 to 6000 generated images.

Based on these results, segmentation performance reaches a peak when the proportion of synthetic images is approximately two to three times that of the real training data. Beyond this range, additional synthetic samples provide diminishing returns, suggesting that excessive augmentation may be unnecessary for practical deployment.

5.4. Comparative analysis with state-of-the-art methods

The performance of the proposed CrackMaskNet was evaluated against four representative state-of-the-art generative models, including VDCrackGAN (Yu et al., 2024), Denoising Diffusion Probabilistic Models (DDPM) (Cano-Ortiz et al., 2024a), Self-Attention Generative Adversarial Networks (SAGAN) (Yao et al., 2024) and WGAN-GP (Zhong et al., 2023). All models were trained and evaluated under the same dataset and experimental settings to ensure a fair comparison. Fig. 6 provides qualitative comparisons of the generated pavement crack images, while quantitative results in terms of Fréchet Inception Distance (FID), Kernel Inception Distance (KID) and generation efficiency are summarized in Table 4.

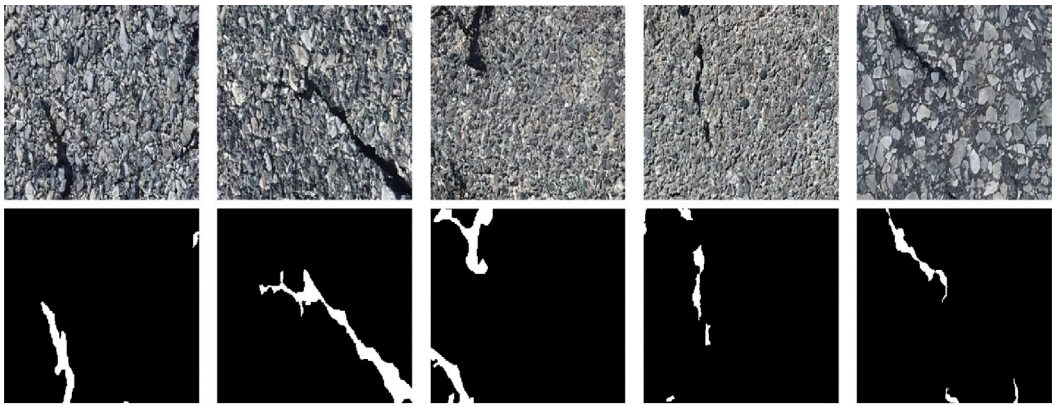
As shown in Table 4, CrackMaskNet achieves the lowest FID and KID scores among all compared methods, with an FID of 39.6 ± 0.8 and a KID of 0.046 ± 0.008, indicating superior alignment between generated and real image distributions. In contrast, GAN-based methods such as VDCrackGAN, SAGAN, and WGAN-GP exhibit substantially higher FID and KID values, suggesting limited capability in capturing the complex texture and structural characteristics of real pavement distresses. DDPM improves generation fidelity compared with GAN-based baselines but remains inferior to CrackMaskNet in both FID and KID.

Table 4
Comparison of generation quality and efficiency with state-of-the-art methods.

Model	FID	KID	Time (ms)
Ours (CrackMaskNet)	**39.6 ± 0.8**	**0.046 ± 0.008**	3154
VDCrackGAN	222.5 ± 5.2	0.279 ± 0.013	**0.93**
DDPM	62.5 ± 1.4	0.157 ± 0.010	70 682
SAGAN	181.9 ± 3.9	0.245 ± 0.010	1.60
WGAN-GP	171.3 ± 2.7	0.236 ± 0.011	1.06

In terms of generation efficiency, GAN-based models demonstrate significantly faster inference speed due to their single-pass generation mechanism. VDCrackGAN, SAGAN, and WGAN-GP require less than 2 ms per image, whereas CrackMaskNet requires 3154 ms per sample, and DDPM exhibits the highest computational cost. This difference is expected, as diffusion-based models rely on iterative denoising processes, and CrackMaskNet further extends this framework to simultaneously generate both pavement images and their corresponding pixel-level segmentation masks.

Despite higher generation time, CrackMaskNet offers a practical advantage by producing image–mask pairs in a single unified process. In contrast, all compared baselines generate images only and require



(a) Synthetic crack images with corresponding segmentation masks on the CRACK500 dataset

(b) Cross-dataset generation of road pothole images and masks on the Indian pothole dataset

(c) Person image and silhouette mask generation for traffic re-identification tasks

Fig. 7. Cross-domain generalization of CrackMaskNet for image–mask joint generation. (a) Synthetic crack images with corresponding segmentation masks on the CRACK500 dataset. (b) Cross-dataset generation of road pothole images and masks on the Indian pothole dataset. (c) Person image and silhouette mask generation for traffic re-identification tasks.

additional manual annotation or separate segmentation models to obtain pixel-level labels. When considering the end-to-end cost of dataset construction for downstream detection and segmentation tasks, CrackMaskNet provides a more efficient and scalable solution by eliminating the need for post-generation annotation.

5.5. Generalization and applicability across domains

To evaluate the generalization capability of CrackMaskNet beyond the training dataset, additional experiments were conducted across three progressively challenging domains. These include a public pavement crack dataset within the same domain CRACK500 (Lau et al., 2020), a cross-dataset road pothole dataset, and a structurally distinct person re-identification dataset used in traffic monitoring. Fig. 7 presents qualitative examples of the generated image–mask pairs, while quantitative results are reported using FID and KID to assess generation fidelity across domains.

5.5.1. Generalization evaluation on the public CRACK500 dataset

To examine in-domain generalization within pavement crack analysis, CrackMaskNet was evaluated on the public CRACK500 dataset (Lau et al., 2020). This dataset contains crack patterns with different surface textures, lighting conditions, and acquisition settings compared with the training data, providing a meaningful benchmark for robustness within the same semantic domain.

As shown in Fig. 7(a), CrackMaskNet successfully generates crack images that preserve realistic pavement textures and crack morphologies, together with accurately aligned segmentation masks. Quantitatively, the model achieves an FID score of 36.6 ± 0.6 and a KID value of 0.044 ± 0.007 on CRACK500, which are comparable to the results obtained on the original dataset. These results indicate that the proposed diffusion-based framework maintains stable generation quality under moderate domain shifts while preserving precise pixel-level annotations.

5.5.2. Application to road pothole generation

To further assess cross-domain adaptability within transportation infrastructure monitoring, CrackMaskNet was applied to the Indian Pothole Dataset (Rahman and Patel, 2020). Unlike cracks, potholes exhibit larger spatial extents, irregular boundaries, and more pronounced texture variations, posing a greater challenge for joint image–mask generation.

As illustrated in Fig. 7(b), the model captures the heterogeneous appearance of potholes, including complex edge structures and transitions between damaged and intact pavement regions. The generated segmentation masks remain well aligned with the pothole regions, enabling

direct use as pixel-level annotations. Quantitative evaluation yields an FID score of 53.1 ± 1.1 and a KID value of 0.117 ± 0.010. Although these values indicate a decrease in generation fidelity compared with crack datasets, the results remain consistent with realistic pothole appearance and demonstrate effective cross-dataset generalization under increased structural variability.

5.5.3. Extension to person re-identification in traffic monitoring

To evaluate generalization beyond pavement analysis, CrackMaskNet was further applied to person image generation for traffic re-identification tasks using a dedicated dataset (Scharwächter et al., 2013). This domain differs substantially from pavement-related imagery in terms of object structure, semantic complexity, and appearance variation. Fig. 7(c) shows that the model is capable of generating person images together with corresponding silhouette masks that preserve body contours and pose information. The generated samples maintain coarse identity-related characteristics such as posture and clothing layout while introducing realistic appearance variations. Quantitatively, this task results in an FID score of 68.7 ± 1.4 and a KID value of 0.183 ± 0.012, reflecting the increased difficulty of modeling articulated human structures and identity-consistent features.

The cross-domain evaluation indicates a consistent generalization pattern. As the target domain increasingly diverges from pavement crack imagery, generation fidelity gradually decreases, as indicated by higher FID and KID values, while the ability to produce spatially aligned image–mask pairs remains stable without architectural modification. These results suggest that the diffusion-based design of CrackMaskNet is robust to moderate domain shifts and remains effective under increased structural and semantic variability. Overall, CrackMaskNet provides a flexible framework for joint image–mask generation across multiple transportation-related visual analysis tasks, achieving a trade-off between generation fidelity and domain complexity.

6. Conclusion

This study proposes CrackMaskNet, a diffusion-based generative framework for pavement distress analysis that targets three practical barriers in automated inspection. These barriers include limited and imbalanced labeled data, high pixel-level annotation cost, and misclassification caused by visual similarity between distress types. The framework formulates pavement appearance and pixel-level annotation as coupled stochastic variables within a unified diffusion process. By jointly modeling both modalities under a shared noise trajectory, the method enables coordinated evolution of visual texture and structural boundaries, allowing class-controllable synthesis of image–mask pairs for direct use in supervised training pipelines.

A central aspect of the framework is the joint denoising mechanism, which refines crack appearance and geometry synchronously under identical diffusion states and label conditioning. This formulation treats segmentation masks as intrinsic generative components rather than external guidance signals, ensuring spatial consistency throughout sampling. Quantitative evaluation on the dataset shows that the generated images achieve an FID of 39.6 ± 0.8 and a KID of 0.046 ± 0.008, while the generated masks reach an mIoU of 0.81 ± 0.02 and a Dice score of 0.89 ± 0.01, indicating coherent image–mask synthesis. The ablation study further suggests that multi-scale skip connections, attention, and adaptive noise-aware scaling contribute complementarily to generation quality, with adaptive scaling providing the largest single-component improvement.

The practical value of the generated data is examined through downstream classification, detection, and segmentation tasks under progressive augmentation. Across representative models, consistent performance gains are observed as synthetic samples are added, with the largest improvements appearing when the synthetic-to-real ratio is approximately two to four for detection and approximately two to three for segmentation. Improvements are particularly evident for the repaired crack category, which is both underrepresented and visually similar to untreated cracks, reducing confusion in multi-class learning settings. In comparisons with representative generative baselines, CrackMaskNet yields lower FID and KID while requiring longer per-sample generation time, reflecting the iterative nature of diffusion sampling and the additional requirement of joint mask generation. However, producing image–mask pairs in a single process reduces reliance on post-generation manual labeling.

Additional experiments evaluate robustness beyond the training distribution. CrackMaskNet maintains comparable fidelity on the CRACK500 dataset with an FID of 36.6 ± 0.6 and a KID of 0.044 ± 0.007. The framework also remains capable of generating aligned image–mask pairs for potholes and person silhouettes under larger domain shifts, with expected decreases in image fidelity as domain complexity increases.

Future work will focus on extending class coverage to additional distress types, improving controllability for visually ambiguous categories, and reducing sampling time while preserving image–mask consistency.

CRediT authorship contribution statement

Yuanyuan Hu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Hancheng Zhang:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Meng Guo:** Writing – review & editing, Validation, Formal analysis, Data curation. **Pengfei Liu:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Formal analysis.

Declaration of competing interest

We declare that we have no financial and personal relationship with other people or organizations that can inappropriately influence our work. There is no professional or other personal interest of any nature or kind in any product, service or company that could be constructed as influencing the manuscript.

Acknowledgment

This paper is funded by German Research Foundation (DFG) under LI 3613/7-2 with Project ID 459436571. The language and grammar of this manuscript were partially refined using a large language model (Claude).

Data availability

The data and code supporting this study have been shared in the manuscript.

References

Abbas, I.H., Ismael, M.Q., 2021. Automated pavement distress detection using image processing techniques. Eng. Technol. Appl. Sci. Res. 11 (5), 7702–7708.

Ashwini, K., Nenavath, H., Jatoth, R.K., 2024. EPQ-GAN: Evolutionary perceptual quality assessment generative adversarial network for image dehazing. IEEE Trans. Intell. Transp. Syst..

Ayman, H., Fakhr, M.W., 2023. Recent computer vision applications for pavement distress and condition assessment. Autom. Constr. 146, 104664.

Bacsa, K., Liu, W., Abdallah, I., Chatzi, E., 2025. Structural dynamics feature learning using a supervised variational autoencoder. J. Eng. Mech. 151 (2), 04024106.

Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., 2018. Demystifying mmd gans. arXiv preprint arXiv:1801.01401.

Bredell, G., Flouris, K., Chaitanya, K., Erdil, E., Konukoglu, E., 2023. Explicitly minimizing the blur error of variational autoencoders. arXiv preprint arXiv:2304.05939.

Cano-Ortiz, S., Iglesias, L.L., del Árbol, P.M.R., Castro-Fresno, D., 2024a. Improving detection of asphalt distresses with deep learning-based diffusion model for intelligent road maintenance. Dev. Built Environ. 17, 100315.

Cano-Ortiz, S., Sainz-Ortiz, E., Lloret Iglesias, L., Martínez Ruiz del Árbol, P., Castro-Fresno, D., 2024b. Leveraging a deep learning generative model to enhance recognition of minor asphalt defects. Sci. Rep. 14 (1), 28904.

Chatterjee, A., Ghosh, S., Ghosh, A., 2025. Context-aware masking and learnable diffusion-guided patch refinement in transformers via sparse supervision for hyperspectral image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2906–2915.

Chen, J., He, Q., Liu, P., Ma, W., Pu, Z., 2025. Enhancing crash frequency modeling based on augmented multi-type data by hybrid VAE-diffusion-based generative neural networks. arXiv preprint arXiv:2501.10017.

Chen, P., Liu, S., Zhao, H., Jia, J., 2020. Gridmask data augmentation. arXiv preprint arXiv:2001.04086.

Chen, T., Ren, J., 2024. Integrating GAN and texture synthesis for enhanced road damage detection. IEEE Trans. Intell. Transp. Syst..

Croitoru, F.-A., Hondru, V., Ionescu, R.T., Shah, M., 2023. Diffusion models in vision: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 45 (9), 10850–10869.

Cui, L., Qi, Z., Chen, Z., Meng, F., Shi, Y., 2015. Pavement distress detection using random decision forests. In: International Conference on Data Science. Springer, pp. 95–102.

Delbracio, M., Milanfar, P., 2023. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. arXiv preprint arXiv:2303.11435.

Fawzi, A., Samulowitz, H., Turaga, D., Frossard, P., 2016. Adaptive data augmentation for image classification. In: 2016 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 3688–3692.

Gong, F., Cheng, X., Fang, B., Cheng, C., Liu, Y., You, Z., 2023. Prospect of 3D printing technologies in maintenance of asphalt pavement cracks and potholes. J. Clean. Prod. 397, 136551.

Gopalakrishnan, K., Khaitan, S.K., Choudhary, A., Agrawal, A., 2017. Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection. Constr. Build. Mater. 157, 322–330.

Han, C., Ma, T., Huyan, J., Tong, Z., Yang, H., Yang, Y., 2024. Multi-stage generative adversarial networks for generating pavement crack images. Eng. Appl. Artif. Intell. 131, 107767.

Hu, Z., He, Y., Shen, Y., Jo, M., Collotta, M., Shen, G., Kong, X., 2024. Vehicular social dynamic anomaly detection with recurrent mulit-mask aggregator enabled VAE. IEEE Trans. Intell. Transp. Syst..

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.

Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H., 2023. Oneformer: One transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2989–2998.

Jing, L., Aiqin, Z., 2010. Pavement crack distress detection based on image analysis. In: 2010 International Conference on Machine Vision and Human-Machine Interface. IEEE, pp. 576–579.

Karras, T., Aittala, M., Aila, T., Laine, S., 2022. Elucidating the design space of diffusion-based generative models. Adv. Neural Inf. Process. Syst. 35, 26565–26577.

Kim, D., Lee, J., 2022. Predictive evaluation of spectrogram-based vehicle sound quality via data augmentation and explainable artificial intelligence: Image color adjustment with brightness and contrast. Mech. Syst. Signal Process. 179, 109363.

Lau, S.L., Chong, E.K., Yang, X., Wang, X., 2020. Automated pavement crack segmentation using u-net-based convolutional neural network. IEEE Access 8, 114892–114899.

Li, Y., Che, P., Liu, C., Wu, D., Du, Y., 2021. Cross-scene pavement distress detection by a novel transfer learning framework. Comput. Aided Civ. Infrastruct. Eng. 36 (11), 1398–1415.

Li, D., Duan, Z., Hu, X., Zhang, D., Zhang, Y., 2023. Automated classification and detection of multiple pavement distress images based on deep learning. J. Traffic Transp. Eng. (Engl. Ed.) 10 (2), 276–290.

Li, G., Wan, J., He, S., Liu, Q., Ma, B., 2020. Semi-supervised semantic segmentation using adversarial learning for pavement crack detection. IEEE Access 8, 51446–51459.

Li, J., Yin, G., Wang, X., Yan, W., 2022. Automated decision making in highway pavement preventive maintenance based on deep learning. Autom. Constr. 135, 104111.

Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al., 2022. Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975.

Liu, J., Huang, X., Yu Liu, H.L., 2022. MixMIM: Mixed and masked image modeling for efficient visual representation learning. arXiv:2205.13137.

Liu, F., Zhang, S., Wang, X., Wei, Y., Qiu, H., Zhao, Y., Zhang, Y., Ye, Q., Wan, F., 2025. Timestep embedding tells: It's time to cache for video diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7353–7363.

Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., Omata, H., 2018. Road damage detection and classification using deep neural networks with smartphone images. Comput. Aided Civ. Infrastruct. Eng. 33 (12), 1127–1141.

Nair, P., Subha, V., Aneesh, R., 2018. Non verbal behaviour analysis for distress detection using texture analysis. In: 2018 International CET Conference on Control, Communication, and Computing. IC4, IEEE, pp. 239–244.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

Radopoulou, S.C., Jog, G.M., Brilakis, I., 2013. Patch distress detection in asphalt pavement images. In: ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction. Vol. 30, IAARC Publications, p. 1.

Rahman, A., Patel, S., 2020. Annotated potholes image dataset. http://dx.doi.org/10.34740/KAGGLE/DSV/973710, URL https://www.kaggle.com/dsv/973710.

Ren, R., Shi, P., Jia, P., Xu, X., 2023. A semi-supervised learning approach for pixel-level pavement anomaly detection. IEEE Trans. Intell. Transp. Syst. 24 (9), 10099–10107.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 658–666.

Sapkota, R., Cheppally, R.H., Sharda, A., Karkee, M., 2025. RF-DETR object detection vs YOLOv12: A study of transformer-based and CNN-based architectures for single-class and multi-class greenfruit detection in complex orchard environments under label ambiguity. arXiv preprint arXiv:2504.13099.

Scharwächter, T., Enzweiler, M., Franke, U., Roth, S., 2013. Efficient multi-cue scene segmentation. In: German Conference on Pattern Recognition. Springer, pp. 435–445.

Schreiner, C., Torkkola, K., Gardner, M., Zhang, K., 2006. Using machine learning techniques to reduce data annotation time. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 50, SAGE Publications Sage CA, Los Angeles, CA, pp. 2438–2442.

Setiawan, A.W., 2020. Image segmentation metrics in skin lesion: accuracy, sensitivity, specificity, dice coefficient, Jaccard index, and Matthews correlation coefficient. In: 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia. CENIM, IEEE, pp. 97–102.

Shi, Y., Cui, L., Qi, Z., Meng, F., Chen, Z., 2016. Automatic road crack detection using random structured forests. IEEE Trans. Intell. Transp. Syst. 17 (12), 3434–3445.

Song, J., Meng, C., Ermon, S., 2020. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR.

Tian, K., Jiang, Y., Diao, Q., Lin, C., Wang, L., Yuan, Z., 2023. Designing BERT for convolutional networks: Sparse and hierarchical masked modeling. arXiv:2301.03580.

Tian, Y., Ye, Q., Doermann, D., 2025. Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524.

Wang, S., Cai, B., Wang, W., Li, Z., Hu, W., Yan, B., Liu, X., 2024. Automated detection of pavement distress based on enhanced YOLOv8 and synthetic data with textured background modeling. Transp. Geotech. 48, 101304.

Wang, K.C., Li, Q., Gong, W., 2007. Wavelet-based pavement distress image edge detection with a trous algorithm. Transp. Res. Rec. 2024 (1), 73–81.

Wang, X., Liu, H., 2024. VAE-ResNet cascade network: An advanced algorithm for stochastic clutter suppression in ground penetrating radar data. IEEE Trans. Geosci. Remote Sens..

Wen, T., Ding, S., Lang, H., Lu, J.J., Yuan, Y., Peng, Y., Chen, J., Wang, A., 2023. Automated pavement distress segmentation on asphalt surfaces using a deep learning network. Int. J. Pavement Eng. 24 (2), 2027414.

Wu, S., Zhang, H., Valiant, G., Ré, C., 2020. On the generalization effects of linear transformations in data augmentation. In: International Conference on Machine Learning. PMLR, pp. 10410–10420.

Xi, Y., Zheng, J., Li, X., Xu, X., Ren, J., Xie, G., 2018. SR-POD: sample rotation based on principal-axis orientation distribution for data augmentation in deep object detection. Cogn. Syst. Res. 52, 144–154.

Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X., Ling, H., 2019. Feature pyramid and hierarchical boosting network for pavement crack detection. IEEE Trans. Intell. Transp. Syst. 21 (4), 1525–1535.

Yao, H., Wu, Y., Liu, S., Liu, Y., Xie, H., 2024. A pavement crack synthesis method based on conditional generative adversarial networks. Math. Biosci. Eng. 21 (1), 903–923.

Yu, G., Zhou, X., Chen, X., 2024. VDCrackGAN: A generative adversarial network with transformer for pavement crack data augmentation. Appl. Sci. 14 (17), 7907.

Zhang, H., Chen, N., Li, M., Mao, S., 2024a. The crack diffusion model: An innovative diffusion-based method for pavement crack detection. Remote. Sens. 16 (6), 986.

Zhang, Z., Li, X., Liu, Y., Wang, Y., Chen, Y., Xue, T., Guo, S., 2025. Egvd: Event-guided video diffusion model for physically realistic large-motion frame interpolation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3923–3932.

Zhang, H., Qian, Z., Zhou, W., Min, Y., Liu, P., 2024b. A controllable generative model for generating pavement crack images in complex scenes. Comput. Aided Civ. Infrastruct. Eng. 39 (12), 1795–1810.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847.

Zhang, Y., Zhang, L., 2024. A generative adversarial network approach for removing motion blur in the automatic detection of pavement cracks. Comput. Aided Civ. Infrastruct. Eng. 39 (22), 3412–3434.

Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J., 2023. Fast segment anything. arXiv preprint arXiv:2306.12156.

Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J., 2023. Unleashing text-to-image diffusion models for visual perception. In: ICCV.

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. Eng. Appl. Artif. Intell. 119, 105808.