Results in Engineering 23 (2024) 102745

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Enhancing pavement crack segmentation via semantic diffusion synthesis model for strategic road assessment

Saúl Cano-Ortiz [a,*], Eugenio Sainz-Ortiz [a], Lara Lloret Iglesias [b], Pablo Martínez Ruiz del Árbol [b], Daniel Castro-Fresno [a]

[a] GITECO Research Group, Universidad de Cantabria, 39005, Santander, Spain
[b] Institute of Physics of Cantabria (UC-CSIC), 39005, Santander, Spain

ARTICLE INFO

Keywords:
Pavement crack segmentation
Generative artificial intelligence
Semantic diffusion synthesis
Road maintenance
Deep learning

ABSTRACT

Computer-aided deep learning has significantly advanced road crack segmentation. However, supervised models face challenges due to limited annotated images. There is also a lack of emphasis on deriving pavement condition indices from predicted masks. This article introduces a novel semantic diffusion synthesis model that creates synthetic crack images from segmentation masks. The model is optimized in terms of architectural complexity, noise schedules, and condition scaling. The optimal architecture outperforms state-of-the-art semantic synthesis models across multiple benchmark datasets, demonstrating superior image quality assessment metrics. The synthetic frames augment these datasets, resulting in segmentation models with significantly improved efficiency. This approach enhances results without extensive data collection or annotation, addressing a key challenge in engineering. Finally, a refined pavement condition index has been developed for automated end-to-end defect detection systems, promoting more effective maintenance planning.

1. Introduction

Industrial advancements and transport infrastructure improvements face constraints such as material ageing, atmospheric conditions, and traffic loads [1]. In road infrastructure, prompt crack detection and repair are crucial. These actions prevent further degradation, extend service life, and minimize costs for maintenance organizations [2]. Conversely, road users reap the benefit of enhanced driving safety, improved comfort, elevated quality of passage with minimal disruptions, and reduced accident risks stemming from road defects [3]. A well-maintained road network is vital for economic prosperity [4]. Therefore, surface pavement distress detection is essential in Structural Health Monitoring, providing an initial assessment of road conditions [5].

Current practices still rely on manual visual inspection by qualified engineers. This incurs significant costs and extended durations for pavement distress detection [6]. Thus, computer-aided visual inspection techniques, mainly based on Deep Learning (DL), have gained significant interest worldwide [7,8] in the field of intelligent road damage detection [9,10]. In surface distress recognition, DL models primarily tackle the following Computer Vision (CV) tasks: classification (26 %), object detection (25 %), and instance (9 %)/semantic (37 %) segmentation [5,11]. Classification focuses on categorizing samples at the image level. Object detection aims to locate and classify multiple instances of distress within an image. Segmentation refers to classification at the pixel level.

Road crack segmentation, a leading trend in DL-based defect inspection due to the prevalence of cracks, provides essential information about crack location and topology for pavement condition analysis [12]. The current state-of-the-art predominantly focuses on advanced segmentation architectures, with a particular interest in binary crack segmentation [13]. The most implemented models are primarily built upon DL-based convolutional neural networks (CNNs). These networks are mainly encoder-decoder-based, such as improved U-Net [12], asymmetric dual-decoder U-Net [2,14,15], ARD-U-Net [16], APF-Net [17], PCSNet [18], improved FasterNet [19], and multi-fusion U-Net [20,21], among others. Additionally, there is a growing interest in the development of segmentation networks utilizing visual transformer layers [22–24]. Unfortunately, the application of DL techniques to real-world pavement crack segmentation faces several challenges.

Collecting high-quality, diverse, and large-scale images is challenging but crucial for ensuring the robustness and generalizability of

[*] Corresponding author.
E-mail address: saul.cano@alumnos.unican.es (S. Cano-Ortiz).

https://doi.org/10.1016/j.rineng.2024.102745
Received 18 June 2024; Received in revised form 9 August 2024; Accepted 14 August 2024
Available online 18 August 2024

data-driven models [25]. Such models require thousands of varied and realistic images to perform effectively. Many state-of-the-art models rely on supervised learning, which demands the labour-intensive creation of segmentation masks for each raw image. This manual annotation process is not only costly but also impractical for engineering applications [26]. Traditional solutions often involve transfer learning and conventional data augmentation. Transfer learning depends on pre-trained models from large datasets, which may not be available or suitable for specific CV tasks [27]. Conventional data augmentation methods, such as rotation and brightness shifting, may fall short in providing the necessary realism and diversity [28]. To address these limitations, emerging approaches utilize generative models [29]. These models can generate realistic and diverse synthetic images, thereby enhancing datasets and reducing the need for manual annotation and extensive data collection.

The most used generative algorithms in the sub-field of road crack generation are variants of Generative Adversarial Networks (GANs) [30] and Variational Autoencoders (VAEs) [31]. A GAN includes a generator that creates high-definition samples from random noise and a discriminator that differentiates between real and generated images. In contrast, a VAE features an encoder that compresses input data into a lower-dimensional latent space and a decoder that reconstructs the original data from this compressed representation. Although there have been fewer implementations in road crack generation, diffusion models have also been explored in this area [31]. A diffusion model [32] operates in two stages: forward and reverse. In the forward stage, noise is progressively added to images based on a noise schedule. During the reverse stage, a network predicts and removes the noise from the images, producing high-quality frames. The following review explores the scarce research on applying generative models to crack image generation.

Xu et al. [33] introduced a dual-step system using a Deep Convolutional GAN (DCGAN) [34] to generate synthetic crack images and a classification CNN model trained on the augmented dataset. Similarly [35], used a DCGAN for synthetic pavement crack image generation and trained an enhanced VGG16 model on the augmented data. Ma et al. [36] employed a GAN to create synthetic crack images, which were then used to augment the dataset for training a YOLOv3 model. In Ref. [9], a VAE was trained to produce crack images, integrating its encoder with the DCGAN generator to use learned latent vectors instead of Gaussian noise, with improved performance validated by a Faster R–CNN network. Zhang et al. [37] developed a two-stage system combining a DCGAN for image augmentation and an attention-guided U-Net for crack segmentation [38]. proposed FeatureGAN, where an autoencoder learns real crack features, and masked Gaussian noise is added to real crack images for the GAN generator to produce synthetic cracks [26]. created multi-stage GANs based on Wasserstein GAN [39] to address crack pixel imbalance by increasing resolution sequentially. Likewise [40], developed a super-resolution GAN to enhance crack image resolution and validate classification results [41]. proposed a three-stage framework addressing sample imbalance by annotating images for crack/non-crack classification and using an encoder-DCGAN to generate crack images.

A key challenge with previous models is their lack of control over generated images, which complicates labelling for crack segmentation tasks and makes synthetic image annotation time-consuming and costly. This lack of control also hinders the ability to create images with specific attributes, limiting their use for tasks requiring precise control. Few studies address the need for conditional generative models to enhance road crack semantic segmentation. Semantic synthesis-oriented models aim to generate realistic images from segmentation maps. Yan et al. [42] introduced CycleADC-Net, which uses CycleGAN [43] for image-to-image translation from low-light to bright domains, followed by training an encoder-decoder segmentation network on this augmented dataset [44]. proposed SynCrack, which employs the Perlin noise algorithm to generate background images and integrates segmentation masks using a weight map strategy [28]. developed a semantic diffusion model that combines background images and segmentation maps to generate synthetic images with seamlessly integrated cracks for complex scenarios.

Conversely, assessing road condition is challenging in pavement crack segmentation and requires developing condition indices from semantic segmentation masks [45]. introduced an index derived from areas identified by an object detection model. However, this method faced precision issues with certain crack-like defects at the bounding box level [46]. proposed an index based solely on the pixel-level areas of predicted masks [47]. developed an index using segmentation masks and the mean crack width estimated via a skeleton-based algorithm. Despite these efforts, the latter two approaches are limited in their ability to provide detailed classifications of various crack types, resulting in indices that are restricted and less informative. Therefore, it is essential to create a pixel-level pavement condition index that incorporates the severity of each crack type, offering a more meaningful assessment for road authorities.

This study addresses key research gaps in crack segmentation, including the limitations of supervised architectures due to the lack of large, diverse, and high-quality datasets. It also aims to develop a conditional generative system for semantic synthesis, reducing the costs associated with manual labelling. Additionally, the study seeks to create a refined pavement condition index based on crack detections by deep learning models. This paper introduces a novel semantic diffusion synthesis model, named RoadPainter. RoadPainter generates a large volume of realistic and diverse images depicting pavement defects from segmentation masks, addressing labelling challenges through its conditional approach. The primary contributions are as follows.

a) Introducing a novel semantic diffusion synthesis with an improved encoder-decoder denoising network based on self-attention layers and spatially-adaptative (de)normalization modules.
b) Optimizing the RoadPainter model based on the architecture complexity, noise schedules, and classifier-free guidance scaling in terms of image quality assessment metrics and computational cost.
c) Demonstrating the enhanced performance of the improved semantic diffusion synthesis architecture compared to state-of-the-art generative models.
d) Validating the enhanced detection performance of various DL-based segmentation algorithms following synthetic image augmentation across multiple benchmark road datasets.
e) Engineering a fine-grained pavement condition index tailored from pixel-level crack segmentation masks.

The rest of the paper is organized as follows. A detailed exposition of state-of-the-art DL-based generative architectures alongside the proposed innovative model is given in Section 2. Section 3 then scrutinizes the model's optimization results, comparing its performance with existing models. Also, it explores its practical utility in intelligent road maintenance with a refined pavement condition index. Finally, Section 4 encapsulates the primary research findings, the limitations, and outlines future research directions.

2. Semantic road crack synthesis

In sub-section 2.1, a concise review of the state-of-the-art semantic synthesis architectures analyzed in this study will be provided for comparison with RoadPainter. The aim is to facilitate comprehension by a broader research audience. Thus far, we have delineated two novel models oriented towards road crack synthesis: SynCrack [44] and the Crack Diffusion Model (CDM) [28]. Given the limited application of semantic synthesis models to pavement cracks, CycleGAN [43] and Pix2Pix [48] have also been included.

2.1. State-of-the-art architectures

GANs consist of two networks: the generator (G) and the discriminator (D). G learns to create synthetic images resembling the real data distribution by transforming random noise. Meanwhile, D distinguishes between real and fake images as a binary classifier. Through adversarial training, G aims to fool D, while D strives to accurately classify real and synthetic images. G and D are typically CNN-based architectures. Once a GAN is trained, the pre-trained G is used as the generative system. A conditional GAN [49] generates fake images based not only on random noise but also on some input condition (e.g., text, semantic mask, etc.).

Pix2Pix is a conditional GAN, specially designed for image-to-image (I2I) translation. An I2I task is semantic synthesis. The G of Pix2Pix is a U-Net [50] with skip connections between mirrored layers, and its D corresponds to a PatchGAN classifier [51]. G receives the semantic mask as input and produces fake images. Then, D receives both real images with real labels and fake images (from the generator) with fake masks.

CycleGAN is capable of performing semantic synthesis in an unpaired manner, meaning it does not require paired image-mask samples. CycleGAN consists of four networks: a generator (G1) that maps masks to synthetic images with its corresponding discriminator (D1), and a generator (G2) that maps synthetic images back to masks with its associated discriminator (D2). The loss function in CycleGAN includes an adversarial loss, which encourages the generation of realistic images, and a cycle-consistency loss, which ensures fidelity between the original and translated images.

SynCrack is not a DL-based generative model; it is a methodology based on traditional image processing operations. First, it uses a workflow based on 2D Perlin noise to create background road images. Then, the segmentation mask is blended or fused with the background image using a weighting map strategy.

CDM is a generative semantic synthesis model based on Denoising Diffusion Probabilistic Models (DDPM). DDPMs operate in two stages: forward and reverse diffusion. During the forward diffusion process, Gaussian noise is gradually added to real images over T steps, referred to as timesteps. The noise inclusion during this process is defined by the noise schedule. Thanks to the reparametrization trick, the noised image at the final timestep can be computed in a single step. The reverse diffusion process aims to learn to denoise the noisy images from timestep T to 0. This process can be simplified: a neural network predicts the mean noise at a given t from the preceding timestep. The loss function at each timestep is the mean squared error (MSE) between the real noise computed during the forward process and the predicted noise. The analytical expressions for sampling new fake images from the pre-trained denoising network are detailed in Ref. [32]. CDM has one main difference from conventional DDPMs: the denoising network is a U-Net that receives as input the element-wise summation of the noisy crack image, the segmentation mask, and the background image (a real image with no defects). Consequently, the pre-trained CDM requires a background and mask image to sample synthetic frames.

The aforementioned networks present various potential limitations that will be discussed in the results. For example, GANs such as Pix2Pix and CycleGAN often suffer from mode collapse, where the generator produces a limited variety of images despite having different inputs. Pix2Pix may produce blurry images or fail to preserve fine details. CycleGAN, on the other hand, faces significant complexity due to the simultaneous training of four networks and may generate unrealistic artifacts, textures, or colors with limited datasets. SynCrack is constrained by a very limited repertoire of backgrounds, which prevents it from capturing real road elements such as highly textured surfaces or shadows. Lastly, CDM's major limitation is its requirement for background images for both training and inference, posing a significant constraint.

2.2. Our approach: RoadPainter

RoadPainter is an advanced conditional DDPM tailored for semantic road crack synthesis. While it utilizes the traditional forward diffusion process to generate noisy images as outlined in Ref. [32], its reverse diffusion process features several innovations. Notably, RoadPainter employs a novel architecture that deviates from the classic U-Net by integrating semantic information through the SPADE mechanism and utilizing ResBlocks and self-attention mechanisms for enhanced multimodal fusion of the semantic mask and crack image. Additionally, it introduces a new approach to incorporating conditional information during the reverse diffusion process (CFG). The following sub-sections will provide detailed insights into these innovations, covering SPADE, CFG, and architecture specifics.

2.2.1. Spatially adaptative (DE)Normalization (SPADE)

Previous generative architectures like Pix2Pix and CycleGAN primarily integrated the semantic layout directly with a noisy image input, which led to the problem of semantic washing. This term refers to the degradation of semantic information as it passes through the network layers, causing the generated images to lose their correlation with the original semantic masks. Consequently, while these models could produce realistic images, they struggled to maintain a strong alignment between the generated images and the intended semantic layouts.

SPADE addresses this issue by introducing a conditional normalization technique that maintains the semantic information throughout the image generation process. Mathematically, SPADE can be viewed as a form of conditional Batch Normalization where the conditioning information is the semantic mask. The process starts with Instance Normalization (InstanceNorm), which normalizes each spatial location independently, maintaining local spatial information. The semantic mask is processed through two convolutional layers to produce spatially adaptive γ and β tensors, ensuring that these parameters are tensors with the same spatial dimensions as the input feature maps. This spatial adaptability allows SPADE to retain and emphasize the semantic structure within the generated images, resulting in outputs that are not only realistic but also semantically consistent with the input masks. The SPADE method is expressed as:

$$x_{SPADE} = InstanceNorm(x) \odot \gamma + \beta \tag{1}$$

In Eq. (1), x refers to the input feature map. This element-wise multiplication and addition ensure that the semantic information is effectively incorporated at each spatial location, preserving the structure and enhancing the fidelity of the generated images.

2.2.2. Classifier-free guidance (CFG) for semantic image synthesis

In this study, a new CFG strategy [52] has been incorporated into the reverse diffusion process to improve conditional integration. This approach calculates the mean predicted noise through linear interpolation between the noise predicted from the conditional input ($\varepsilon_\theta(y_t)$)) and the noise predicted from the unconditional input (the crack image alone). The CFG scale, denoted as "s" in Eq. (2), governs this interpolation. Inspired by Ref. [53], the condition is defined by the semantic mask, while a null mask ($y_t = 0$) is used in non-conditional cases to represent a background image of a road without cracks. The mean predicted noise is reformulated as:

$$\varepsilon_\theta = \varepsilon_\theta(y_t) + s[\varepsilon_\theta(y_t) - \varepsilon_\theta(y_t = 0)] \tag{2}$$

The incorporation of CFG allows for refined control over the influence of the semantic mask on the generated output. This leads to improved quality and accuracy of the synthesized images, enhancing their fidelity and ensuring more effective balancing of semantic information.

2.2.3. *Multi-modal U-Net*

The novel architecture is illustrated in Fig. 1. The proposed U-Net architecture features an encoder composed of ResBlock modules [54]. Each ResBlock processes the noisy images (x_t) and timesteps (t) as inputs. The number of output channels for each block is determined by the model's channel configuration and a channel multiplication vector, which are specified hyperparameters. Noisy images are first passed through a 2D convolutional layer (Conv2D) for feature extraction, while timesteps are processed through a series of layers including positional embedding, dense layers, and a non-linear Sigmoid Linear Unit (SiLU) activation function.

In the encoder, the noisy input traverses two residual blocks, each comprising Group Normalization (GroupNorm2D), a SiLU activation function, and a down-sampling Conv2D layer. The second block incorporates a dropout layer between the SiLU activation and Conv2D to reduce overfitting. Meanwhile, encoded timesteps are processed through SiLU and dense layers to capture temporal dynamics. The output from the last residual block, which pertains to the noisy image, is combined with the corresponding timestep data using a scale-shift normalization technique, enhancing the feature-temporal correlation within the denoising network. To preserve critical low-level details, the original noisy image is added to the output tensor from the residual block before further propagation. Following a sequence of ResBlock modules, a self-attention mechanism [55] is employed to allow the encoder to assess the relevance of various elements in the noisy-temporal feature map, thereby facilitating the capture of long-range dependencies and improving contextual understanding.

The bottleneck of the network consists of two ResBlock modules interspersed with a self-attention mechanism. The ResBlock structure in this stage mirrors that used in the encoder, with two key modifications: the incorporation of SPADE for conditional normalization instead of GroupNorm2D, and the convolutional layers maintain the spatial dimensions without alteration. SPADE, as discussed previously, utilizes semantic-based conditional normalization, adapting the feature maps according to the semantic masks. This ensures semantic consistency in the generated images.

In the decoder, ResBlock modules with SPADE are used similarly to the encoder, with convolutional layers performing up-sampling. The decoder architecture is nearly symmetrical to the encoder, but it incorporates SPADE and up-sampling layers to align the generated output with the desired semantic structure. This semantic conditioning in the decoder ensures that the output image is consistent with the intended semantic content, guiding the synthesis process to produce meaningful results. Finally, GroupNorm2D, SiLU activation, and Conv2D are applied to generate an output tensor of the same dimension as the input, representing the mean noise tensor for a given time step.

In summary, RoadPainter improves semantic road crack synthesis by integrating SPADE for semantic consistency, CFG for refined conditional control, and a novel U-Net architecture with ResBlocks and self-attention mechanisms. This approach, validated through extensive experiments detailed in the following sections, demonstrates enhanced alignment between generated images and semantic masks, resulting in high-quality and accurate outputs.

3. Experimental setup

3.1. *Benchmark crack datasets*

Binary crack segmentation datasets, including DeepCrack [56], CrackSC [57], CFD [58], and Crack500 [29], have been utilized to validate various generative architectures and segmentation models. Additionally, the benchmark dataset Mosquitonet [59] was employed to validate the refined and detailed pavement condition index.

The DeepCrack dataset contains 537 raw crack images (300 for training and 237 for testing) with human-based segmentation annotations. The percentage of crack versus non-crack pixels is 2.91 %/97.09 %

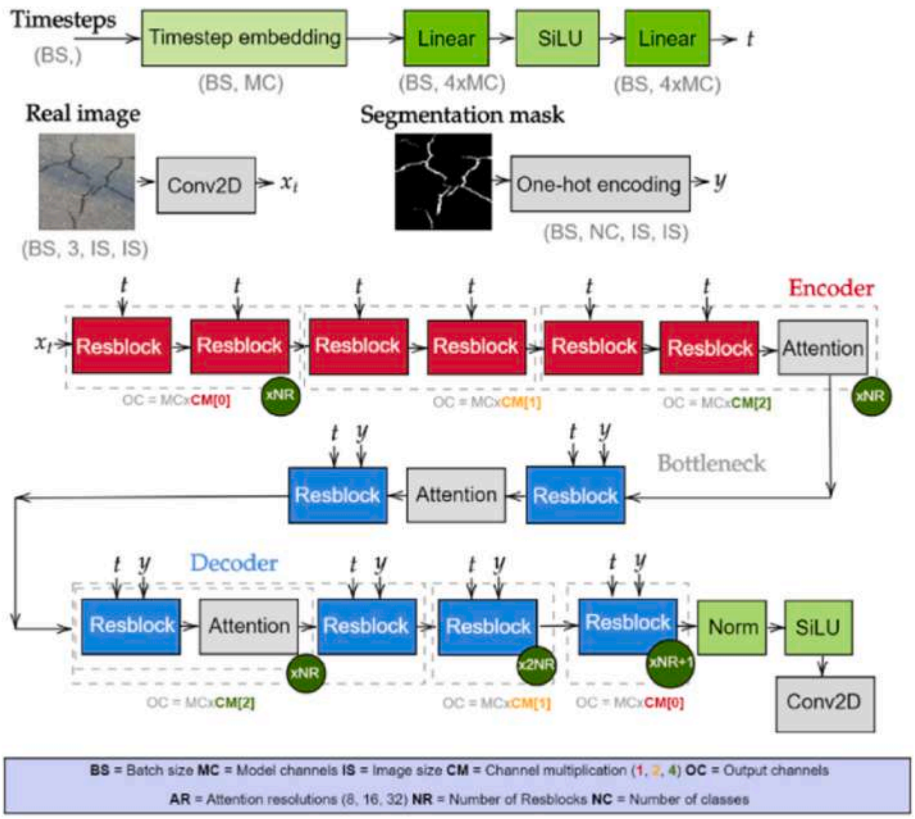

Fig. 1. Proposed denoising network of RoadPainter.

for the training set and 4.33 %/95.67 % for the test set, with images sized at 544x384 pixels. This benchmark database includes various textures, scenes, and scales. The Crack Forest Dataset (CFD) includes 118 annotated road crack images, each 480x320 pixels, captured using a smartphone. These images feature disturbances such as shadows, oil spots, and water stains, with a train-test split of 90%–10 %.

The Crack500 dataset consists of 500 images, each 2000x1500 pixels, captured with cost-effective smartphones on the main campus of Temple University. These images are typically cropped into 640x360 pixel patches, reflecting diverse lighting conditions and shadows that increase segmentation complexity. The dataset was used with 1896 patches for training and 1124 for testing. The CrackSC dataset, in contrast to CFD and Crack500, is specifically designed to address challenges in low-volume local roads with pronounced shadows and dense crack formations. It includes additional complexities such as tree shadows, fallen leaves, and abundant moss. This dataset comprises 197 images of pavement surface cracks, captured using an iPhone 8 along Enoree Ave, Columbia, SC, with a 9:1 train-test split.

The Mosquitonet dataset contains 7099 images, each with dimensions of 640x640 pixels, accompanied by annotations in multiple object detection formats. These top-down view images were collected using a vehicle-mounted camera and include 13 types of distress categorized into three families: distress (e.g., potholes or longitudinal cracks), repair (e.g., patches or sealed cracks), and sewer (e.g., manholes). The dataset encompasses a variety of brightness and weather conditions. To validate the refined pavement condition index, a subset of 100 patches from Mosquitonet was annotated for binary crack segmentation.

3.2. Performance metrics

Image quality assessment (IQA). To ensure the reliability and realism of the synthetic frames generated by generative architectures, it is crucial to employ IQA metrics. These metrics provide an objective evaluation of visual fidelity and structural integrity, allowing for a comprehensive assessment of image quality against real-world standards. The IQA metrics analyzed in this study include Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Fréchet Inception Distance (FID), and Learned Perceptual Image Patch Similarity (LPIPS).

PSNR measures the quality of reconstruction by comparing the peak signal value (255) to the mean squared error (MSE) between the real image (x_{real}) and the generated image (x_{fake}) (Eq. (2)). It is typically expressed in decibels (dB), with higher values indicating better quality. SSIM, a perception-based metric, assesses image quality based on luminance, contrast, and structural similarity (Eq. (3)). SSIM values range from 0 (completely dissimilar) to 1 (identical).

$$PSNR = \frac{1}{MxN} \sum_{i=1}^{M} \sum\nolimits_{j=1}^{N} 10 \, log_{10} \left(\frac{PV^2}{MSE(x_{real}, x_{fake})} \right) \tag{2}$$

$$SSIM = \frac{\left(2\mu_{x_{real}}\mu_{x_{fake}} + C_1\right)\left(2\sigma_{x_{real}}\sigma_{x_{fake}} + C_2\right)}{\left(\mu^2_{x_{real}} + \mu^2_{x_{fakel}} + C_1\right)\left(\sigma^2_{x_{real}} + \sigma^2_{x_{fakel}} + C_2\right)} \tag{3}$$

FID measures the similarity between the distributions of real images (x_{real}) and generated images (x_{fake}) by comparing their associated features extracted using a pre-trained Inceptionv3 model (Eq. (4)). A lower FID score indicates a closer statistical resemblance between the two distributions. LPIPS assesses similarity based on human perception. It calculates the Euclidean distance between the feature representations of real and generated images obtained from a pre-trained AlexNet model. A lower LPIPS score signifies greater visual similarity between x_{real} and x_{fake} according to human perception.

$$FID = \left\|\mu_{real} - \mu_{fake}\right\|^2 + Tr\left(\sigma_{real} + \sigma_{fake} - 2\sqrt{\sigma_{real}\sigma_{fake}}\right) \tag{4}$$

In the preceding equations, μ_{real} and μ_{fake} represent the mean features, while σ_{real} and σ_{fake} denote the variance of the extracted features from Inceptionv3. Tr refers to the trace. Additionally, $\mu_{x_{real}}$ and $\mu_{x_{fake}}$ correspond to the mean of the real and generated images, respectively, and $\sigma_{x_{real}}$ and $\sigma_{x_{fake}}$ denote their variances. C_1 and C_2 are constants.

Segmentation metrics. In this article, a binary segmentation problem is addressed, where pixel values of the image background are set to 0 and crack pixels to 1. Accuracy (Eq. (5)) is calculated as the average mean absolute error between predicted and ground truth pixels. However, accuracy alone is insufficient in cases of high pixel imbalance, so additional metrics are included. A True Positive (TP) occurs when both the predicted pixel value and the ground truth pixel value are 1; a False Positive (FP) occurs when the real label is 0 and the predicted value is 1; a False Negative (FN) occurs when the prediction is 0 and the real label is 1; and a True Negative (TN) arises when both the label and the prediction are 0. Based on these parameters, the following metrics are defined: Precision (Eq. (6)) measures the quality of predictions, Recall (Eq. (7)) reflects how many of the true labels have been correctly identified, and the F1-score (Eq. (8)) provides a balance between Precision and Recall. Finally, Intersection over Union (IoU, Eq. (9)) quantifies the overlap between the set of predicted pixel values (P) and the set of ground truth pixel values (GT) as the ratio of their intersection to their union. The mean IoU (mIoU) is the average of the IoU calculated across all categorized classes.

$$Accuracy = \frac{1}{N} \sum_{i=1}^{N} \left(x^i_{pred} - x^i_{gt} \right) \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 - score = \frac{2TP}{2TP + FN + FP} \tag{8}$$

$$IoU = \frac{P \cap GT}{P \cup GT} = \frac{TP}{TP + FP + FN} \tag{9}$$

In the previous expressions, x^i_{pred}, x^i_{gt}, and N refer to the predicted, ground truth pixel-level label, and the number of pixels, respectively.

3.3. Environment and programming details

The programming language used is Python 3.8.10. The computer vision libraries include OpenCV 4.7.0.72 and Pillow 9.5.0. The deep learning framework is PyTorch 2.2.1 with CUDA 12.1. Figures in the paper are created using Seaborn 0.12.2, Matplotlib 3.7.1, and Matcha Online Math Editor. The computer station utilized is a Dell Alienware Aurora with a GeForce RTX 3080 Ti GPU.

3.4. Workflow

The workflow of this study is illustrated in Fig. 2. Initially, the proposed architecture, RoadPainter, was optimized using the DeepCrack dataset. The model was trained for 300 epochs with a batch size of 4 and a dropout ratio of 0.1. The Mean Squared Error (MSE) was used as the loss function, and the Adam optimizer was employed with a learning rate of 1e-6, β_1 of 0.5, and β_2 of 0.999, on images sized 128x128 pixels.

The optimization process was conducted in sequential steps. Evaluation metrics included IQA metrics, visual inspection of synthetic images, and computational cost. Initial efforts focused on identifying the optimal architecture by varying channel multipliers and model channels. For the optimal configuration, various noise schedules -linear, cosine, sigmoid, and stable diffusion-were explored [60]. Subsequently,

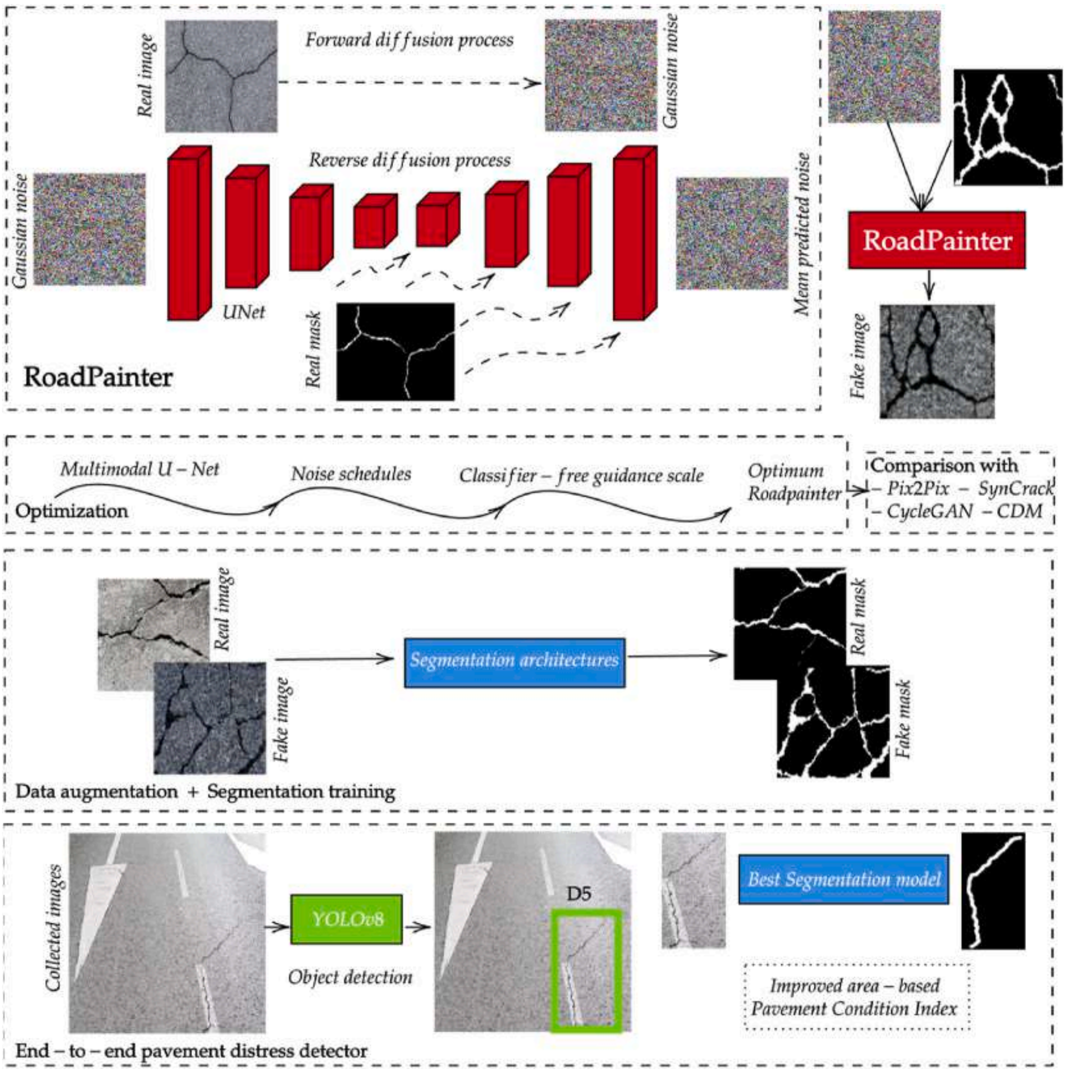

Fig. 2. Workflow.

the impact of the improved CFG scale was assessed. All results presented are based on the test split.

The performance of the optimal RoadPainter model was benchmarked against state-of-the-art models using the DeepCrack, CFD, CrackSC, and Crack500 datasets. Models such as CycleGAN, Pix2Pix, SynCrack, and CDM were evaluated using default hyperparameters from their respective original publications. Following validation of the RoadPainter model's superior performance through IQA metrics and visual inspection, it was employed to generate synthetic crack images from manually crafted semantic masks. This process involved using the model to enhance binary information images with texture.

After image generation, data augmentation was performed on synthetic images across open-source datasets. This augmentation process included both synthetic RGB images and their corresponding semantic masks in the training set, while test images remained unaltered. Five segmentation architectures -U-Net [50], PAN [61], PSPNet [62], LinkNet [63], and FPN [64]-were trained both before and after augmentation. These models were trained for 300 epochs using the Adam optimizer with a learning rate of 1e-4, β_1 of 0.5, and β_2 of 0.999. The Dice loss function was utilized, with a batch size of 32 and an image size of 128 pixels.

Finally, the most robust segmentation architecture, trained with the augmented Crack500 dataset, was selected to validate the proposed methodology for computing the refined pavement condition index. A deep learning-based architecture was utilized with the Mosquitonet dataset to detect various road defects. The chosen object detector is YOLOv8 [65]. YOLOv8 was trained with a learning rate of 1e-5, using the Adam optimizer with β_1 of 0.5 and β_2 of 0.999, a batch size of 32, for 300 epochs. The loss function employed is a weighted loss that integrates Complete-IoU for bounding box regression and Focal Losses for completeness and classification.

Bounding boxes predicted by YOLOv8 were cropped and resized, with only those classified as crack types retained. The crack categories, as defined in Ref. [59], include seven distinct types. The best pre-trained segmentation model was then used to generate binary segmentation masks from the predicted cropped bounding boxes. These masks were employed to quantify the potential refinement of the pavement condition index proposed in Ref. [59], both before and after applying semantic segmentation.

4. Results and discussion

4.1. RoadPainter optimization

4.1.1. Architecture

First, the impact of channel multiplier expansion strategies (CM, as shown in Fig. 1) within the proposed multimodal attention-guided U-Net architecture for semantic image synthesis is examined. In this initial set of experiments, all network-related hyperparameters were kept fixed to assess the effect of increasing the number of ResBlock modules in both the encoder and decoder networks. Specifically, the focus was on lengthening the channel multiplier vector and modifying its elements to boost the output channels of the corresponding modules. This analysis aims to evaluate the trade-off between computational efficiency and channel multiplier expansion. The results are presented in Table 1.

Table 1 demonstrates a clear trend of increasing computational complexity associated with more elaborate channel multiplication configurations. This trend is evident in the rising number of parameters, training and sampling times, and overall model capacity. Although it is generally anticipated that more complex models will yield improved image quality metrics, this assumption is not always validated. The subsequent discussion on IQA metrics will address this hypothesis.

For every configuration, PSNR values are higher, indicating greater similarity between generated and ground truth samples in terms of pixel intensity. However, PSNR alone may not accurately reflect the quality of segmentation masks, as it focuses solely on pixel intensity and overlooks higher-level features such as structure, shape, or semantic meaning. Therefore, high PSNR values may coexist with artifacts or inconsistencies that are not immediately visible but can compromise overall segmentation quality. Conversely, SSIM assesses structural similarity by considering luminance, contrast, and local pixel intensity variations. Results show substantial similarity, though the most complex configuration exhibits approximately 5.1 % lower SSIM performance compared to simpler configurations. Despite SSIM's advantages over PSNR, it may not fully address issues related to semantic information control or guidance, as high SSIM does not guarantee accurate semantic segmentation.

FID measures the perceptual similarity between generated and real samples. A notable trend is the deterioration of FID with increased model complexity. The configurations (1, 2, 3, 4) and (1, 1, 2, 3, 4), which are the least complex, show exceptional FID performance, nearing the minimum value. The most complex configuration results in the lowest LPIPS value, but differences between configurations are marginal. Given that IQA metrics favour the least complex models, the (1, 2, 3, 4) configuration was chosen for its optimal FID performance and lower computational cost. Attempts to increase complexity with additional self-attention modules were abandoned due to memory allocation errors. Table 2 displays the results of adjusting the model's channel configurations.

Contrary to expectations, where increased complexity typically results in diminished FID, the results revealed a more nuanced relationship. Notably, the configuration with 128 channels yielded the lowest FID (Table 1), while simpler configurations, such as those with 32 channels, demonstrated superior FID values (↑127.4 %). Conversely, both lighter (64 and 96 channels) and more complex (160 channels) configurations were close to the optimal configuration (128 channels), but they did not surpass it in terms of FID. Beyond the quantitative IQA metrics, a qualitative visual assessment of the generated images was conducted to evaluate their fidelity, reliability, and segmentation mask accuracy. Fig. 3 depicts the various images generated for each configuration.

The RoadPainter configurations highlight two key aspects: realism and texture. The simplest model can create cracks but often produces unrealistic colors, such as green and yellow, which are not typical of road surfaces. Consequently, the model with 32 channels, which yields poor FID scores, is discarded due to its unrealistic images. As the number of channels increases, texture quality improves. However, the most complex model often results in nearly opaque images or struggles with pavement textures, particularly in defects like block or alligator cracking. Models with 96 and 128 channels provide more convincing textures, while the 64-channel model lacks realism. After evaluating multiple images, the 128-channel model was determined to be the best, offering logical colors, high realism, and convincing texture. Although not the lightest model, the (1, 2, 3, 4) configuration with 128 channels is deemed optimal.

4.1.2. Noise schedules

Fig. 4 illustrates the square root of the cumulative product of α_t (see Ref. [32]) over timesteps, indicating the evolution of the forward diffusion process from the original raw image to pure Gaussian noise at $T = 1000$. The corresponding signal-to-noise ratio (SNR) is depicted on the right plot, balancing noise, and image power throughout training. Typically, noise schedules begin with non-zero SNR, causing a disparity between training and inference. When $t = T$ during training, a small amount of signal persists, comprising low-frequency information. In contrast, noise schedules with zero terminal SNR closely resemble inference behavior, aligning with pure noise input at $t = T$ during training. Fig. 4 displays the logarithmic representation of different noise schedules, demonstrating adherence to this requirement while showing distinct modulation of the forward diffusion process for each schedule.

Table 3 presents the IQA metrics for various noise schedules. The results demonstrate relatively consistent trends across all noise schedules, with RoadPainter's optimal configuration slightly outperforming others in the stable diffusion noise schedule, particularly in terms of PSNR. However, the most significant impact is observed in the FID values, where the sigmoid schedule performs the worst. The nearly linear schedule also surpasses the cosine schedule. Notably, the stable diffusion schedule achieves an FID of 1.11, which is nearly six times lower than the best result obtained, highlighting a substantial disparity. This result underscores the importance of incorporating noise schedules as an additional degree of freedom in the optimization space and tuning hyperparameters for semantic image synthesis models, especially in crack synthesis. Consequently, the stable diffusion noise schedule has been selected.

Fig. 5 presents a selection of synthetic images generated from ground truth masks. These images showcase various geometries (e.g., mesh, diagonal, transversal, longitudinal) and exhibit a range of thicknesses, lengths, perspectives, and patterns. Notably, the texture of the pavement becomes discernible upon closer inspection, highlighting the model's ability to generate cracks with multiple logical endpoints, leading to diverse backgrounds. This variability significantly enhances the value of synthetic images for data augmentation by providing both considerable diversity and high realism. Unlike traditional data augmentation techniques, which often rely on geometric alterations, this approach leverages a pre-trained generative model to produce an infinite number of frames that meet two criteria: they closely resemble real images and are conditioned by a semantic mask. This capability is particularly valuable in contexts like road maintenance, where the collection and labeling of images are major challenges. In essence, this method offers an efficient solution for data collection and labeling.

Table 1
Computational aspects and image quality assessment metrics (calculated for the test partition) using the DeepCrack dataset, focusing on channel multiplications.

Channel multiplication	Parameters (M)	Training time/cost (h/MiB)	Inference time/cost (h/MiB)	PSNR	SSIM	FID	LPIPS
(1, 2, 3, 4)	134	4.00/5940	0.227/4774	56.58 ± 0.72	0.98 ± 0.03	6.23 ± 0.51	0.0003 ± 0.0001
(1, 1, 2, 3, 4)	143	4.60/5574	0.272/5228	56.36 ± 0.74	0.98 ± 0.01	6.63 ± 0.44	0.0003 ± 0.0001
(1, 1, 2, 2, 4, 4)	187	5.40/6794	0.296/6234	56.62 ± 0.69	0.98 ± 0.02	7.69 ± 0.53	0.0004 ± 0.0001
(0.5, 1, 1, 2, 2, 4, 4)	197	5.47/6820	0.316/6384	56.19 ± 0.70	0.93 ± 0.03	9.82 ± 0.47	0.0002 ± 0.0001

Table 2
Computation aspects and image quality assessment metrics (computed for test partition) with DeepCrack datasets in terms of model channels.

Model channels	Parameters (M)	Training time/cost (h/MiB)	Inference time/cost (h/MiB)	PSNR	SSIM	FID	LPIPS
32	15	2.00/1550	0.070/1276	55.66 ± 0.26	0.96 ± 0.01	14.16 ± 0.56	0.0005 ± 0.0001
64	42	2.63/2840	0.147/2858	56.57 ± 0.80	0.98 ± 0.01	7.01 ± 0.50	0.0004 ± 0.0001
96	82	4.00/4408	0.227/4784	56.39 ± 0.56	0.98 ± 0.02	8.85 ± 0.43	0.0002 ± 0.0001
128	134	4.63/5940	0.255/4774	56.58 ± 0.72	0.98 ± 0.03	6.23 ± 0.51	0.0003 ± 0.0001
160	197	7.72/8142	0.462/8704	54.44 ± 0.69	0.96 ± 0.03	8.08 ± 0.58	0.0004 ± 0.0001

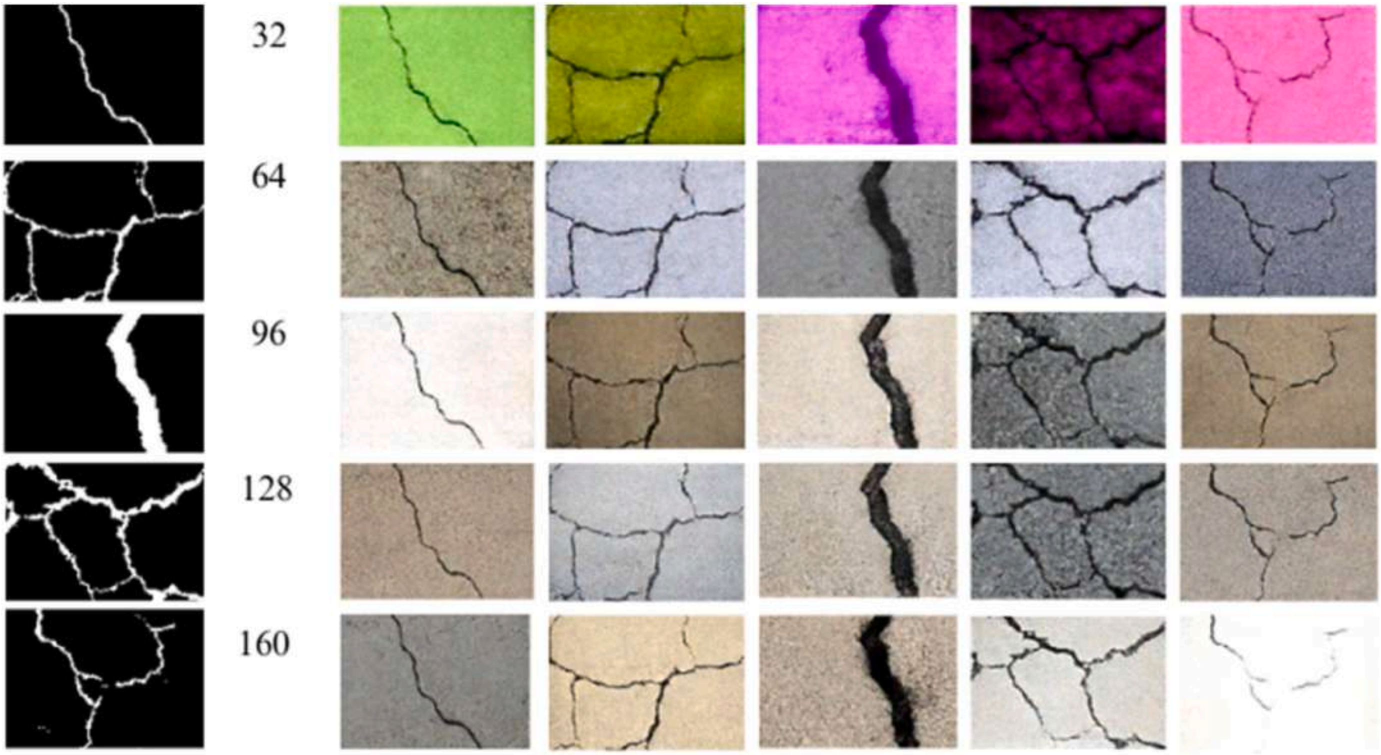

Fig. 3. Representation of synthetic cracks with varying features (e.g., widths, geometries) sampled from diffusion models trained with different model channels (MC).

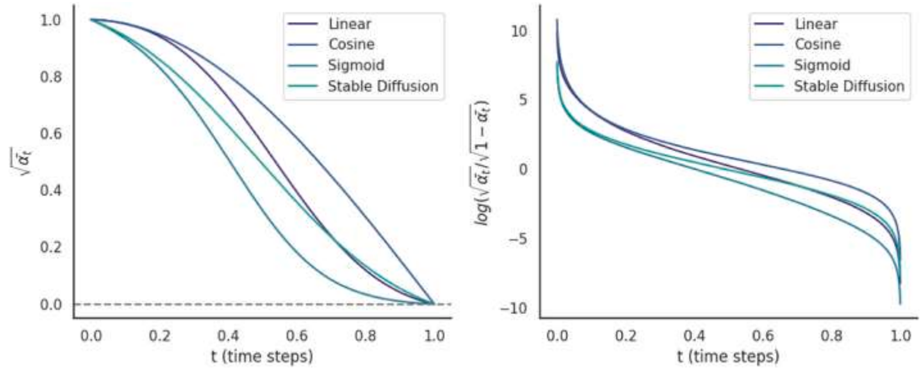

Fig. 4. Modified noise schedules with rescaling to fit pure noise at t = T.

4.1.3. *CFG scale*

In Table 4, it can be observed that the differences in relation to all metrics are very minor. However, for the case of s = 3, an FID is obtained that is 5.41 % lower compared to the best configuration from the previous sub-analysis, along with a superior PSNR of 2.19 %. Therefore, despite the slight differences, this hyperparameter has been established to conclude the overall optimal configuration of RoadPainter. In conclusion, the optimal configuration has a channel multiplier vector of (1, 2, 3, 4) with 128 channels, employing a stable diffusion noise schedule, and a CFG scale of 3. Subsequently, to ascertain the robustness and efficacy of the proposed generative model, a comparative analysis has been conducted against state-of-the-art generative models across various open-source datasets pertaining to pavement cracks.

Table 3
IQA metrics for different noise schedules.

		PSNR	SSIM	FID	LPIPS
Schedule	Linear	56.58 ± 0.72	0.98 ± 0.03	6.23 ± 0.51	0.0003 ± 0.0001
	Sigmoid	55.59 ± 0.11	0.97 ± 0.06	10.02 ± 0.49	0.0003 ± 0.0001
	Cosine	55.81 ± 0.57	0.97 ± 0.01	9.65 ± 0.57	0.0003 ± 0.0001
	Stable Diffusion	58.47 ± 0.89	0.99 ± 0.01	1.11 ± 0.44	0.0002 ± 0.0001

4.2. Comparison with state-of-the-art approaches

Table 5 presents IQA metrics for various state-of-the-art models evaluated on different open-source datasets. RoadPainter consistently outperforms the other models, demonstrating significantly higher PSNR and SSIM scores, as well as substantially lower FID and LPIPS values. In contrast, Pix2Pix and CycleGAN exhibit lower performance across all datasets, with reduced PSNR and SSIM scores and increased FID and LPIPS values. SynCrack and CDM display intermediate performance, with results varying across datasets. Overall, RoadPainter delivers superior image quality and fidelity compared to the other models.

For the CFD dataset, which comprises a relatively small volume of images, the results are more comparable. Although RoadPainter shows slightly better performance according to IQA metrics, it often struggles with semantic correlation, as illustrated in Fig. 6. In comparison, Pix2Pix and CycleGAN achieve conditional translation but face their own limitations. Pix2Pix tends to suffer from mode collapse in the background, resulting in a lack of diversity in synthetic images and potentially degrading the performance of segmentation architectures trained on these images. Meanwhile, CycleGAN avoids mode collapse but often produces backgrounds that are highly unrealistic.

For the CrackSC dataset, Pix2Pix does not exhibit mode collapse in the background (see Fig. 6). However, like CycleGAN, it demonstrates limited generalization with new samples, resulting in outputs that resemble pavement backgrounds with slightly more texture than those from earlier datasets. RoadPainter, on the other hand, excels by generating a diverse range of textures, including varying colors, tones, and intricate details such as shadows or dust within the cracks.

The Crack500 dataset, which includes a large volume of high-quality, high-resolution images with diverse and textured content, presents a different scenario. Here, Pix2Pix produces both realistic backgrounds and cracks with clear semantic correlation. Despite this, mode collapse is evident, as the background remains largely similar regardless of the input. CycleGAN generates highly unrealistic images, occasionally achieving semantic similarity but failing to capture the characteristic darker color of cracks relative to the background. In contrast,

Table 4
IQA metrics as a function of the CFG scale.

		PSNR	SSIM	FID	LPIPS
CFG scale	1.5	58.47 ± 0.89	0.99 ± 0.01	1.11 ± 0.44	0.0002 ± 0.0001
	2.0	57.49 ± 0.43	0.98 ± 0.02	1.10 ± 0.43	0.0001 ± 0.0000
	2.5	58.52 ± 0.92	0.99 ± 0.02	1.09 ± 0.47	0.0002 ± 0.0001
	3.0	59.75 ± 0.96	0.99 ± 0.02	1.05 ± 0.33	0.0002 ± 0.0001

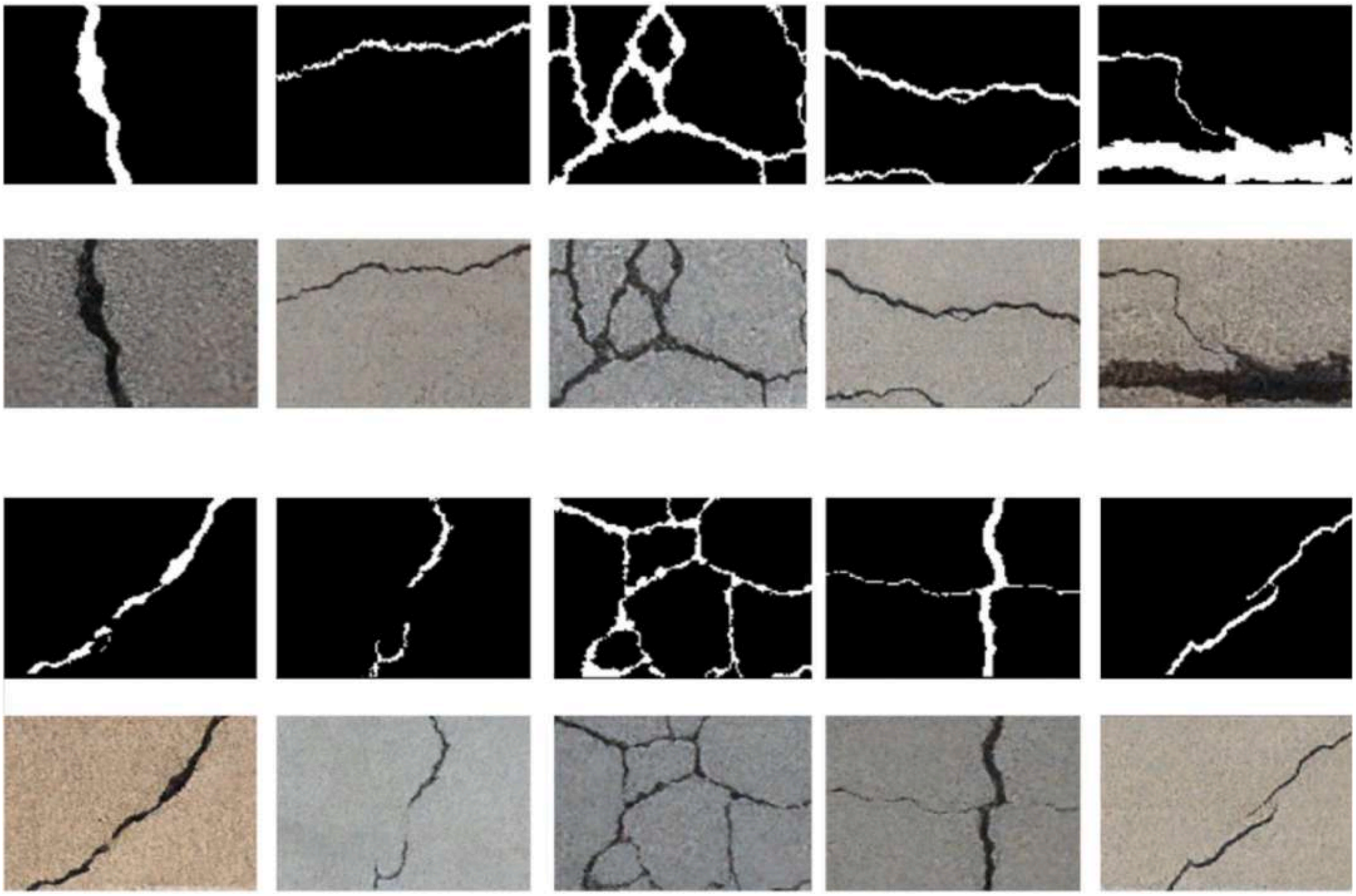

Fig. 5. Real test segmentation masks from DeepCrack and synthetic crack images generated by RoadPainter.

Table 5
IQA metrics of synthetic images for different generative models and benchmark datasets.

Dataset	Model	PSNR	SSIM	FID	LPIPS
DeepCrack	Pix2Pix	15.03 ± 0.36	0.23 ± 0.01	137.13	0.57 ± 0.09
	CycleGAN	14.85 ± 0.49	0.27 ± 0.03	158.16	0.64 ± 0.02
	RoadPainter	59.75 ± 0.92	0.99 ± 0.02	1.05 ± 0.33	0.0002 ± 0.0001
	SynCrack	14.73 ± 1.65	0.27 ± 0.01	115.35 ± 0.69	0.58 ± 0.02
	CDM	12.87 ± 0.46	0.28 ± 0.08	136.49 ± 0.98	0.61 ± 0.08
CrackSC	Pix2Pix	19.76 ± 0.63	0.19 ± 0.01	88.79	0.42 ± 0.08
	CycleGAN	18.51 ± 0.20	0.51 ± 0.04	84.69	0.51 ± 0.03
	RoadPainter	67.60 ± 0.97	0.99 ± 0.01	2.10 ± 0.01	0.0003 ± 0.0001
	SynCrack	18.48 ± 0.95	0.24 ± 0.01	136.82±	0.40 ± 0.06
	CDM	16.61 ± 0.05	0.17 ± 0.04	50.31 ± 0.63	0.78 ± 0.10
Crack500	Pix2Pix	15.29 ± 0.28	0.061 ± 0.003	73.90 ± 8.96	0.44 ± 0.02
	CycleGAN	13.60 ± 0.09	0.08 ± 0.01	89.70 ± 0.03	0.82 ± 0.08
	RoadPainter	70.05 ± 2.50	0.99 ± 0.01	1.55 ± 0.01	0.0004 ± 0.0001
	SynCrack	16.42 ± 0.90	0.11 ± 0.01	187.27 ± 6.70	0.53 ± 0.03
	CDM	15.52 ± 0.87	0.12 ± 0.03	62.98 ± 0.11	0.82 ± 0.02
CFD	Pix2Pix	23.04 ± 0.19	0.63 ± 0.08	163.11 ± 26.00	0.34 ± 0.03
	CycleGAN	22.76 ± 0.13	0.67 ± 0.09	155.35 ± 32.15	0.50 ± 0.05
	RoadPainter	44.30 ± 1.75	0.68 ± 0.01	98.11 ± 19.33	0.27 ± 0.01
	SynCrack	17.38 ± 3.98	0.41 ± 0.09	140.87 ± 2.14	0.41 ± 0.02
	CDM	13.95 ± 1.60	0.40 ± 0.08	134.07 ± 3.60	0.64 ± 0.03

RoadPainter consistently delivers highly realistic synthetic images with notable diversity, featuring distinct backgrounds and avoiding issues such as mode collapse and limited generalization.

Fig. 7 displays synthetic images generated by state-of-the-art models specifically designed for the semantic synthesis of road cracks. CDM successfully captures the correlation with the semantic mask; however, its backgrounds often appear highly unrealistic, displaying whitish tones around the defects and a consistently grayish background. Additionally, CDM requires both a semantic mask and a real background image without defects for generating new images, which may limit its applicability in real-world engineering contexts. In contrast, while SynCrack produces images that may seem realistic, it fails to accurately represent the true tones of the pavement and lacks the depiction of artifacts such as shadows or fine details. Based on IQA metrics and visual inspection, it is evident that the RoadPainter model outperforms these state-of-the-art models.

Fig. 8 presents additional images generated by RoadPainter across various datasets, showcasing its strengths and limitations. The size and resolution of datasets influence the model's creative capabilities. RoadPainter excels at producing highly textured and realistic images, even in cases where cracks are not distinctly differentiated from the background. Future research will aim to enhance the model's performance on smaller datasets. Notably, images from Crack500 demonstrate exceptional texture and realism, closely resembling the original images. Results from DeepCrack are omitted, as they were previously discussed in Fig. 5.

4.3. Improved segmentation efficiency with augmented datasets

Each model was augmented with synthetic images generated by its corresponding pre-trained generative model on its training split. Tables 6–9 present the segmentation metrics for various segmentation architectures and benchmark datasets. For each dataset, with N images in the training partition, N/2 synthetic images were added, resulting in 3N/2 images in the new training partitions, where 66.7 % are real and the remaining 33.3 % are synthetic.

To generate these synthetic images, we created a dataset using the Paint tool, featuring a consistent black background with white pixels representing cracks, thereby producing synthetic crack semantic masks. We introduced diversity by incorporating various crack geometries, widths, and positions, including longitudinal, transversal, irregular, block, and alligator cracking (see Fig. 9). This pseudo-labeling method proved highly efficient, allowing us to control the generated images and enhance the learning of data-driven models by increasing the volume of diverse images. Unlike manual dataset annotation, this approach streamlines the process and ensures greater variety in the generated images.

Table 6 shows that while all architectures achieved high baseline accuracy (above 0.97) even without augmentation, RoadPainter consistently enhanced this metric across all models. However, accuracy, which is simply the ratio of correctly classified pixels, can be misleading in imbalanced datasets where cracks are significantly outnumbered by background pixels. This issue is evident in DeepCrack and several other datasets. For images depicting mesh-type defects, such as block cracking or alligator cracking, the proportion of crack-associated pixels is relatively higher compared to the total number of pixels. In contrast, most images contain isolated cracks, such as longitudinal or transverse cracks, where the crack pixel ratio fluctuates between 5 % and 15 %. This imbalance suggests that accuracy alone may not be the most appropriate metric for these datasets.

Precision, which measures the proportion of true positives among all predicted positives, showed the most significant improvements with RoadPainter augmentation. Architectures like LinkNet and U-Net experienced substantial gains in precision (↑49.7 % and ↑53.2 %, respectively), indicating a notable reduction in false positive crack detections. While precision is crucial, effective crack detection also requires high recall, meaning the model must capture a large proportion of actual cracks. The observed improvements in recall, alongside precision gains, suggest that the overall accuracy and the ability to identify all true crack instances have been balanced.

The F1-score, which is the harmonic mean of precision and recall, achieved the highest gain with FPN (↑23.6 %), reflecting significant improvement in this trade-off. This metric assesses the average overlap between predicted and ground truth crack masks. Additionally, the improvements in mIoU with RoadPainter augmentation indicate better overall segmentation quality, capturing both the presence and precise location of cracks more accurately. LinkNet showed the most significant improvement in mIoU (↑43.2 %). Although mIoU provides a comprehensive view of segmentation quality, it may not capture specific crack characteristics. Overall, these results underscore that the enhanced LinkNet with synthetic images from RoadPainter achieves the best metrics in the case of DeepCrack.

Table 7 illustrates that while significant improvements are observed in various metrics, the change in accuracy with RoadPainter augmentation may appear modest. This subtle shift is likely attributable to the limitations of accuracy as a metric in imbalanced crack-pixel datasets. Notably, a substantial positive trend is evident in precision for all architectures following the incorporation of synthetic data. Precision is particularly critical in imbalanced scenarios, as it reflects the reduction in false positive crack detections. Architectures such as U-Net and LinkNet achieved the most significant gains in precision (approximately 3.5 times), underscoring the effectiveness of RoadPainter in reducing the misclassification of background pixels as cracks.

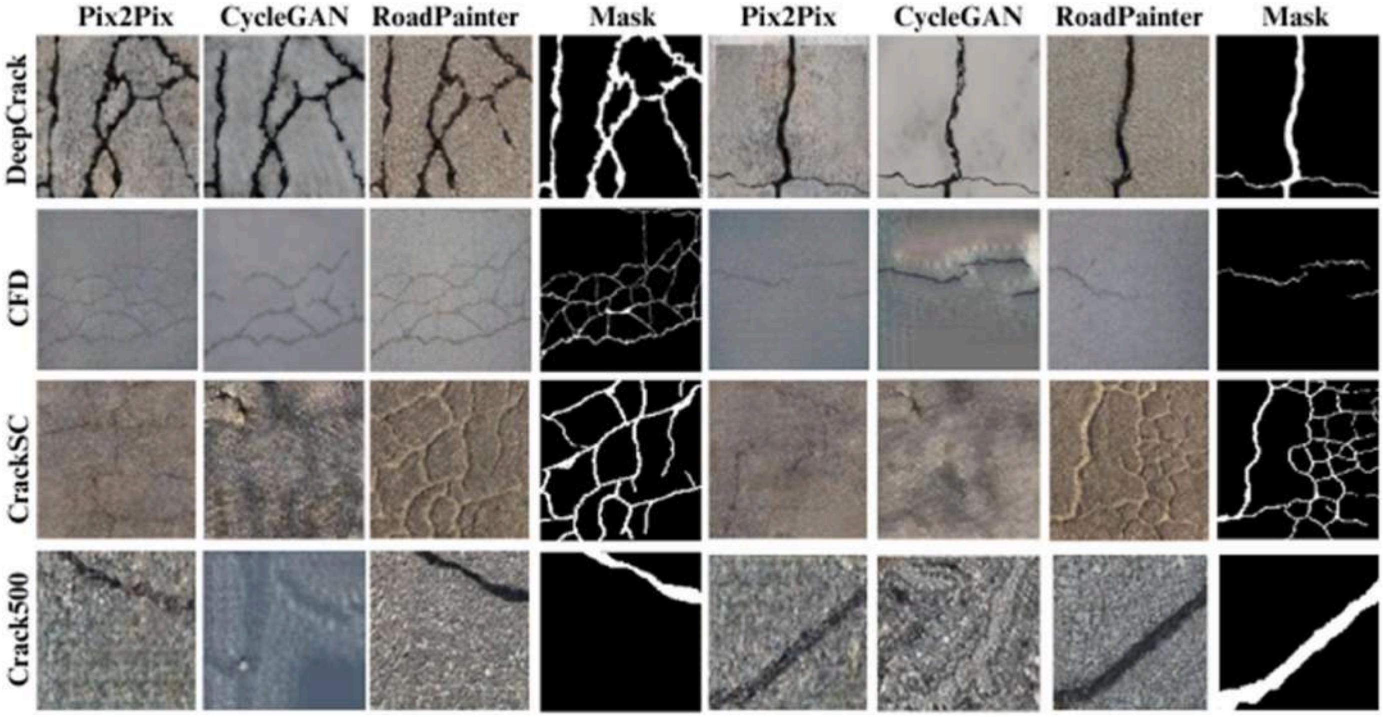

Fig. 6. Comparison of synthetic samples from test segmentation masks for several public crack datasets in terms of state-of-the-art and proposed generative architectures.

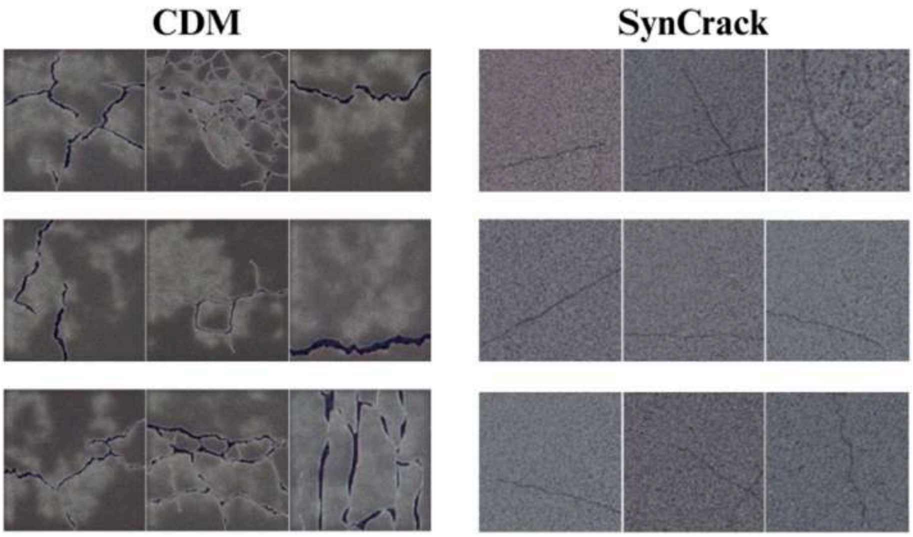

Fig. 7. Synthetic images from state-of-the-art crack-oriented semantic synthesis models.

In addition to reducing false positives, RoadPainter enhances the models' ability to capture a greater proportion of actual cracks. U-Net and LinkNet once again demonstrate the most notable improvements in recall, reflecting their enhanced capability to identify true cracks despite the class imbalance. All models also show substantial gains in the F1-score with RoadPainter augmentation. However, improvements in mIoU, which measures the overlap between predicted and ground truth crack masks, are relatively modest compared to other metrics. This is likely due to the CrackSC dataset's higher proportion of very thin or faint cracks, which are inherently more challenging to segment accurately.

Regarding Table 8, the improvements in recall and F1-score for the CFD dataset are more moderate compared to those observed with CrackSC. This suggests that the CFD dataset may have a higher proportion of easily detectable cracks or a more balanced class distribution. While RoadPainter continues to enhance crack detection performance, the initial performance may have been relatively high due to these dataset characteristics. In contrast to the CrackSC dataset, the improvements in mIoU are more pronounced in the CFD dataset. This indicates that the CFD dataset likely contains a higher prevalence of well-defined, thicker cracks, which benefit more from the additional training data provided by RoadPainter. Despite the varying degrees of improvement across different metrics, RoadPainter augmentation consistently boosted the performance of all architectures for the CFD dataset. This is likely because the dataset is quite small, with a 9:1 split applied, making performance improvements in the segmentation metrics less noticeable.

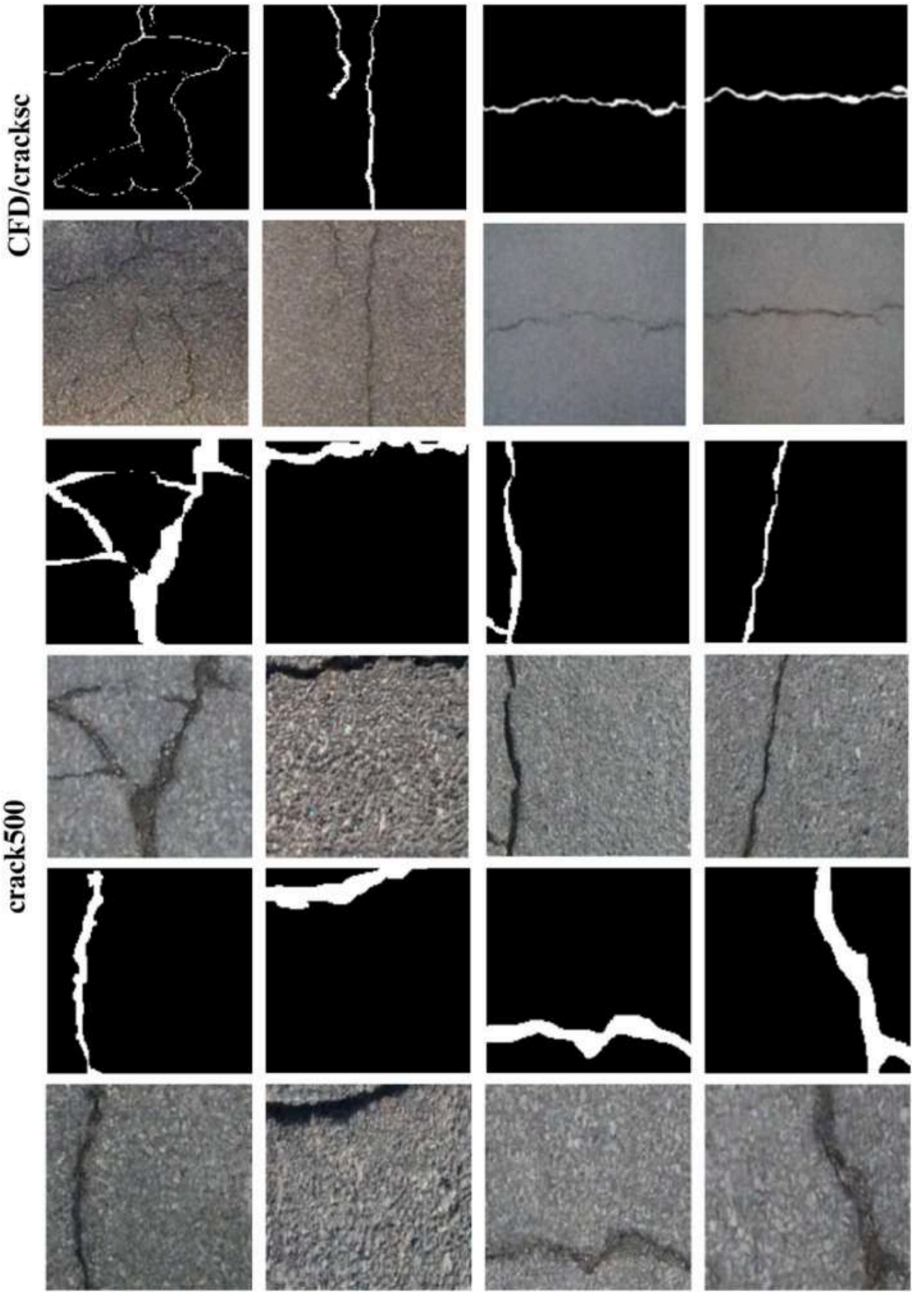

Fig. 8. Crack images generated by RoadPainter from test segmentation masks of various crack benchmark datasets.

Table 6
Standard binary segmentation metrics for several state-of-the-art architectures, before and after applying image augmentation with synthetic frames from the DeepCrack dataset, generated using the optimized RoadPainter model.

	Configuration	Accuracy	Precision	Recall	F1-score	mIoU
DeepCrack	FPN	0.974	0.447	0.870	0.594	0.423
	FPN&RoadPainter	0.981	0.697	0.889	0.734	0.580
	U-Net	0.976	0.464	0.883	0.603	0.432
	U-Net&RoadPainter	0.982	0.695	0.911	0.691	0.596
	LinkNet	0.975	0.451	0.850	0.595	0.424
	LinkNet&RoadPainter	0.983	0.691	0.926	0.755	0.607
	PSPNet	0.972	0.420	0.666	0.555	0.385
	PSPNet&RoadPainter	0.976	0.615	0.701	0.647	0.496
	PAN	0.974	0.449	0.787	0.593	0.422
	PAN&RoadPainter	0.981	0.683	0.854	0.718	0.560

Table 7
Standard binary segmentation metrics for several state-of-the-art segmentation architectures, before and after applying image augmentation with synthetic frames from the CrackSC dataset, generated using the optimized RoadPainter model.

Dataset	Configuration	Accuracy	Precision	Recall	F1-score	mIoU
CrackSC	FPN	0.975	0.122	0.294	0.202	0.113
	FPN&RoadPainter	0.980	0.305	0.589	0.312	0.185
	U-Net	0.981	0.132	0.307	0.212	0.118
	U-Net&RoadPainter	0.983	0.447	0.515	0.388	0.241
	LinkNet	0.979	0.119	0.352	0.198	0.110
	LinkNet&RoadPainter	0.980	0.418	0.547	0.397	0.248
	PSPNet	0.974	0.109	0.201	0.177	0.100
	PSPNet&RoadPainter	0.981	0.250	0.394	0.245	0.139
	PAN	0.975	0.116	0.283	0.192	0.110
	PAN&RoadPainter	0.980	0.319	0.562	0.330	0.198

Table 8
Standard binary segmentation metrics for several state-of-the-art segmentation architectures, before and after applying image augmentation with synthetic frames from the CFD dataset, generated using the optimized RoadPainter model.

Dataset	Configuration	Accuracy	Precision	Recall	F1-score	mIoU
CFD	FPN	0.968	0.100	0.817	0.190	0.104
	FPN&RoadPainter	0.978	0.202	0.822	0.323	0.192
	U-Net	0.979	0.170	0.785	0.330	0.160
	U-Net&RoadPainter	0.983	0.250	0.850	0.377	0.232
	LinkNet	0.981	0.190	0.813	0.351	0.191
	LinkNet&RoadPainter	0.982	0.225	0.822	0.380	0.213
	PSPNet	0.968	0.089	0.609	0.175	0.080
	PSPNet&RoadPainter	0.972	0.145	0.695	0.238	0.135
	PAN	0.972	0.109	0.728	0.215	0.110
	PAN&RoadPainter	0.979	0.204	0.793	0.323	0.192

Table 9
Standard binary segmentation metrics for several state-of-the-art segmentation architectures, before and after applying image augmentation with synthetic frames from the Crack500 dataset, generated using the optimized RoadPainter model.

Dataset	Configuration	Accuracy	Precision	Recall	F1-score	mIoU
Crack500	FPN	0.969	0.660	0.653	0.686	0.522
	FPN&RoadPainter	0.967	0.723	0.701	0.694	0.531
	U-Net	0.968	0.663	0.679	0.692	0.529
	U-Net&RoadPainter	0.970	0.723	0.704	0.709	0.550
	LinkNet	0.967	0.633	0.686	0.692	0.530
	LinkNet&RoadPainter	0.968	0.729	0.754	0.710	0.551
	PSPNet	0.962	0.605	0.640	0.649	0.480
	PSPNet&RoadPainter	0.964	0.659	0.676	0.670	0.499
	PAN	0.967	0.658	0.695	0.688	0.525
	PAN&RoadPainter	0.969	0.703	0.708	0.707	0.547

Table 9 presents the results for the Crack500 dataset, which is significantly larger than the previously used datasets. As with the previous tables, this table illustrates the potential benefits of image augmentation with synthetic cracks generated by the optimized RoadPainter model. Compared to the smaller CrackSC dataset, the improvements in all metrics for Crack500 are more modest. This is likely because Crack500 provides the models with a larger volume of real-world crack data during training. Despite this, RoadPainter augmentation still leads to noticeable improvements in precision, recall, and F1-score across all architectures. The results from the Crack500 dataset demonstrate the continued effectiveness of RoadPainter augmentation, even with a larger and potentially more diverse dataset.

Comparing the results across Tables 6–8 underscores the impact of dataset size on the effectiveness of synthetic data augmentation. While RoadPainter resulted in more substantial improvements with the smaller CrackSC dataset, its positive impact persists with the larger Crack500 dataset. This suggests that RoadPainter augmentation generalizes well across various data scenarios, validating the hypothesis that this innovative diffusion model enhances crack segmentation performance. The solution addresses challenges related to image collection and labeling, thereby supporting the data-driven nature of deep learning architectures for road crack segmentation.

This novel approach, in contrast to traditional image collection and labeling methods, offers several advantages. It enables the rapid creation of a large volume of diverse, high-quality images within a shorter timeframe and at a reduced cost. Additionally, it allows for precise control over the types of images generated, which is particularly valuable when a model struggles with specific crack types. In such scenarios, targeted images can be created by drawing specific semantic masks, thereby improving class balance and enhancing the performance of the segmentation architecture.

4.4. A refined pavement condition index

Most crack segmentation studies focus on designing high-performance architectures but often overlook aspects that are crucial for developing effective road maintenance strategies. Specifically, a few studies create pavement condition indices from segmentation results to guide strategic road maintenance decisions.

In a previous study [59], a pavement condition index was developed based on detections from an object detection model, referred to as the Area-based Pavement Distress Index (ASPDI). This index utilized a

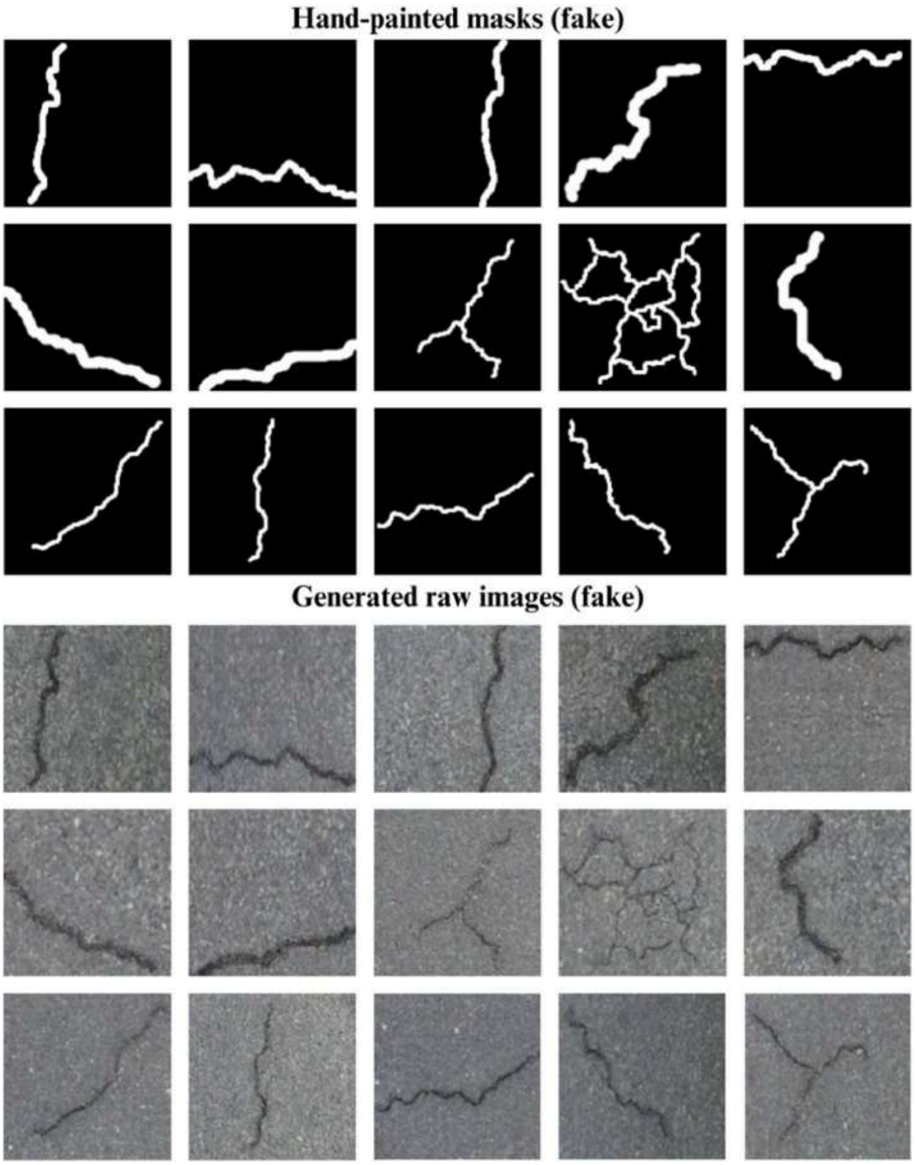

Fig. 9. Synthetic samples generated by the trained RoadPainter (using the Crack500 dataset) from hand-painted segmentation masks for data augmentation.

pre-trained object detection model that outputs multiple bounding boxes for each image, each encompassing various defects and classifying the type of distress. The ASPDI was then calculated based on these predictions, with values ranging from 0 to 100 %, where 100 % represents ideal road conditions and 0 % indicates a critical need for repair. The index is derived from the weighted sum of the areas within the predicted bounding boxes, with weights representing the severity of each defect as indicated by the predicted label. This approach incorporates fine-grained classification through severity coefficients, addressing the limitation of many state-of-the-art solutions.

However, analysis revealed a significant issue related to crack-type defects (e.g., diagonal cracking, irregular cracking) in calculating their area. This calculation often proved unrealistic, resulting in lower ASPDI scores for images with less severe defects. To address this issue, an object detection network, specifically YOLOv8, was utilized. From the YOLOv8 detections with the Mosquitonet test split, only those classified as crack-type defects were programmatically cropped. For these defects, the LinkNet & RoadPainter model, pre-trained with Crack500, was applied to generate predicted masks for the cropped patches. This refinement aimed to improve ASPDI calculations by using the area of crack-type pixels (white) within the bounding box dimensions rather than the entire bounding box area itself. This approach seeks to enhance the accuracy of the ASPDI calculation by providing a more precise measure of the crack areas.

To validate the effectiveness of the refined ASPDI calculation, 100 patches with crack-type defects from the Mosquitonet test split were selected and manually annotated for binary segmentation. Fig. 10 presents histograms of ASPDI values calculated under four conditions for this subset: (1) for the annotated or ground truth bounding boxes (object detection approach), (2) for the predicted bounding boxes using YOLOv8, (3) for the ground truth segmentation masks, and (4) for the predicted segmentation masks obtained from YOLOv8's cropped predictions processed with LinkNet. In the object detection approach, ASPDI values frequently fall below 80, indicating potential concerns. This is particularly evident because the test subset mainly consists of isolated crack-type defects (e.g., longitudinal, transversal, diagonal), resulting in a noticeable tail in the histogram where ASPDI values diverge from expected norms.

Conversely, the results derived from predicted segmentation masks show a significant improvement. Most instances are concentrated in a less concerning range, between 80 and 100. The few remaining instances correspond to two types of crack defects -block cracking and alligator

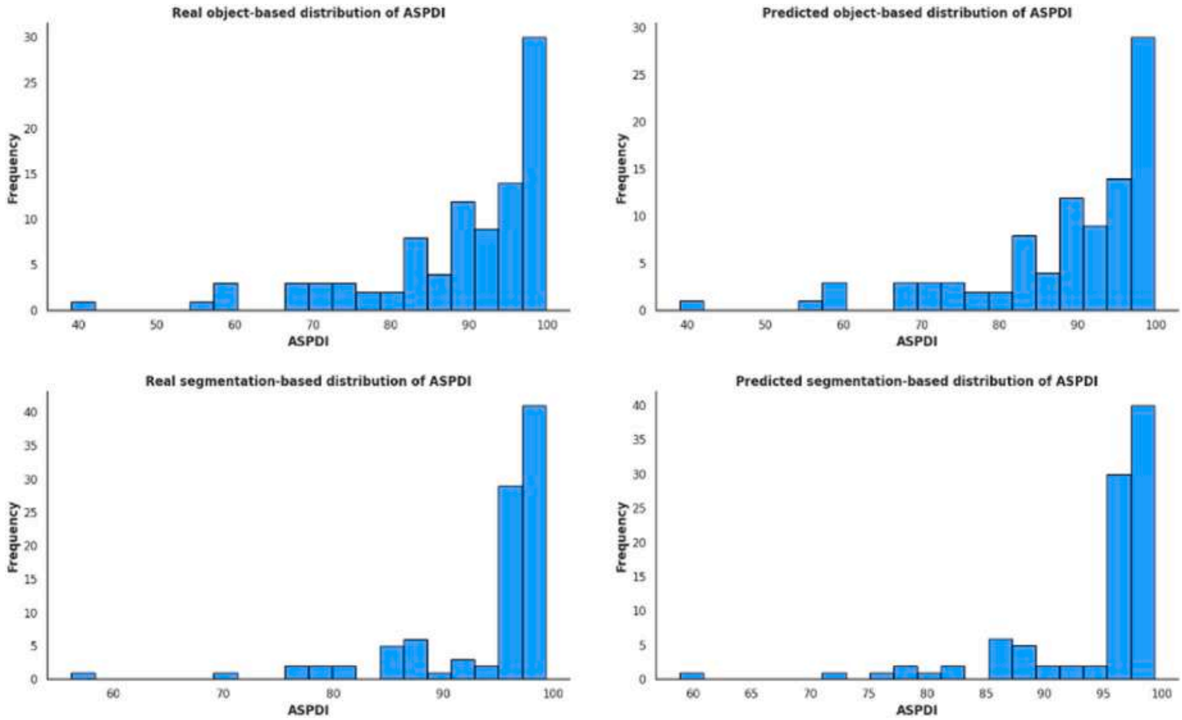

Fig. 10. ASPDI distribution for the test split of the Mosquitonet benchmark dataset, comparing real object detection/annotation ground truths with predictions from YOLOv8 (object detection) and augmented LinkNet (segmentation) for crack-type road defects.

cracking-where the area calculated from the predicted bounding boxes closely matches that from the segmentation masks. By refining the ASPDI index to focus on pixel-level analysis, we retain the detailed classification provided by the original ASPDI while achieving a more accurate measure of crack area. This refinement results in a more realistic and meaningful ASPDI value, which is essential for effective decision-making in the development of strategic road maintenance plans.

Fig. 11 illustrates various scenarios to clarify the previous results, featuring detections from YOLOv8 alongside the corresponding predicted semantic masks from the enhanced LinkNet model. It presents four scenarios: (I) a case where the refinement is minimal (first paired row, first two columns); (II) a case where the refinement is noticeable (first paired row, columns three and four); (III) a case where the refinement is highly significant (second paired row, columns one and two); and (IV) a case involving mesh-type cracks, where the difference between bounding box and segmentation is less pronounced due to the substantial area of these cracks (second paired row, columns three and four).

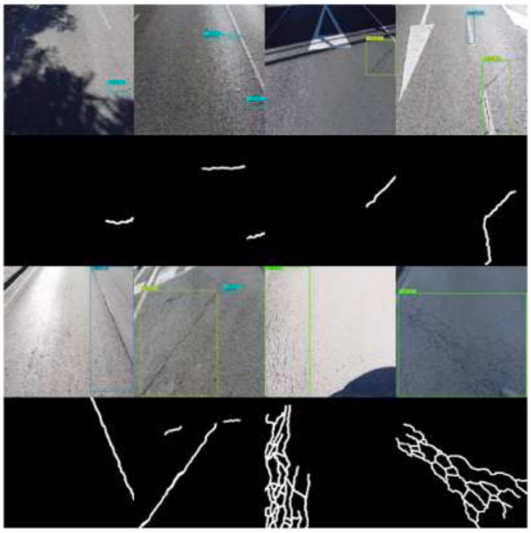

Fig. 11. Comparison of predicted bounding boxes with pre-trained YOLOv8 and predicted segmentation masks with improved pre-trained LinkNet to see area-based challenges concerning the calculus of the refined pavement condition index.

To summarize, the refined ASPDI index provides a more accurate assessment of crack areas by leveraging pixel-level analysis from YOLOv8 and LinkNet models. This enhancement not only improves the effectiveness of pavement maintenance strategies but also serves as a valuable metric for evaluating and comparing deep learning models in pavement distress detection.

5. Conclusions

This paper introduces a novel deep learning-based generative diffusion architecture for synthesizing crack images from segmentation masks. A multimodal, attention-based U-Net has been designed to incorporate both images and masks. The masks fuse semantic information using SPADE modules, and conditioning sampling is ensured with the modified CFG strategy. This system addresses the challenges associated with acquiring a large volume of diverse and realistic images. Moreover, as a conditional model, it eliminates the need for labelling the synthetic images, thereby reducing the costly annotating process. Additionally, it is a controllable model that can be guided through semantic masks to address specific problems related to pavement cracks. The main sub-conclusions are.

a) RoadPainter can generate a substantial volume of high-definition, conditionally cross-correlated, and diverse crack images.
b) The optimized RoadPainter model demonstrates superior IQA metrics with a more efficient configuration on the DeepCrack dataset, utilizing 134 million parameters, 128 channels, and a (1,2,3,4) channel multiplication vector.

c) Various noise schedules (linear, sigmoid, cosine, and stable diffusion) were investigated with a null SNR. The stable diffusion noise schedule yielded the most favorable metric values. Additionally, different CFG scales were assessed, with a CFG scale of 3 achieving the best results: 59.7 PSNR, 0.99 SSIM, 1.05 FID, and 0.0002 LPIPS on the DeepCrack dataset.
d) The optimal configuration of RoadPainter was evaluated in terms of IQA and visual inspection, demonstrating superior performance across all four datasets (DeepCrack, CFD, CrackSC, and Crack500) when compared to state-of-the-art semantic synthesis models, including Pix2Pix, CycleGAN, CDM, and SynCrack.
e) Standard segmentation metrics for five different segmentation architectures (U-Net, LinkNet, PSPNet, PAN, FPN) showed clear improvement after augmentation with the four benchmark datasets.
f) A refined pavement condition index for pavement distress detection has been developed and validated.

Future work will delve into latent diffusion models to reduce the complexity of diffusion models and explore the use of diffusion models with transformer-based denoising networks, rather than convolutional-based networks, to enhance segmentation performance for small crack datasets.

CRediT authorship contribution statement

Saúl Cano-Ortiz: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Eugenio Sainz-Ortiz:** Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Lara Lloret Iglesias:** Writing – review & editing, Validation, Supervision, Resources. **Pablo Martínez Ruiz del Árbol:** Writing – review & editing, Validation, Supervision, Resources. **Daniel Castro-Fresno:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work has been co-financed by the Ministry of Science and Innovation (ES) through the State Plan for Scientific and Technical Research and Innovation 2021-23 under the project MAPSIA [TED2021-129749B–I00]. Also, by the Horizon Europe Research and Innovation Framework program of the European Union under the project LIAISON [101103698].

References

[1] A. Saka, et al., GPT models in construction industry: o pportunities, limitations, and a use case validation, Developments in the Built Environment 17 (Mar. 2024) 100300, https://doi.org/10.1016/j.dibe.2023.100300.
[2] Z. Al-Huda, et al., Asymmetric dual-decoder-U-Net for pavement crack semantic segmentation, Autom Constr 156 (Dec. 2023) 105138, https://doi.org/10.1016/j.autcon.2023.105138.
[3] L. Yang, S. Bai, Y. Liu, H. Yu, Multi-scale triple-attention network for pixelwise crack segmentation, Autom Constr 150 (October 2022) (2023) 104853, https://doi.org/10.1016/j.autcon.2023.104853.
[4] S. Zhou, C. Canchila, W. Song, Deep learning-based crack segmentation for civil infrastructure: data types, architectures, and benchmarked performance, Autom Constr 146 (Feb. 2023) 104678, https://doi.org/10.1016/j.autcon.2022.104678.
[5] J. Guo, P. Liu, B. Xiao, L. Deng, Q. Wang, Surface defect detection of civil structures using images: review from data perspective, Autom Constr 158 (Feb. 2024) 105186, https://doi.org/10.1016/j.autcon.2023.105186.
[6] X. Wang, T. Wang, J. Li, Advanced crack detection and quantification strategy based on CLAHE enhanced DeepLabv3+, Eng. Appl. Artif. Intell. 126 (Nov. 2023) 106880 https://doi.org/10.1016/j.engappai.2023.106880.
[7] J. Ding, W. Li, L. Pei, M. Yang, C. Ye, B. Yuan, Sw-YoloX: an anchor-free detector based transformer for sea surface object detection, Expert Syst. Appl. 217 (May 2023) 119560, https://doi.org/10.1016/j.eswa.2023.119560.
[8] J. Ding, W. Li, L. Pei, M. Yang, A. Tian, B. Yuan, Novel pipeline integrating cross-modality and motion model for nearshore multi-object tracking in optical video surveillance, IEEE Trans. Intell. Transport. Syst. (2024) 1–13, https://doi.org/10.1109/TITS.2024.3373370.
[9] L. Pei, Z. Sun, L. Xiao, W. Li, J. Sun, H. Zhang, Virtual generation of pavement crack images based on improved deep convolutional generative adversarial network, Eng. Appl. Artif. Intell. 104 (July) (2021) 104376, https://doi.org/10.1016/j.engappai.2021.104376.
[10] S.-Y. Lee, T.H.M. Le, Y.-M. Kim, Prediction and detection of potholes in urban roads: machine learning and deep learning based image segmentation approaches, Developments in the Built Environment 13 (Mar. 2023) 100109, https://doi.org/10.1016/j.dibe.2022.100109.
[11] R. Nyirandayisabye, H. Li, Q. Dong, T. Hakuzweyezu, F. Nkinahamira, Automatic pavement damage predictions using various machine learning algorithms: evaluation and comparison, Results in Engineering 16 (Dec. 2022) 100657, https://doi.org/10.1016/j.rineng.2022.100657.
[12] Q. Du Nguyen, H.-T. Thai, Crack segmentation of imbalanced data: the role of loss functions, Eng. Struct. 297 (Dec. 2023) 116988, https://doi.org/10.1016/j.engstruct.2023.116988.
[13] Z. Pan, S.L.H. Lau, X. Yang, N. Guo, X. Wang, Automatic pavement crack segmentation using a generative adversarial network (GAN)-based convolutional neural network, Results in Engineering 19 (Sep. 2023) 101267, https://doi.org/10.1016/j.rineng.2023.101267.
[14] Y. Zhang, C. Liu, Network for robust and high-accuracy pavement crack segmentation, Autom Constr 162 (Jun. 2024) 105375, https://doi.org/10.1016/j.autcon.2024.105375.
[15] J. Liang, X. Gu, D. Jiang, Q. Zhang, CNN-based network with multi-scale context feature and attention mechanism for automatic pavement crack segmentation, Autom Constr 164 (Aug. 2024) 105482, https://doi.org/10.1016/j.autcon.2024.105482.
[16] Y. Gao, H. Cao, W. Cai, G. Zhou, Pixel-level road crack detection in UAV remote sensing images based on ARD-Unet, Measurement 219 (Sep. 2023) 113252, https://doi.org/10.1016/j.measurement.2023.113252.
[17] M. Ma, L. Yang, Y. Liu, H. Yu, An attention-based progressive fusion network for pixelwise pavement crack detection, Measurement 226 (Feb. 2024) 114159, https://doi.org/10.1016/j.measurement.2024.114159.
[18] J. Liang, Q. Zhang, X. Gu, Lightweight convolutional neural network driven by small data for asphalt pavement crack segmentation, Autom Constr 158 (Feb. 2024) 105214, https://doi.org/10.1016/j.autcon.2023.105214.
[19] Y. Huang, Y. Liu, F. Liu, W. Liu, "A Lightweight Feature Attention Fusion Network for Pavement Crack Segmentation," Computer-Aided Civil and Infrastructure Engineering, May 2024, https://doi.org/10.1111/mice.13225.
[20] J. Shang, et al., Automatic Pixel-level pavement sealed crack detection using Multi-fusion U-Net network, Measurement 208 (Feb. 2023) 112475, https://doi.org/10.1016/j.measurement.2023.112475.
[21] X. Wen, S. Li, H. Yu, Y. He, Multi-scale context feature and cross-attention network-enabled system and software-based for pavement crack detection, Eng. Appl. Artif. Intell. 127 (Jan. 2024) 107328, https://doi.org/10.1016/j.engappai.2023.107328.
[22] F. Guo, Y. Qian, J. Liu, H. Yu, Pavement crack detection based on transformer network, Autom Constr 145 (Jan. 2023) 104646, https://doi.org/10.1016/j.autcon.2022.104646.
[23] J. Wang, et al., Dual-path network combining CNN and transformer for pavement crack segmentation, Autom Constr 158 (Feb. 2024) 105217, https://doi.org/10.1016/j.autcon.2023.105217.
[24] J. Wang, et al., Dual-path network combining CNN and transformer for pavement crack segmentation, Autom Constr 158 (Feb. 2024) 105217, https://doi.org/10.1016/j.autcon.2023.105217.
[25] S. Cano-Ortiz, L. Lloret Iglesias, P. Martinez Ruiz del Árbol, D. Castro-Fresno, Improving detection of asphalt distresses with deep learning-based diffusion model for intelligent road maintenance, Developments in the Built Environment 17 (Mar. 2024) 100315, https://doi.org/10.1016/j.dibe.2023.100315.
[26] C. Han, T. Ma, J. Huyan, Z. Tong, H. Yang, Y. Yang, Multi-stage generative adversarial networks for generating pavement crack images, Eng. Appl. Artif. Intell. 131 (May 2024) 107767, https://doi.org/10.1016/j.engappai.2023.107767.
[27] S. Cano-Ortiz, P. Pascual-Muñoz, D. Castro-Fresno, Machine learning algorithms for monitoring pavement performance, Autom Constr 139 (Jul. 2022) 104309, https://doi.org/10.1016/j.autcon.2022.104309.
[28] H. Zhang, Z. Qian, W. Zhou, Y. Min, P. Liu, "A Controllable Generative Model for Generating Pavement Crack Images in Complex Scenes," Computer-Aided Civil and Infrastructure Engineering, Mar. 2024, https://doi.org/10.1111/mice.13171.
[29] F. Guo, Y. Qian, J. Liu, H. Yu, Pavement crack detection based on transformer network, Autom Constr 145 (Jan. 2023) 104646, https://doi.org/10.1016/j.autcon.2022.104646.
[30] I.J. Goodfellow, et al., Generative adversarial networks, arXiv:1406.2661 [stat.ML] (Jun. 2014).

[31] D.P. Kingma, M. Welling, An introduction to variational autoencoders, Foundations and Trends® in Machine Learning 12 (4) (2019) 307–392, https://doi.org/10.1561/2200000056.

[32] J. Ho, A. Jain, P. Abbeel, Denoising Diffusion Probabilistic Models, Jun. 2020 arXiv:2006.11239 [cs.LG].

[33] B. Xu, C. Liu, Pavement crack detection algorithm based on generative adversarial network and convolutional neural network under small samples, Measurement 196 (Jun. 2022) 111219, https://doi.org/10.1016/j.measurement.2022.111219.

[34] A. Radford, L. Metz, S. Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv:1511.06434 [cs.LG], Nov. 2015.

[35] Y. Que, et al., Automatic classification of asphalt pavement cracks using a novel integrated generative adversarial networks and improved VGG model, Eng. Struct. 277 (Feb. 2023) 115406, https://doi.org/10.1016/j.engstruct.2022.115406.

[36] D. Ma, H. Fang, N. Wang, C. Zhang, J. Dong, H. Hu, Automatic detection and counting system for pavement cracks based on PCGAN and YOLO-MF, IEEE Trans. Intell. Transport. Syst. 23 (11) (Nov. 2022) 22166–22178, https://doi.org/10.1109/TITS.2022.3161960.

[37] T. Zhang, D. Wang, A. Mullins, Y. Lu, Integrated APC-GAN and AttuNet framework for automated pavement crack pixel-level segmentation: a new solution to small training datasets, IEEE Trans. Intell. Transport. Syst. 24 (4) (Apr. 2023) 4474–4481, https://doi.org/10.1109/TITS.2023.3236247.

[38] X. Zhang, B. Peng, Z. Al-Huda, D. Zhai, FeatureGAN: combining GAN and autoencoder for pavement crack image data augmentations, Int. J. Image Graph. Signal Process. 14 (5) (Oct. 2022) 28–43, https://doi.org/10.5815/ijigsp.2022.05.03.

[39] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved Training of Wasserstein GANs," arXiv:1704.00028 [cs.LG], Mar. 2017.

[40] B. Yuan, Z. Sun, L. Pei, W. Li, M. Ding, X. Hao, Super-resolution reconstruction method of pavement crack images based on an improved generative adversarial network, Sensors 22 (23) (Nov. 2022) 9092, https://doi.org/10.3390/s22239092.

[41] Q. Song, L. Liu, N. Lu, Y. Zhang, R.C. Muniyandi, Y. An, A three-stage pavement image crack detection framework with positive sample augmentation, Eng. Appl. Artif. Intell. 129 (Mar. 2024) 107624, https://doi.org/10.1016/j.engappai.2023.107624.

[42] Y. Yan, S. Zhu, S. Ma, Y. Guo, Z. Yu, CycleADC-Net: a crack segmentation method based on multi-scale feature fusion, Measurement 204 (Nov. 2022) 112107, https://doi.org/10.1016/j.measurement.2022.112107.

[43] J. Song, P. Li, Q. Fang, H. Xia, R. Guo, Data augmentation by an additional self-supervised CycleGAN-based for shadowed pavement detection, Sustainability 14 (21) (Nov. 2022) 14304, https://doi.org/10.3390/su142114304.

[44] R. Rill-García, E. Dokladalova, P. Dokládal, Syncrack: improving pavement and concrete crack detection through synthetic data generation, in: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, SCITEPRESS - Science and Technology Publications, 2022, pp. 147–158, https://doi.org/10.5220/0010837300003124.

[45] H. Ranjbar, P. Forsythe, A.A.F. Fini, M. Maghrebi, T.S. Waller, Addressing practical challenge of using autopilot drone for asphalt surface monitoring: road detection, segmentation, and following, Results in Engineering 18 (Jun. 2023) 101130, https://doi.org/10.1016/j.rineng.2023.101130.

[46] H. Majidifard, Y. Adu-Gyamfi, W.G. Buttlar, Deep machine learning approach to develop a new asphalt pavement condition index, Constr Build Mater 247 (Jun. 2020) 118513, https://doi.org/10.1016/j.conbuildmat.2020.118513.

[47] E. Ibragimov, Y. Kim, J.H. Lee, J. Cho, J.-J. Lee, Automated pavement condition index assessment with deep learning and image analysis: an end-to-end approach, Sensors 24 (7) (Apr. 2024) 2333, https://doi.org/10.3390/s24072333.

[48] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, Nov. 2016 arXiv:1611.07004 [cs.CV].

[49] J. Kang, S. Feng, Pavement cracks segmentation algorithm based on conditional generative adversarial network, Sensors 22 (21) (Nov. 2022) 8478, https://doi.org/10.3390/s22218478.

[50] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional Networks for Biomedical Image Segmentation, May 2015 arXiv:1505.04597 [cs.CV].

[51] U. Demir, G. Unal, Patch-Based Image Inpainting with Generative Adversarial Networks, Mar. 2018 arXiv:1803.07422 [cs.CV].

[52] J. Ho, T. Salimans, Classifier-Free Diffusion Guidance, Jul. 2022 arXiv:2207.12598 [cs.LG].

[53] Y. Du, et al., ArSDM: Colonoscopy Images Synthesis with Adaptive Refinement Semantic Diffusion Models, Sep. 2023 arXiv:2309.01111 [cs.CV].

[54] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, Dec. 2015 arXiv:1512.03385 [cs.CV].

[55] A. Dosovitskiy, et al., An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, Oct. 2020 arXiv:2010.11929 [cs.CV].

[56] Y. Liu, J. Yao, X. Lu, R. Xie, L. Li, DeepCrack: a deep hierarchical feature learning architecture for crack segmentation, Neurocomputing 338 (Apr. 2019) 139–153, https://doi.org/10.1016/j.neucom.2019.01.036.

[57] Y. Shi, L. Cui, Z. Qi, F. Meng, Z. Chen, Automatic road crack detection using random structured forests, IEEE Trans. Intell. Transport. Syst. 17 (12) (Dec. 2016) 3434–3445, https://doi.org/10.1109/TITS.2016.2552248.

[58] F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, H. Ling, Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection, Jan. 2019 arXiv:1901.06340 [cs.CV].

[59] S. Cano-Ortiz, L. Lloret Iglesias, P. Martinez Ruiz del Árbol, P. Lastra-González, D. Castro-Fresno, An end-to-end computer vision system based on deep learning for pavement distress detection and quantification, Constr Build Mater 416 (Feb. 2024) 135036, https://doi.org/10.1016/j.conbuildmat.2024.135036.

[60] S. Lin, B. Liu, J. Li, X. Yang, Common Diffusion Noise Schedules and Sample Steps Are Flawed, May 2023 arXiv:2305.08891 [cs.CV].

[61] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid Scene Parsing Network, Dec. 2016 arXiv:1612.01105 [cs.CV].

[62] H. Li, P. Xiong, J. An, L. Wang, Pyramid Attention Network for Semantic Segmentation, May 2018 arXiv:1805.10180 [cs.CV].

[63] A. Chaurasia, E. Culurciello, LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation, Jun. 2017, https://doi.org/10.1109/VCIP.2017.8305148.

[64] A. Kirillov, R. Girshick, K. He, P. Dollár, Panoptic Feature Pyramid Networks, Jan. 2019 arXiv:1901.02446 [cs.CV].

[65] D. Reis, J. Kupec, J. Hong, A. Daoudi, Real-Time Flying Object Detection with YOLOv8, May 2023 arXiv:2305.09972 [cs.CV].