The transformative roles of generative artificial intelligence in vision techniques for structural health monitoring

A state-of-the-art review

Duan, Shundi; Tan, Xiao; Guo, Pengwei; Guo, Yurong; Bao, Yi

DOI
10.1016/j.aei.2025.103719

Publication date
2025

Document Version
Final published version

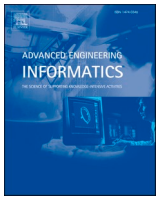
Published in
Advanced Engineering Informatics

Advanced Engineering Informatics 68 (2025) 103719

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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



Review article

The transformative roles of generative artificial intelligence in vision techniques for structural health monitoring: A state-of-the-art review

Shundi Duan [a,b], Xiao Tan [c], Pengwei Guo [d,e,*], Yurong Guo [a,b,*], Yi Bao [e]

[a] College of Civil Engineering, Hunan University, No. 1 Lushan Road, Changsha 410082, China
[b] Key Laboratory of Building Safety and Energy Efficiency, Hunan University, China
[c] College of Water Conservancy and Hydropower Engineering, Hohai University, No. 1 Xikang Road, Nanjing 210024, China
[d] Department of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, Delft 2628 CN, The Netherlands
[e] Department of Civil, Environmental and Ocean Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, United States

ARTICLE INFO

Keywords:
Generative artificial intelligence
Structural health monitoring
Image restoration
Data augmentation
Multi-modal generative AI
Large language model

ABSTRACT

As urbanization accelerates, aging infrastructure demands more advanced inspection methods for structural health monitoring. The growing integration of artificial intelligence (AI) and computer vision technologies has significantly enhanced damage detection accuracy while simultaneously reducing inspection time and operational costs. Despite these advantages, the adoption of AI-based technologies in infrastructure maintenance remains limited due to challenges related to data. One major issue is the lack of comprehensive, task-specific annotated datasets. Another is the poor quality of images captured by drones or mobile devices, which are often affected by noise, blurring, and inconsistent lighting. Although recent advances in generative AI offer promising support for structural health monitoring, it remains unclear which models are best suited for specific tasks.

This study examines the use of generative AI in structural health monitoring, focusing on key challenges such as limited datasets and low-quality image restoration. The review covers a range of generative AI technologies, outlining their principles, strengths, limitations, and representative applications to support the selection of appropriate tools for specific tasks. Generative AI models enable accurate image segmentation and structural anomaly detection using limited training data. The paper also explores new opportunities for integrating multi-modal generative AI to enhance human–computer interaction in support of structural health monitoring. A framework is proposed to streamline the use of generative AI technologies for data augmentation, image restoration, damage inspection, and human–computer interaction in structural health monitoring.

1. Introduction

Aging infrastructure poses serious risks to public safety, economic growth, and the efficient use of resources. According to the ASCE, 6.8 % of the 623,000 bridges in the U.S. are in poor condition [1], especially in disaster-prone regions. Timely inspection is essential to identify early

Abbreviations: AI, Artificial Intelligence; CNN, Convolutional Neural Network; VGG, Visual Geometry Group; ResNet, Residual Network; Faster R-CNN, Faster Region-Based Convolutional Neural Network; YOLO, You Only Look Once; FCN, Fully Convolutional Networks; GANs, Generative Adversarial Networks; DCGAN, Deep Convolutional Generative Adversarial Network; WGAN, Wasserstein Generative Adversarial Network; WGAN-GP, Wasserstein Generative Adversarial Network with Gradient Penalty; BCE, Binary Cross-Entropy; AdaIN, Adaptive Instance Normalization; GPU, Graphics Processing Unit; SRGAN, Super-Resolution Generative Adversarial Network; CycleGAN, Cycle-Consistent Adversarial Network; ERFNET, Efficient Residual Factorized Convnet; CGNET, Context Guided Network; LEDNET, Lightweight Encoder-Decoder Network; SGAN, Semi-Supervised GAN; CGANs, Conditional GANs; ESRGAN, Enhanced Super-Resolution Generative Adversarial Networks; CDU-Net, Context-Encoding Network; LSTM, Long Short-Term Memory; VAEs, Variational Autoencoders; LLM, Large Language Model; mAP, Mean Average Precision; MSE, Mean Square Error; IS, Inception Score; FID, Fréchet Inception Distance; IoU, Intersection over Union; PSNR, Peak Signal-to-Noise Ratio; SSIM, Structural Similarity Index; DDPMs, Denoising Diffusion Probabilistic Models; LDM, Latent Diffusion Model; CLIP, Contrastive Language-Image Pre-training; BERT, Bidirectional Encoder Representations from Transformers; SDIGLM, Structural Damage Identification on Generative Large Language Models; LoRA, Low-Rank Adaptation; VQA, Visual Question Answering; UAV, Unmanned Aerial Vehicle; ITC, Image-Text Contrastive; GT, Ground Truth.

* Corresponding authors at: Department of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, Delft 2628 CN, the Netherlands (P. Guo) College of Civil Engineering, Hunan University, No. 1 Lushan Road, Changsha 410082, China (Y. Guo).
E-mail addresses: pengweiguo@tudelft.nl (P. Guo), yurongguo@hnu.edu.cn (Y. Guo).

https://doi.org/10.1016/j.aei.2025.103719
Received 26 February 2025; Received in revised form 24 July 2025; Accepted 25 July 2025
Available online 28 July 2025

damage and reduce long-term repair costs. However, traditional inspection methods are labor-intensive and time-consuming [2]. Recent advances in deep learning and computer vision enable faster and more automated damage assessments [3]. Drones equipped with cameras can capture images of infrastructure [4,5], which are analyzed using AI models such as CNNs (e.g., VGG, ResNet) for crack detection [6]. Object detection models such as Faster-RCNN and YOLO localize damage [7,8], while segmentation models like Transformers, U-Net, and DeepLabV3 + help measure crack widths from annotated datasets [9–11].

The quality of datasets is critical for both training deep learning models and ensuring their ability to generalize across different environments [12]. Limited data availability remains a major challenge. In many scenarios, data may be scarce or difficult to obtain. For example, fiber-reinforced concrete tends to generate dense microcracks, which models trained on conventional concrete datasets often fail to detect accurately [13]. Additionally, small-scale datasets are insufficient for training highly accurate detection models, limiting their effectiveness in real-world applications. Data imbalance remains a critical issue, as most datasets are heavily skewed toward cracks, while defects like corrosion and exposed rebar are scarce, reducing model accuracy in detecting these underrepresented categories [14]. Additionally, poor image quality caused by factors such as inadequate lighting, motion blur, shadows, and noise also hinders accurate analysis [15]. These challenges highlight the need for synthetic data and image restoration techniques to enable reliable AI deployment in real-world scenarios.

Generative AI techniques, especially GANs, have gained extensive attention for their capabilities to produce high-quality synthetic data through adversarial training between the generator and the discriminator networks [16]. In structural health monitoring, generative AI refers to machine learning techniques that generate realistic damage representations by learning patterns from existing data. Various GANs have already shown promise for multiple tasks [17]. For example, DCGAN and WGAN-GP can generate simulated damage data that augments existing datasets, offering more diverse features in terms of texture, shape, and pixel intensity [18,19]. StyleGAN has been employed to enhance the performance of crack recognition through style transfers, thereby improving the accuracy of deep learning models in infrastructure inspection [20]. Additionally, SRGAN has proven effective for super-resolution reconstruction, converting low-resolution images into high-resolution ones to boost the precision of inspection models [21,22]. The primary data analyzed in the reviewed studies are image-based, acquired through drones and handheld devices [23,24]. Although crack detection remains the most studied application [25,26], some works have addressed other damage types, such as corrosion, spalling, and exposed rebar [14]. Collectively, these advanced techniques address key limitations in current workflows, including data scarcity and image quality issues, leading to more efficient and reliable inspection strategies. Despite recent advances in generative AI for data augmentation and image enhancement, the optimal models for specific tasks remain uncertain.

Advanced AI technologies such as vision-language models and LLMs are opening new opportunities for innovation in infrastructure inspection. Vision-language models integrate visual and textual data, facilitating accurate defect classification and automated reporting [27]. In computer vision-based structural health monitoring, integrating LLMs can improve human–machine interaction and interpretation of complex inspection data. A recent study employed multi-modal GPT-4o mini for zero-shot detection of fatigue cracks in steel bridges [28]. It enables the chatbot to interactively analyze images, interpret damage descriptions, and provide real-time feedback on both visual data and textual prompts. Contextual object detection with LLMs further enhances inspection processes by enabling models to identify and interpret objects within complex scenes using contextual information [29]. Together, these technologies significantly improve efficiency, precision, and contextual understanding of inspection workflows. Notably, vision-language models and LLMs are emerging technologies with limited use in damage detection. A comprehensive review is needed to explore current developments, identify key challenges, and outline future research directions.

To address these challenges, this study systematically evaluates the application of generative AI in computer vision-based structural health monitoring, focusing on bridges, pavements, and buildings. This review covers multiple damage types, though most studies focus on cracks, with limited studies on cavities, spalling, corrosion, and exposed rebar. The contributions of this research are summarized as follows: (1) Provide an in-depth analysis of the benefits and limitations of various generative AI models and their specific applications in damage assessment. (2) Conduct a comprehensive comparison to determine the most suitable methods for addressing different scenarios effectively. (3) Review innovative methods with potential applications in structural health monitoring and propose future directions for leveraging generative AI to enhance inspection practices. (4) A framework is proposed to streamline the use of generative AI technologies for data augmentation, image restoration, damage inspection, and human–computer interaction in structural health monitoring.

In summary, generative AI plays a transformative role in computer vision-based structural health monitoring by expanding the capabilities of image generation, enhancement, and interpretation. It addresses limitations such as data scarcity, imbalance, and low image quality by generating realistic synthetic data and simulating diverse defect scenarios. In addition, the advancement of visual language models and large language models improves human–computer interaction by enabling more intuitive interpretation of visual data. Their integration of extensive text-based knowledge supports more informed, context-aware decision-making, helping shift infrastructure maintenance from reactive to predictive approaches.

2. Overview

2.1. Statistical analysis

Based on the scope of the research, a keyword search was conducted focusing on the application of generative AI in damage inspection of pavements, building structures, and bridges. The targeted damage types include cracks, spalling, cavities, corrosion, and exposed rebar. The topics covered include data augmentation, image restoration, image segmentation, and multi-modal generative AI. The search utilized keyword combinations such as {“Generative AI”} and {“Concrete” or “Pavement” or “Bridge”} and {“Damage” or “Crack” or “Defect”}. The resulting literature will serve as key references for this review article. To ensure the selected research literature exhibits high relevance, the following steps will be implemented: (1) An initial keyword search was conducted in the Scopus database within the time frame of 2020–2025 to ensure the selected literature represents the latest research trends. (2) All relevant literature identified through the keyword search will be exported, including information such as publication year, DOI, and keywords. (3) Duplicate entries were removed based on DOI. (4) The first 100 articles were reviewed and served as the primary reference sources for this study. The keyword search initially retrieved 520 articles. After removing 85 duplicates and 176 irrelevant entries, 259 unique articles (Fig. 1) were retained. Together with other supplementary references, resulting in 133 references.

Fig. 1 shows the number of publications on generative AI in infrastructure maintenance from 2020 to 2025, highlighting a clear upward trend in related research in recent years. There are two main reasons behind the rapid development of generative AI in infrastructure maintenance. First, the scale of real-world data on infrastructure is limited, including damage data from bridges, roads, and buildings. Many datasets only consist of a few hundred to a few thousand images [30–32]. This scarcity necessitates the use of generative AI to augment datasets, thereby improving the performance of deep learning models tailored for structural health monitoring. Second, the rapid advancement of

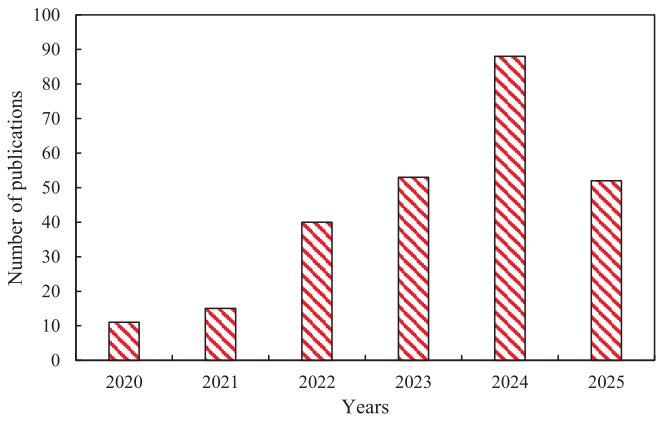


Fig. 1. Publications on generative AI-assisted infrastructure maintenance from 2020 to 2025.

generative AI reflects that performance in tasks like damage detection has largely converged around established models such as DeepLabV3+, SegFormer, and YOLO. Many practical challenges, such as damage inspection, pothole detection, and corrosion assessment, have been addressed satisfactorily. As a result, it is becoming increasingly difficult for new publications to offer novel contributions in these areas. In contrast, the field of generative AI remains in a dynamic and fast-evolving phase, continually generating fresh concepts and applications. For example, text-to-image generation and large language models are not commonly applied in structural health monitoring.

A comprehensive knowledge map for generative AI in structural health monitoring is developed by analyzing core concepts such as generative adversarial networks and their connections to deep learning, crack detection, pavement cracks, image restoration, synthetic data, super-resolution, semantic segmentation models, and others. These relationships are visualized in Fig. 2. This map visualizes the relationships between key concepts, methodologies, and applications within the field. Keywords from each reference are categorized and connected, highlighting the most frequently occurring terms and their linkages to related topics. The most frequent keyword in the knowledge map is "generative adversarial network," indicating its prominence as the dominant model in the literature. This model is closely associated with key civil engineering applications such as "crack "detection," crack segmentation," "semantic segmentation." Cracks remain the primary focus of structural damage analysis. A temporal analysis of keyword co-occurrence reveals clear shifts over time. From 2020 to 2022, research was heavily centered on GANs, which appeared in approximately 70 percent of papers, often in conjunction with terms like "crack detection" and "data augmentation," reflecting their central role in synthetic data generation for structural inspection tasks. However, from 2023 to 2025,

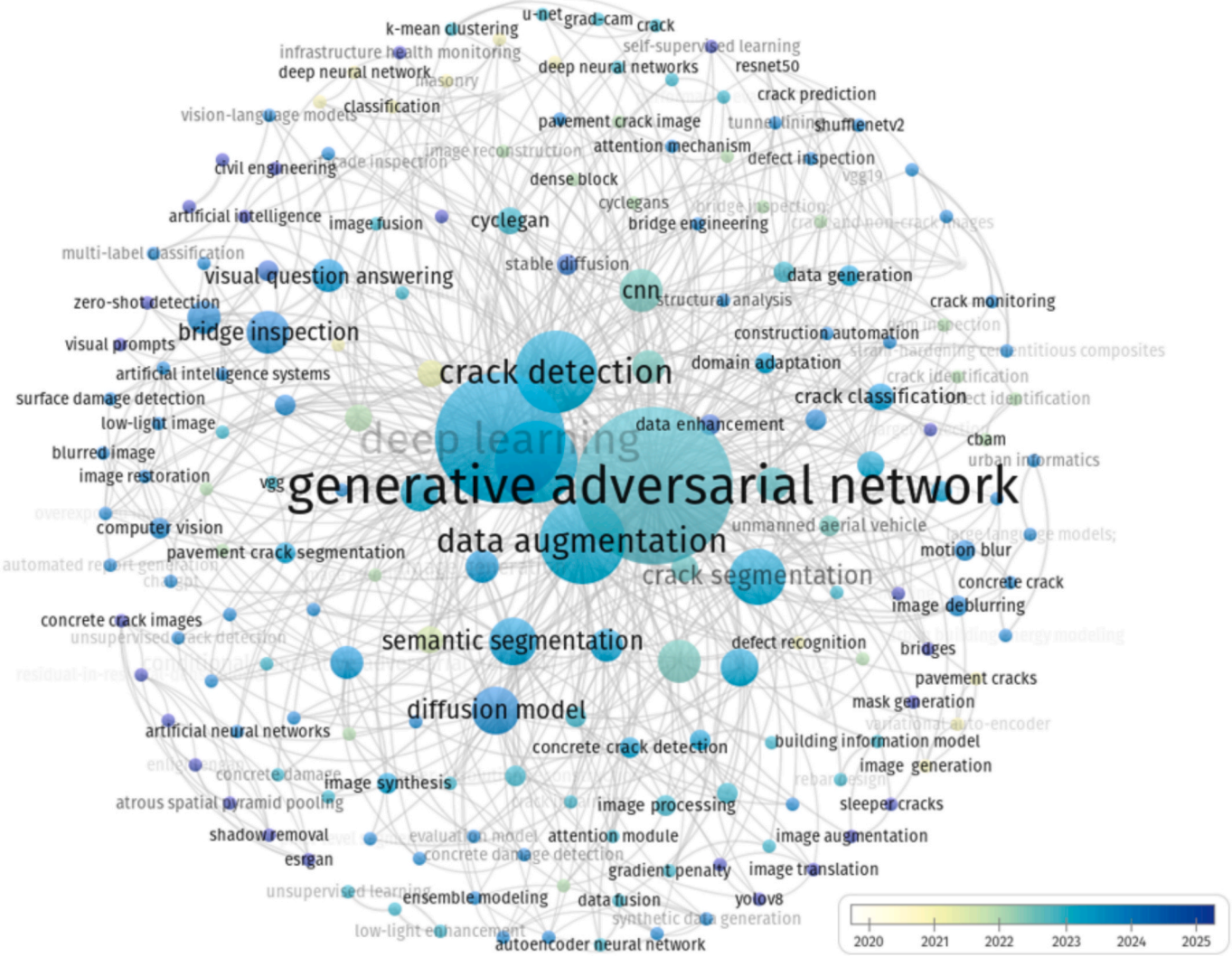


Fig. 2. Keyword co-occurrence analysis over time.

keywords like “diffusion model” and “large language model” surged in frequency, marking a transition toward higher-fidelity image synthesis, such as Stable Diffusion, and AI-assisted analysis and reporting, such as ChatGPT. During this period, GAN mentions declined by about 20 percent. Domain-specific terms such as “bridge inspection” and “pavement crack” remained consistently present but showed evolving associations. Notably, co-occurrences between “LLM” and “visual question answering” began to emerge after 2024, suggesting growing interdisciplinary integration. Meanwhile, core technical terms like “deep learning” and “semantic segmentation” persisted across all years, underscoring their foundational importance to the field.

2.2. Roadmap of generative AI development

Fig. 3 presents a detailed roadmap outlining the evolution and application of generative AI from its inception in 2013 to projected advancements through 2025. The roadmap is structured into five main periods: the VAE, GANs, Diffusion Models, LLM, and Multi-modal generative AI, which are subsequently introduced as follows: (1) **VAEs** combine deep learning with Bayesian inference by mapping inputs to a probabilistic latent space and sampling from it to generate variable outputs. The decoder reconstructs these to resemble the original data. VAEs are widely used in structural health monitoring, damage detection, data augmentation, and predictive maintenance [33]. (2) **GAN models** are crucial in structural health monitoring by generating high-quality synthetic data, enhancing image resolution, and identifying anomalies. They expand training datasets with synthetic images of defects, boosting the performance of AI models [34]. GAN models include specialized variants tailored to specific tasks, such as DCGANs for data generation [35], Conditional GANs for image translation [36], StyleGANs for style transfer [37], and SRGANs for image super-resolution [38]. (3) **Diffusion models** are a type of generative model that progressively transform simple noise distributions into complex data distributions [39]. Stable Diffusion particularly excels in text-to-image generation, which is able to convert textual descriptions into highly detailed and realistic images [40]. (4) **Multi-modal generative AI** combines domains like text, image, audio, and video generation, enabling cohesive cross-media outputs [41]. Examples include DALL·E [42] and Imagen [43], which convert text into detailed images, and Meta’s Make-a-Video [44], which generates videos from text prompts. LLMs like GPT-4, built on transformer architectures, excel in language tasks such as text generation and summarization [45]. In civil engineering, they automate documentation [46], assist in design [47], and predict maintenance needs [48], enhancing accuracy, efficiency, and innovation in infrastructure development. In recent years, GPT-4 has evolved into a multi-modal model capable of handling not only text data but also processing data related to images, audio, and more.

3. Generative AI applications

Section 3 is organized into five main sections that collectively represent a progressive workflow for applying generative AI techniques in structural health monitoring. Section 3.1 begins with dataset augmentation, which addresses the challenge of limited or imbalanced training data and lays the foundation for robust model development. Section 3.2 covers image restoration, which enhances data quality by mitigating issues such as noise and poor lighting conditions. When combined, data augmentation and image restoration contribute to the creation of high-quality datasets. Building on this, Section 3.3 focuses on image segmentation, which relies on improved data quality to accurately isolate structural features or defects, enabling precise localization of damage. Finally, Section 3.4 discusses multi-modal generative AI, which integrates visual and textual modalities to enhance interpretation. Section 3.5 summarizes key challenges and future research directions. These sections trace a path from data enhancement to high-level insight, highlighting the impact of generative AI on structural health monitoring.

3.1. Dataset augmentation

Generative AI models have emerged as a powerful technique for data augmentation, providing an effective way to enlarge datasets. The primary issues with the datasets are data scarcity and data imbalance [49]. Data scarcity refers to a lack of sufficient data for training, while data imbalance occurs when some classes are underrepresented, leading to poor performance on those classes. Traditional methods used in computer vision tasks for data augmentation include image cropping, flipping, rotation, and scaling [50]. These methods are easy to use but limit the diversity of augmented images, potentially leading to repetitive datasets. In contrast, synthetic data from generative AI models can introduce variations that were not present in the original dataset, helping to create more robust models that generalize better to unseen data [51].

3.1.1. Unsupervised dataset augmentation

The original GAN is a classic unsupervised neural network model [52]. To provide a deeper understanding of GANs, this paper will provide a detailed discussion from several perspectives, including the fundamental concept and model architecture [16,53]. GANs comprise a generator and a discriminator [54]. The objective of GANs is to train a generator to create highly realistic data through adversarial training [55]. The generator takes random noise as input and generates synthetic data, while the discriminator attempts to distinguish between the real data and the data generated by the generator [13,56]. Fig. 4 illustrates the GAN model used for data generation.

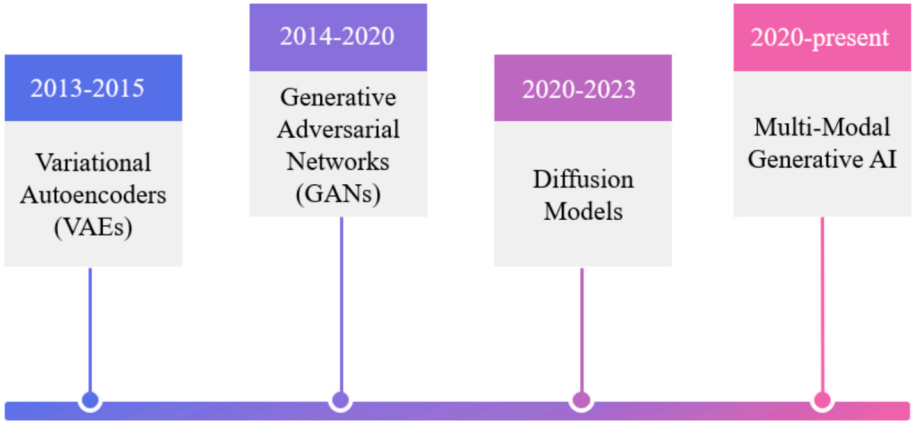


Fig. 3. Roadmap of the development of generative AI techniques.

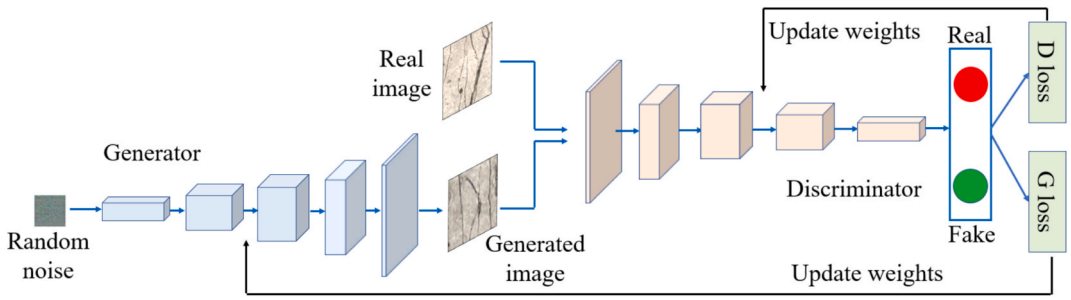


Fig. 4. Illustration of a GAN model for data generation [57].

A typical GAN used for data generation is DCGAN. DCGAN includes both a generator and a discriminator [58]. In the generator, transpose convolutional layers (also known as deconvolutional layers) up-sample the noise to the desired image size [59]. In the discriminator, convolutional layers down-sample the input image, with a Sigmoid function in the output layer to produce a probability score indicating whether the image is real or fake [60]. DCGAN training begins with initializing random weights for both the generator and discriminator. The generator creates a fake image from random noise and feeds it to the discriminator [61]. The discriminator is trained using both real images and the fake images produced by the generator [62]. It calculates the loss based on a classification loss function (e.g., BCE loss) and updates its weights to minimize this loss [60]. This process is repeated over many epochs until the generator produces images that are sufficiently realistic [63]. To address the vanishing gradient problem, the WGAN replaces the BCE loss with Wasserstein loss [64]. The goal of Wasserstein loss is to minimize the Wasserstein distance between the real and generated data distributions. Further improvements are made by incorporating a gradient penalty term into the Wasserstein loss, resulting in the WGAN-GP model [18,65]. This gradient penalty term is added to the critic's loss function to enforce the Lipschitz constraint, which is crucial for stable training. The gradient penalty encourages the gradient norm to be close to 1, ensuring that the critic adheres to the Lipschitz continuity requirement.

StyleGAN introduces a novel style-based generator architecture [20,66]. StyleGAN transforms the latent vector z into an intermediate latent space w. Intermediate latent space controls the style at each convolutional layer through AdaIN [20]. AdaIN enables the application of styles at different layers, allowing control over coarse, middle, and fine features of the generated images. The network generates images from the intermediate latent code w, with styles applied at each layer to influence the final image. Additionally, StyleGAN uses progressive growing, starts with a low resolution and gradually increases the resolution of generated images during training. This technique aids in stabilizing the training process and enhances image quality.

The data generation framework is summarized in Fig. 5. Initially, the original dataset is collected and fed into a generative AI model. This model generates additional synthetic images, expanding the original image dataset. The enlarged dataset is then used to train various models: classification models (e.g., VGG-16 and ResNet), object detection models (e.g., Faster R-CNN and YOLO) using the dataset with bounding boxes, and semantic segmentation models (e.g., U-Net and SegNet) using the dataset with masks. The DCGAN, WGAN, and WGAN-GP models are types of unsupervised GANs designed to generate data without requiring labeled datasets. This ability is particularly valuable because acquiring labeled data is often a challenging, time-consuming, and costly process. By leveraging these unsupervised GANs, researchers can efficiently produce high-quality synthetic data, improve the performance of machine learning models, and facilitate various applications where labeled data is scarce or unavailable.

Generative AI models have been utilized to enhance datasets for various tasks such as classification, object detection, and semantic segmentation. In [67], DCGAN was used to generate high-resolution images at 256 × 256 pixels. The size of dataset increased from 4,160 to 9,600 images. This study categorized the dataset into five types of pavement defects: horizontal crack, vertical crack, alligator crack, pothole, and non-crack. The VGG16 model was employed to classify these pavement defects. By augmenting the dataset with images generated using DCGAN, the classification accuracy of VGG16 increased from 88.6 % to 91.4 %, demonstrating the significant impact of generated data on improving model performance. In [18], GAN models were used to augment datasets for pavement crack detection. This study proposed an improved WGAN-GP model to generate 512 × 512 pixels pavement images, addressing the issue of data scarcity. The study creates a synthesized dataset of grooved pavement crack images by combining generated crack images with real images. The robustness of the improved WGAN-GP model was validated using Faster R-CNN, YOLOv3, and YOLOv4 models for region-level detection, increasing the mAP scores from 68.0 %, 75.6 %, and 72.9 % to 74.6 %, 82.0 %, and 80.2 %, respectively. In [16], a GAN model was developed to augment a multi-

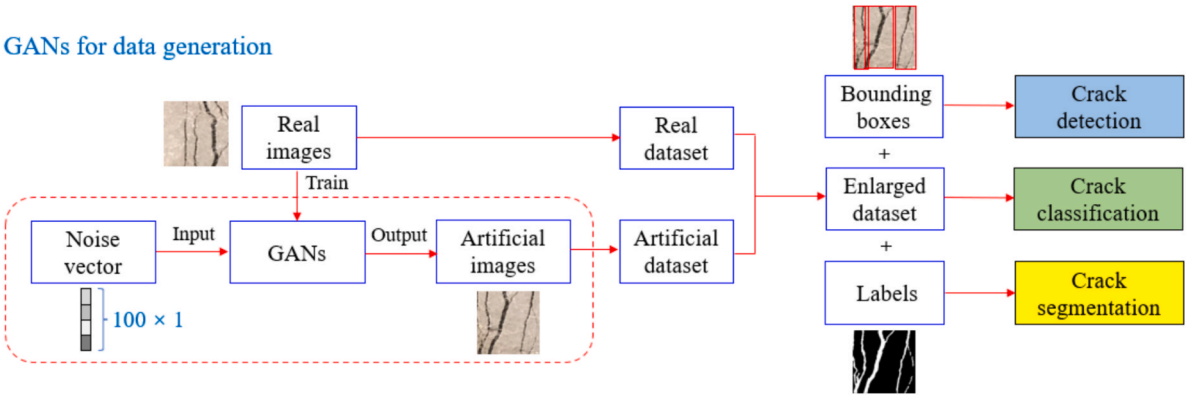


Fig. 5. Illustration of a data generation framework using GANs for improving deep learning performance.

class image dataset for damage classification. The augmented dataset improved the average accuracy of the VGG model from 85.7 % to 97.6 %. Table 1 summarizes the studies that apply GANs for data augmentation, including the specific GAN models used, the types of applications, as well as the corresponding detection and classification models. Additionally, it quantitatively presents the performance metrics, highlighting the accuracy of each model.

3.1.2. Unpaired dataset augmentation

Unpaired data augmentation GANs enhance datasets without requiring paired samples. When image-to-image translation pairs are hard to obtain, models like CycleGAN use unpaired data from two domains, leveraging dual generators and discriminators with cycle consistency to preserve original content [75]. A key example is CycleGAN, which uses dual generators and discriminators with cycle consistency to translate between domains while preserving original content [76–78]. In [77], the study addressed the challenge of detecting pavement cracks under shadow interference, which hinders the performance of deep learning models. The authors used CycleGAN to generate realistic shadowed crack images from unshadowed ones, without the need for paired training data, leading to improved segmentation performance of the U-Net model. In [79], CycleGAN was used to generate 500 synthetic damage images, including cracks and spalling in concrete structures. Using this augmented dataset, the damage segmentation accuracy of the DeepLabV3 + model improved from 75 % to 90 %. In [80], The CycleGAN was used to augment pavement damage data, including cracks and potholes, while boosting the mAP of YOLOv5 from 77.0 % to 85.0 %. In [81], CycleGAN was utilized to generate synthetic images and synthetic labels, effectively doubling the crack dataset size from 1,703 to 3,406. The expanded dataset improved crack segmentation accuracy of attention U-Net model, raising the mAP score from 95.2 % to 97.5 %. The illustration of using CycleGAN for data augmentation is shown in Fig. 6.

3.1.3. Semi-supervised dataset augmentation

In the original GAN, classifiers typically categorize input data into real or fake [82]. However, in a semi-supervised GAN, the discriminator is enhanced into a multi-class classifier that not only distinguishes between real and fake images but also categorizes real images into one of the $N + 1$ classes, where N represents the number of classes in the training dataset, with the additional class representing fake samples generated by the generator [83–85]. Compared to supervised learning methods, semi-supervised learning significantly reduces the need for labeled data while maintaining high classification accuracy [86]. For instance, a semi-supervised GAN a semi-supervised GAN can classify whether test samples contain a crack or not [87]. The ratio of labeled to unlabeled samples can be adjusted as a variable parameter, and additional unlabeled samples can also be incorporated to further enhance the performance of the model. The ratio of labeled to unlabeled samples ranged from 1/5 to 1/30, resulting in classification accuracy exceeding 93.5 %. The highest accuracy of 0.953 and F1-score of 0.976 were achieved by using only 1/5 of the labeled samples combined with an additional 10,000 unlabeled samples [87]. The illustration of semi-supervised SGAN is shown in Fig. 7.

3.1.4. Supervised dataset augmentation

Supervised dataset augmentation relies on paired data, where each input corresponds directly to a labeled output [88]. In [79], CGANs are used to generate concrete damage (cracks and spalling) from hand-painted semantic masks. The study utilized models such as pix2pix, OASIS, and pix2pixHD for data augmentation, and compared the quality of generated images using IS and FID. The pix2pixHD model achieved a higher IS score of 2.41 and a lower FID score of 121.3, indicating better performance in generating synthetic data. Images generated using the pix2pixHD model were further used for crack segmentation. These generated images were labeled to train segmentation models such as FCN, PSPNet, and DeepLabV3+, resulting in improved mIoU scores from 82 % to 90 %, 85 % to 89 %, and 75 % to 90 %, respectively [79]. In [89], an L1-CGAN was proposed to generate bridge damage images such as rebar exposure based on segmentation masks. The model was trained using 208 concrete bridge images and produced 840 synthetic samples. The augmented dataset increased the mIoU of SegNet from 65.7 % to 81.4 %. In [36], pix2pix was applied to generate image with crack from segmentation map. The augmented dataset was further used to train segmentation models, including FCN, U-Net, and DeepLabV3 + . The enlarged dataset consisted of 7,800 synthesized crack images and 7,800 real crack images. The results demonstrated that the models trained with synthesized images achieved mIoU scores exceeding 74 %. An illustration of using CGAN for data generation is shown in Fig. 8.

3.1.5. Text to image

Stable Diffusion generates detailed images from text descriptions using a latent diffusion process [40]. As an LDM, it combines VAEs, U-Net architectures, and transformer-based text encoders to create high-quality images. The model employs a forward diffusion process to add

Table 1
Summary of GAN models for data generation.

No.	Year	GAN model	Application	Task	Deep learning	Accuracy (%)	Ref.
1	2021	DCGAN	Pavement crack	Object detection	Faster R-CNN	84.9 to 87.8	[19]
2	2022	GAN	Building crack, rebar exposure, delamination, leakage	Classification	VGG16	85.7 to 97.6	[16]
3	2022	GAN	Building crack, rebar exposure, delamination, leakage classification	Classification	ResNet-50	75 to 96.1	[16]
4	2022	GAN	Building crack, rebar exposure, delamination, leakage classification	Classification	MobileNetV2	68.9 to 96.9	[16]
5	2022	DCGAN	Building crack	Classification	Deep CNN	91.4 to 92.8	[68]
6	2022	StyleGAN	Bridge crack	Classification	ConvNeXt	Up to 100	[20]
7	2022	StyleGAN	Bridge crack	Classification	ResNet-152	Up to 99.9	[20]
8	2023	APC-GAN	Pavement crack	Semantic segmentation	U-Net	81.2 to 83.6	[69]
9	2023	WGAN-GP	Pavement crack	Object detection	YOLOv4	72.9 to 80.2	[18]
10	2023	DCGAN	Pavement crack	Classification	VGG16	88.6 to 91.4	[67]
11	2023	DCGAN	Bridge crack and pitting	Classification	Normal CNN	72.1 to 76.0	[70]
12	2024	HGAN	Building crack	Semantic segmentation	LinkNet	92.5 to 97	[13]
13	2024	HGAN	Building crack	Semantic segmentation	Vision transformer	93.7 to 98.2	[13]
14	2024	SEGAN	Building crack	Semantic segmentation	U-Net	83.4 to 94.5	[71]
15	2024	MCT2GAN	Building crack	Semantic segmentation	DeepCrack	68.4 to 71.8	[72]
16	2025	DCGAN	Building crack, spalling, leakage	Object detection	YOLOv5	75.6 to 81.4	[73]
17	2025	MaskGAN	Building crack	Object detection	YOLOv5-seg	88.2 to 98.6	[74]

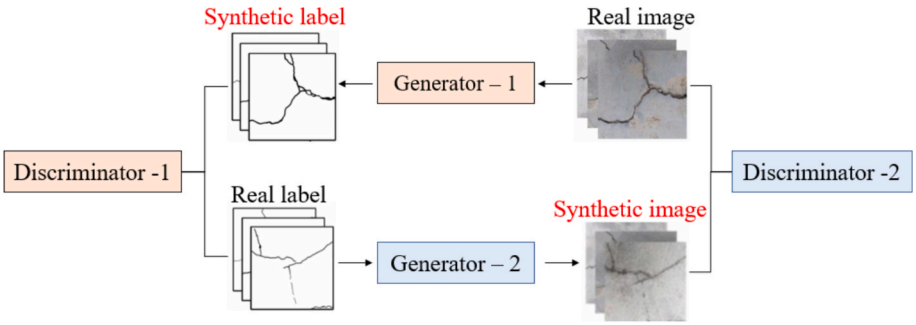


Fig. 6. Illustration of a CycleGAN architecture for image generation [81].

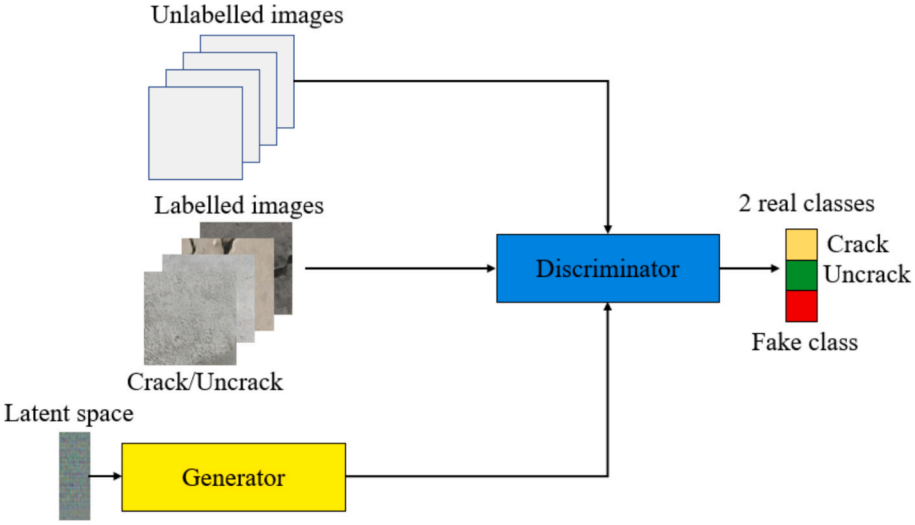


Fig. 7. Illustration of a SGAN architecture for data generation [87].

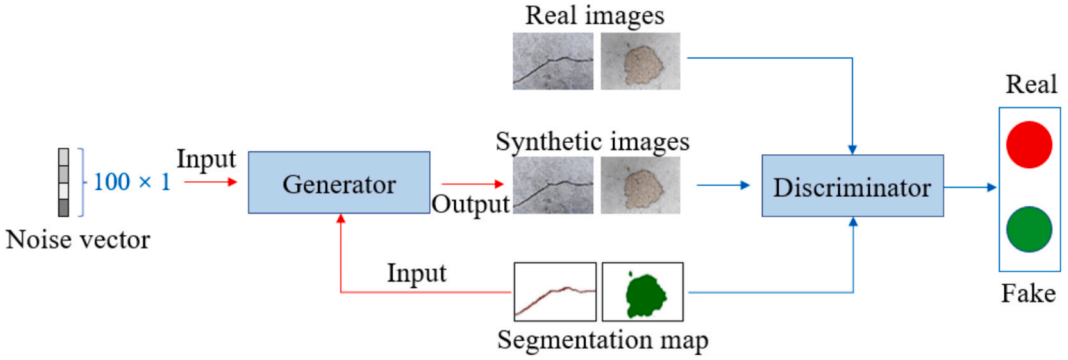


Fig. 8. Illustration of a CGAN architecture for data generation [36].

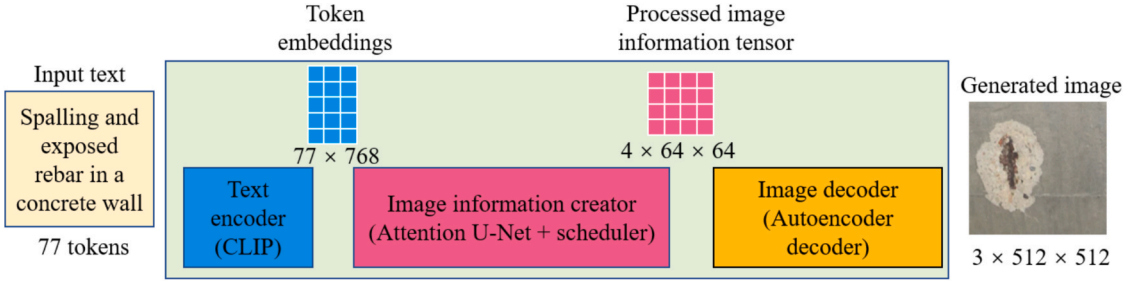


Fig. 9. Illustration of a Stable Diffusion architecture for text-to-image generation [90].

noise and a reverse process to denoise and reconstruct the image. The VAE maps images into latent space, while the U-Net handles denoising using skip connections, self-attention, and cross-attention, allowing for text-conditioned guided image generation. Fig. 9 illustrates the text-to-image generation process using a Stable Diffusion model [90]. The encoder extracts key information from the input text, which is transformed into latent space and decoded into the final image. Stable Diffusion holds significant potential for application in computer vision-based structural health monitoring. It can generate synthetic data to train AI models for anomaly detection, addressing challenges related to data scarcity. As reported in [90], the Stable Diffusion model was used to generate images of concrete surface damage, including cracks, spalling, and exposed rebar. This approach synthesized new damage images by pairing text and image data. To fine-tune Stable Diffusion, a training dataset of 678 images was assembled, and fine-tuning was performed using low-rank adaptation. As a result, a method for synthesizing highly diverse and high-quality concrete damage images was developed. In [91], crack images were generated by fine-tuning a Stable Diffusion model with text prompts. The synthetic images were subsequently used to train a crack detection neural network, achieving up to a 35.30 % improvement in F1 score and an average increase of 21.34 % compared to baseline methods. In conclusion, improving generalizability is crucial, as models trained on a single dataset often underperform when applied to different datasets or conditions. Stable Diffusion shows strong generalization when effectively prompted, enabling image generation across diverse textures and environments.

Despite their potential, several challenges limit the widespread adoption of text-to-image models. A major obstacle is that models like Stable Diffusion typically require paired image-text data for fine-tuning, which is often scarce or costly to obtain in specialized domains like structural health monitoring. Furthermore, generating high-resolution output demands substantial computational resources (e.g., GPU resources), posing difficulties for deployment in resource-constrained environments.

3.1.6. Performance metrics for dataset augmentation

The performance of data generation is typically assessed using the *IS* and *FID* [92]. The *IS* measures image quality and diversity, with higher scores indicating better performance. *FID* assesses the similarity to real images, where lower scores indicate better performance.

$$IS = expE_{x\sim p_G} KL(p(y|x)\|p(y)) \tag{1}$$

where G represents the generative model, D represents the Inception classifier, $p(y|x)$ represents the class distribution generated by a given input image x, $p(y)$ is the average class distribution of all input images, and KL represents the Kullback-Leibler divergence.

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

where μ_r and μ_g represent the mean values of real images and synthetic images, respectively; Σ_r and Σ_g represent the covariance matrices for the real images and the synthetic images, respectively.

3.1.7. Summary of dataset augmentation

Compared to traditional data augmentation methods based on image processing, data augmentation leveraging generative AI demonstrates superior performance in subsequent detection and segmentation tasks. Different generative AI models are applied based on specific scenarios, as summarized in Table 2: (1) Unsupervised data augmentation: When no labeled data is available, models like DCGAN can be utilized. However, these models have limitations in generating high-resolution images and often struggle to capture complex data distributions. (2) Unpaired dataset augmentation: When labeled data is scarce, unpaired dataset augmentation methods, such as CycleGAN, are promising for generating data without requiring paired samples. (3) Semi-supervised dataset augmentation: This approach is typically used in classification tasks involving large amounts of unlabeled data, leveraging both labeled and unlabeled data for better model performance (e.g., semi-supervised GAN). (4) Supervised dataset augmentation: While capable of generating high-quality images, this method requires paired datasets for training (e.g., Conditional GANs), which can be a limiting factor in certain applications. (5) Text-to-image: Stable Diffusion is the most used text-to-image model for generating images of concrete damage based on text prompts. It can produce diverse and highly customizable images. However, it is computationally intensive and requires text-image pairs for fine-tuning.

3.2. Image restoration

Fig. 10 illustrates various image restoration tasks using generative AI, including image denoising, image super-resolution, low-light enhancement, overexposure correction, and image deblurring. For image denoising, generative AI effectively removes random noise from the input, producing a cleaner and more detailed image [93]. In low-light enhancement, the model brightens underexposed images, revealing hidden details while preserving natural lighting conditions [94]. Overexposure correction addresses regions with excessive brightness, recovering lost details and restoring image balance. Image deblurring improves the sharpness of blurred images by reconstructing edges and fine details [94]. Furthermore, super-resolution upscales low-resolution images, enhancing details and improving clarity [95]. Each of these tasks highlights the capacity of generative AI to transform degraded or suboptimal images into visually enhanced outputs, specifically tailored to address distinct image restoration challenges. The restored images can be utilized for image-based infrastructure inspection.

3.2.1. Image denoising

Denoising is a supervised learning task that trains a model using noisy images as inputs and clean images as targets. GANs for denoising use adversarial training, where a generator creates clean images, and a discriminator distinguishes between real images and generated images. In [96], GANs were used to remove shadows, and their denoising capability significantly enhanced the quality of shadowed crack images and boosted U-Net segmentation accuracy (IoU) from 0.152 to 0.879. Diffusion models recently demonstrated superior performance in image

Table 2
Summary of data augmentation methods (DA: Data augmentation).

No.	Strategy	Applications	Pros	Cons	Models
1	Unsupervised DA	Generates samples from unlabeled data	No labeling cost, improves diversity	May produce unrealistic samples	DCGAN, WGAN-GP
2	Unpaired DA	Translate styles across unrelated datasets	Enables cross-domain augmentation	Require complex training, may introduce artifacts	CycleGAN
3	Semi-supervised DA	Combines few labeled + many unlabeled samples	Efficient use of limited labeled data	Depending on initial label quality	Semi-supervised GAN
4	Supervised DA	Handwriting mask to generate images	High label consistency	Less flexible, requires full labels	CGANs, DDPM
5	Text-to-image	Generates images from text prompts	Customizable outputs	Computationally expensive, needs text-image alignment	Stable Diffusion

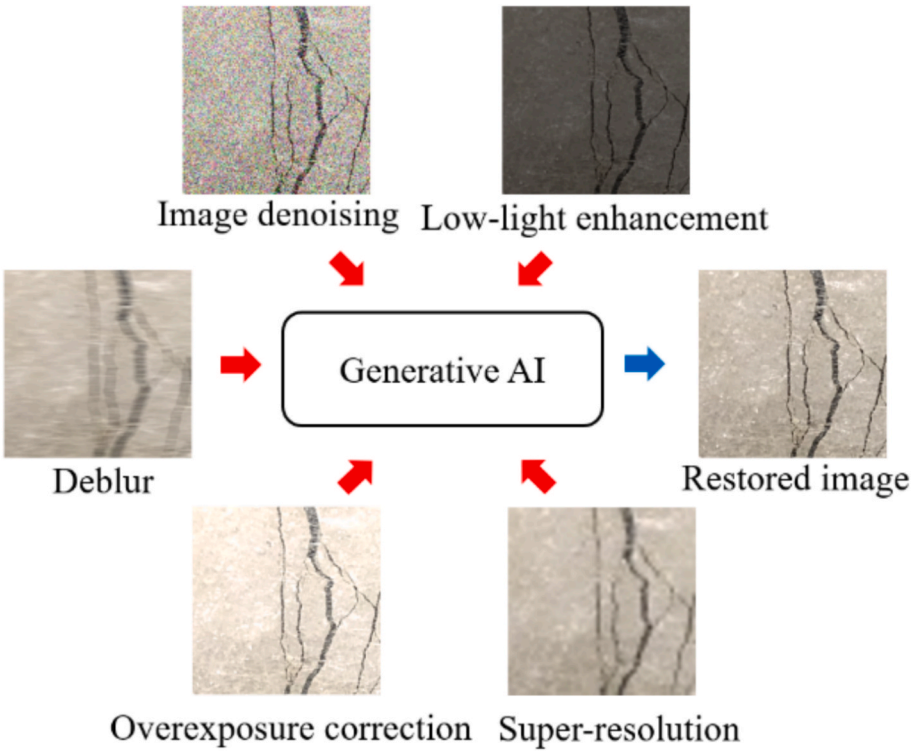


Fig. 10. Illustration of image restoration tasks using generative AI.

denoising by employing a Markov chain process to progressively add and reverse noise, ensuring greater stability in handling large data variations [97]. Specifically, diffusion models have been applied to crack inpainting, automatically restoring missing crack information and preserving detailed features even in high-noise environments [98]. This approach outperforms traditional inpainting methods like PatchMatch, Contextual Attention, and Repaint, achieving PSNR improvements of 20.5 %, 13.4 %, and 4.1 %, respectively. Despite advancements in GANs and diffusion models, the application of image denoising techniques in infrastructure maintenance remains underexplored. To adapt these methods for noisy field conditions, domain-specific validation could be performed using annotated datasets captured under shadowed environments or other noisy conditions. Furthermore, combining synthetic noise augmentation (e.g., Gaussian noise and random shadows) with real-world samples can enhance model robustness and generalizability in practical scenarios.

3.2.2. *Low-light enhancement*

Generative AI enhances visibility and quality in images taken under low-light conditions. The goal of such generative AI models is to improve brightness, contrast, and detail in low-light images while preserving natural colors and reducing noise. In [94], a conditional generative model is developed to enhance the illumination of concrete crack images. The conditional generative model incorporates a self-attention layer in the skip connections and utilizes ResNet as the foundational block, while also gradient penalty loss. Evaluation results demonstrate that this approach surpasses state-of-the-art methods, achieving a SSIM of 0.95 and a PSNR of 31.4. The enhanced images were subsequently used to train a deep learning model for crack identification and localization in concrete images. The performance metrics for crack segmentation, with an IoU exceeding 0.98 and an F1 score surpassing 0.99, demonstrate exceptionally high accuracy in crack assessment tasks. In [99], a model called N-LoLiGan is introduced. This model utilizes feature loss to guide training, ensuring the preservation of image textures. Experimental results show that the developed model excels in enhancing low-light images, greatly improving the visibility and clarity of crack features. Moreover, when these enhanced images were used to train object detection models like YOLACT and YOLOv5s, the average detection accuracy increased from 0.63 to 0.89 and from 0.90 to 0.97, respectively, demonstrating a substantial improvement in detection performance.

3.2.3. *Overexposed correction*

Overexposure correction using generative AI restores details in images that are excessively bright or washed out [100]. The overexposure correction is also based on conditional generative models. The generator reconstructs lost details in overexposed areas by using context from better-exposed regions, adjusting luminance and contrast for a natural appearance. The discriminator compares these adjustments to real properly exposed images and provides feedback to help the generator improve over time. This iterative process enhances the overall quality of the images by gradually refining the correction. In [94], a generative adversarial network is used to restore overexposed images with cracks. The segmentation accuracy on overexposed images was initially measured with an IoU of 0.934 and an F1 score of 0.941. The accuracy improved significantly with the restored images, achieving an IoU of 0.989 and an F1 score of 0.994.

3.2.4. *Image deblurring*

In image processing, removing motion blur is a critical but challenging task, often caused by factors like rapid object movement or camera shake [101]. Conditional generative models have been proposed to address this issue through effective deblurring. In [102], Motion blur from robotic car imaging of pavement cracks posed a major challenge, reducing deep learning model accuracy. To address this, DeblurGAN emerged as a promising solution for deblurring pavement crack images. The model was trained using original images and artificially blurred images generated by a random trajectory blurring algorithm [103]. Deblurred images tested with DeepLabV3 + improved IoU from 0.18 to 0.43 and F1 score from 0.26 to 0.58. Image deblurring significantly improved crack detection on building façades [24]. For instance, deblurring raised segmentation accuracy from an IoU of 0.895 and F1

score of 0.891 to 0.982 and 0.991, respectively, highlighting its effectiveness in restoring blurred images [94].

3.2.5. Super-resolution reconstruction

Low-resolution images often miss fine cracks and subtle defects crucial for evaluating structural integrity. SRGAN presents a promising solution for addressing this challenge [83]. This model can generate high-resolution images from low-resolution inputs, significantly enhancing image details. Super-resolution can be applied to enhance the resolution of low-resolution crack images, improving image details and quality [21]. The reconstructed images achieved a PSNR of 33.1 dB and SSIM of 0.820, surpassing Bicubic (30.24 dB, 0.775) in quality. When trained on the super-resolution dataset, segmentation models achieved a 15.6 % higher F1-score and a 23.8 % improvement in IoU compared to models trained on Bicubic-reconstructed images [21]. Further enhancements were made by incorporating a self-attention mechanism into the model, resulting in a PSNR of 27.5 dB and an SSIM of 0.865 [38]. This surpassed both Bicubic interpolation (PSNR: 21.7 dB, SSIM: 0.663) and the original framework (PSNR: 26.9 dB, SSIM: 0.847). For crack classification tasks, ResNet50 trained on these high-resolution images achieved 98.2 % accuracy [104]. In conclusion, adversarial network-based super-resolution processing has substantially improved the accuracy and reliability of crack detection and segmentation [21]. The illustration of the super-resolution reconstruction using generative AI model is illustrated in Fig. 11.

3.2.6. Performance metrics for image restoration

The performance of image reconstruction, including super-resolution and image restoration, is typically assessed using PSNR and SSIM metrics. The PSNR, defined in Eq. (3), was used to evaluate image quality [15]. PSNR quantifies the logarithmic value of the MSE between the original and reconstructed images, relative to the maximum possible pixel value. A higher PSNR indicates better image quality and lower distortion.

$$PSNR = 20\log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \tag{3}$$

where MAX_I is the maximum pixel value of the original image; and MSE is defined in Eq. (4):

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(f(i,j) - g(i,j))^2 \tag{4}$$

where f is the matrix data of the original image; g is the matrix data of the low-quality image; m is the number of rows of pixels with i as the row index; and n is the number of columns of pixels with j as the column index.

The SSIM, defined in Eq. (5), is a metric used to assess the similarity between two images [15]. SSIM evaluates image quality by considering three key components: luminance, contrast, and structure. The SSIM values range from 0 to 1, with a value of 1 indicating the highest possible image quality.

$$SSIM(x,y) = \frac{\left(2\mu_x\mu_{xy} + C_1\right)\left(2\sigma_{xy} + C_2\right)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

where x is the matrix data from a window in the target image; y is the matrix data from a window in the reference image; C_1 and C_2 are small constants introduced to avoid division by zero, with $C_1 = 0.0001$ and $C_2 = 0.0009$; μ_x and μ_y are the mean values of x and y, respectively; σ_x and σ_y are the variances of x and y, respectively; and σ_{xy} is the covariance between x and y.

The SSIM was calculated using a sliding window approach. In each calculation, a window of size $N \times N$ was taken from the target and reference images, and the SSIM index was computed based on the window. Small window size (e.g., 5×5) captures fine details and local variations, and large window size (e.g., 11×11) captures broader information.

3.2.7. Summary of image restoration

Image restoration includes tasks like denoising, low-light enhancement, overexposure correction, deblurring, and super-resolution. Table 3 summarizes the case studies on image restoration applications in structural health monitoring. Super-resolution is the most common task, primarily applied to crack detection and segmentation. Low-light enhancement and deblurring show significant performance gains (e.g., 63.0 % to 97.0 % for YOLOv5s [99]). While crack detection dominates the research focus, other critical damage types including spalling and corrosion have received significantly less attention. The effectiveness of image restoration depends critically on the availability of paired datasets (input–output pairs) for training. However, such datasets are often difficult to obtain because data collection is often time-consuming, especially for real-world image degradations such as noise, motion blur, and lighting variations. These characteristics pose significant challenges for creating comprehensive datasets that encompass all possible scenarios, ultimately restricting model generalizability to unseen conditions. To address these limitations, researchers have employed synthetic data augmentation techniques. These include adjusting pixel values to simulate overexposure or low-light conditions [94], as well as applying convolution kernels to reproduce focus issues or motion blur effects [102]. While these methods can increase dataset size, ensuring adequate diversity in synthetic datasets.

3.3. Image segmentation

Image segmentation in structural health monitoring, such as damage

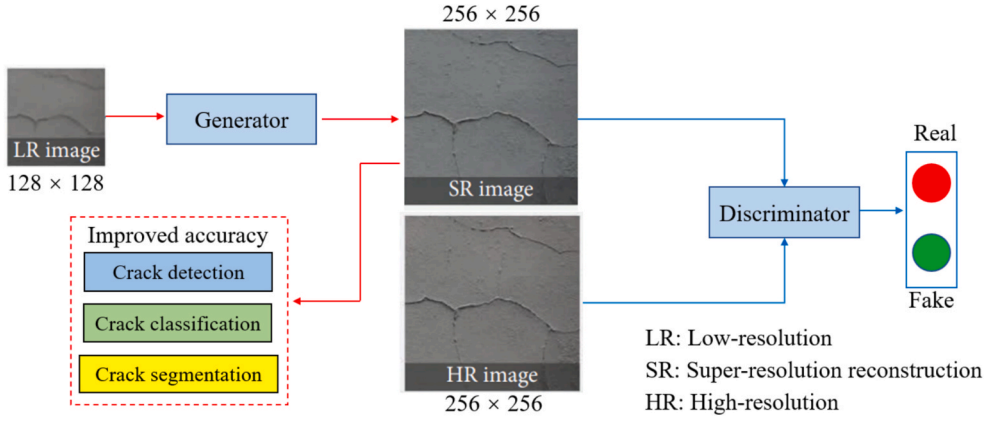


Fig. 11. Illustration of a SRGAN architecture for image reconstruction [105].

Table 3
Summary of generative-AI models for image restoration.

No.	Year	GAN model	Tasks	Application	Deep learning	Metrics (%)	Ref.
1	2020	SRGAN	Super-resolution	Building crack classification	ResNet-50	Up to 98.2	[104]
2	2020	CGAN	Image deblurring	Building crack	–	–	[24]
6	2022	ESRGAN	Super-resolution	Building crack segmentation	CDU-Net	69.5 to 82.3	[21]
4	2022	SRGAN	Super-resolution	Pavement crack detection	Faster-RCNN	76.7 to 87.9	[38]
5	2022	Improved SRGAN	Super-resolution	Pavement crack detection	Faster-RCNN	76.7 to 91.2	[38]
6	2023	Diffusion model	Image denoising	Pavement crack	–	–	[98]
7	2023	SRGAN	Super-resolution	Building crack segmentation	FDDWNet	75.8 to 77.2	[105]
8	2023	N-LoLiGan	Low-light enhancement	Building crack detection	YOLACT	63.0 to 89.0	[99]
9	2023	N-LoLiGan	Low-light enhancement	Building crack detection	YOLOv5s	90.0 to 97.0	[99]
10	2023	CGAN	Image deblurring	Pavement crack segmentation	MIMO-UNet	31.6 to 60.3	[102]
11	2024	ARCGAN	Low-light enhancement	Building crack segmentation	LinkNet	15.3 to 99.4	[94]
12	2024	ARCGAN	Image deblurring	Building crack segmentation	LinkNet	89.5 to 98.2	[94]
13	2024	CGAN	Image deblurring	Building crack segmentation	UperNet	83.4 to 86.8	[101]
14	2024	ARCGAN	Overexposed correction	Building crack segmentation	LinkNet	93.4 to 98.9	[94]
15	2025	IEnlightenGAN	Image denoising	Building crack segmentation	U-Net	39.9 to 98.6	[96]
16	2025	SRGAN	Super-resolution	Pavement crack segmentation	DeepLabV3+	85.0 to 90.0	[106]

detection, aims to identify and isolate damaged regions within images for accurate analysis [107]. This process often relies on deep learning models to classify each pixel as either belonging to a damage or the background [108]. GAN-based image segmentation has demonstrated superior performance compared to CNN-based models like U-Net and DeepLabV3+, particularly in addressing the challenge of limited labeled training datasets. By leveraging adversarial learning, GANs can achieve high-quality segmentation results even with smaller datasets [108,109]. In contrast, traditional models such as CNN-based segmentation model often require labor-intensive manual segmentation during the data preparation phase, making GANs a more efficient and effective alternative. A GAN for image segmentation requires a paired dataset with each input image matched to its corresponding segmentation mask, as illustrated in Fig. 12. The generator learns to produce segmentation masks from input images, while the discriminator assesses the realism of the generated masks. The training process combines adversarial loss from the discriminator and pixel-wise loss (such as L1 loss) to guide the generator toward producing accurate and realistic segmentation outputs. This approach enables the direct application of generative models for image segmentation. Image segmentation performance is typically evaluated using standard metrics such as F1 score and IoU.

In [109], CrackSegAN was developed for crack segmentation, utilizing a U-Net-based generator and a discriminator. The generator creates a binary crack map from a given crack image, while the discriminator compares the ground-truth masked image with the predicted masked image using a multi-scale L1 loss. A joint loss function combining multi-scale L1 loss and Dice loss was introduced for addressing the significant class imbalance in pavement crack images. The generator aimed to minimize the joint loss while the discriminator focused on maximizing the multi-scale L1 loss. Through this process, both the generator and discriminator improved, eventually reaching an equilibrium where the generator produced crack maps indistinguishable from the ground truth. The developed model achieved a F1 score of 0.978, surpassing models like CU-GAN, pix2pix, and DeepCrack. In [110], CrackGAN was developed for direct crack segmentation. The generator used asymmetric U-Net architecture to produce binary images while the discriminator improved the performance of the model. The developed model tested on the CrackForest dataset achieved an F1 score of 0.919 and surpassed other models such as FCN. Compared to traditional segmentation networks such as FCN, CrackGAN incorporated a discriminator that evaluated the generated crack masks, helping to produce more realistic and continuous results with better detail preservation, particularly effective for detecting thin and narrow cracks. In [39], a diffusion model was developed for pavement crack segmentation and trained on 1,037 images with corresponding binary labels. The model achieved the highest IoU score of 0.841, outperforming benchmark models such as U-Net, DeepLabV3+, and SegFormer. Diffusion models outperform conventional deep learning models in challenging scenarios with complex backgrounds (e.g., shadows) and discontinuous annotations. This advantage stems from the architectural limitations of models like U-Net or SegFormer, which rely on local receptive fields, either through convolution kernels or window-based attention. These models often struggle to capture long-range dependencies, leading to fragmented crack detections due to limited contextual understanding. In contrast, diffusion models reconstruct clean segmentation masks from noise by learning the joint probability distribution of all pixels. This enables them to model global crack continuity more effectively. Unlike CNN-based models that process image patches independently, diffusion models iteratively refine predictions while maintaining structural coherence, allowing them to infer and fill in missing crack segments based on the learned data distribution [39]. The comparison between generative AI models with conventional semantic segmentation models is summarized in Table 4.

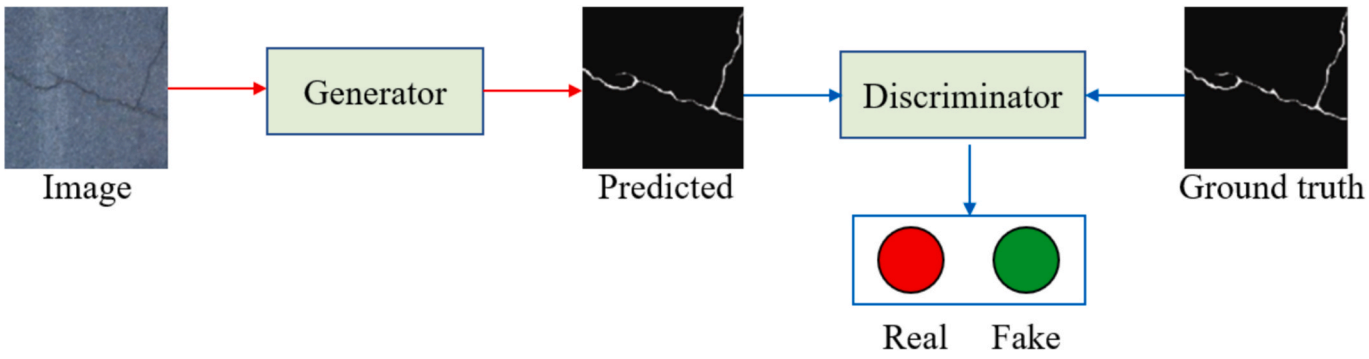


Fig. 12. Illustration of a GAN architecture for image segmentation.

Table 4
Comparison between generative AI models with semantic segmentation models.

Models	Accuracy	Category	Note	Ref
CrackGAN	91.9 %	GAN-based	GANs outperform CNN-based segmentation models, as the discriminator enhances the realism, continuity, and detail of generated crack masks, particularly for thin and narrow cracks.	[110]
FCN	89.0 %	CNN-based		
Crackforest	87.7 %	CNN-based		
CrackSegAN	84.1 %	GAN-based		[109]
DeepCrack	82.5 %	CNN-based		
RoadPainter	71.8 %	Diffusion model	Diffusion models outperform CNNs and Transformers in complex scenes with shadows or annotation gaps by modeling global pixel distributions. They capture long-range dependencies and reconstruct coherent crack masks, inferring missing segments and correcting errors.	[107]
LinkNet	59.5 %	CNN-based		
PSPNet	55.5 %	CNN-based		
CrackDiff	84.1 %	Diffusion model		[39]
SegFormer	83.3 %	Transformer-based		
DeepLabV3+	83.4 %	CNN-based		

3.4. Multi-modal data integration

3.4.1. Visual question answering

The development of VQA systems has been advanced through multi-modal models like CLIP, which align visual and textual data in a shared feature space [27,111]. CLIP's architecture includes separate encoders for images and text, which are trained using ITC learning to align features effectively, as illustrated in Fig. 13**(a)**. In [27], the CLIP model was trained on 610,197 training samples, 130,122 validation samples, and 131,774 test samples. The visual encoder options include ResNet and Vision Transformers, while the text encoder options include LSTM and BERT. The dataset contains 62 possible answer candidates, covering bridge member types, and damage classifications. Questions in the dataset are categorized into three types: (1) Yes/No Questions (e.g., "Is there corrosion?"). (2) Member-Class Questions (e.g., "What is the component in the image?"). (3) Damage-Class Questions (e.g., "What type of damage is present?"). Among the tested configurations, the combination of a Vision Transformer and BERT achieved the highest accuracy: 99.4 % on binary tasks, 83.0 % on member identification, and 77.6 % on damage identification. Fig. 13**(b)** illustrates the results of the Q&A system, where the input consists of an image and a corresponding question. The system processed the image through a visual encoder and the question through a text encoder, fusing the extracted features to generate an accurate answer. This output reflected the ability of the system to effectively align visual and textual modalities to address the query based on the input image. A similar study was conducted in [112], introducing BridgeCLIP, an innovative framework that adapts the pre-trained vision-language model CLIP for automatic bridge inspection through multi-label image classification. In the BridgeCLIP framework, the training phase uses both images and textual descriptions to help the model learn domain-specific knowledge. During prediction, only the image is provided as input. The model has learned from the text during training. It can identify different damage types in the image without any extra text input. In [28], the study proposed leveraging a large language model to develop a VQA system aimed at enhancing human-robot collaboration in UAV-assisted bridge inspections. The system outputs a classification result that determines whether a specified object or damage type is present in the image. With a peak accuracy of 83.33 %, though lower than that of well-trained segmentation models such as ResNet (89.4 %), the approach still demonstrates strong potential for enhancing the precision and safety of UAV-based bridge inspections. This highlights the value of vision-language models in addressing specialized, domain-specific tasks where human-robot collaboration and semantic understanding are essential. In [113], a study proposed a deep learning-based framework for estimating the causes of bridge damage using VQA. The dataset includes 22 distinct types of damage. A domain-specific VQA model was developed and trained on over 440,000 bridge images, enabling it to answer questions related to types of damage. The model demonstrated high accuracy (99.1 %) in damage classification

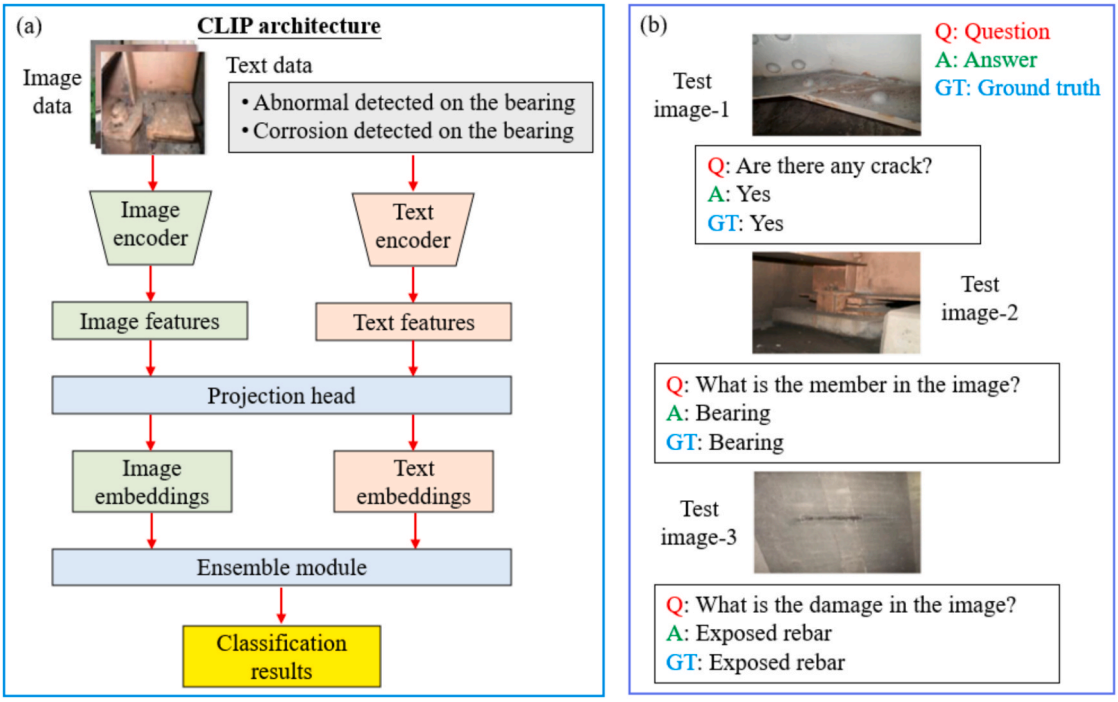


Fig. 13. Illustration of the VQA system [27]: (a) CLIP model architecture. CLIP employs separate vision and text encoders to map image and text inputs into a shared embedding space. The model then selects the text label with the highest similarity score, thereby enabling image classification; (b) Q&A results. The predicted class is integrated into the question-answering program, where the answer corresponds to the classification result produced by CLIP.

based on yes/no questions. In [114], the study explored the feasibility of using VQA for post-disaster damage detection based on aerial footage captured by UAVs. A VQA model combining a CNN with a Bag of Words was proposed to enable image-based question answering. The model was evaluated on a custom dataset collected after the hurricane, achieving an overall accuracy exceeding 92 %. In [115], the study presented the development of a web-based bridge inspection system that automatically generates explanatory texts describing bridge damage from inspection images using a deep learning-based image captioning model. By combining a CNN for visual feature extraction with an LSTM for text generation, the system outputs clear descriptions such as "cracking detected on the bottom of the slab," making inspection results more accessible to engineers. The system can continuously improve its accuracy through user feedback. This approach bridges the gap between visual analysis and textual interpretation in structural health monitoring and offers practical potential for automating field inspections and report generation.

Despite their potential, several challenges hinder the broad adoption of VQA. A major limitation is the substantial amount of annotated training data required, which is particularly burdensome in resource-limited settings where labeling domain-specific engineering imagery is both costly and time-consuming. Technically, the reliance on large-scale architectures like Vision Transformers and BERT introduces significant computational demands, making real-time deployment on edge devices such as UAVs or mobile robots difficult. While a few-shot learning can alleviate data constraints, it often compromises robustness and accuracy, limiting its suitability for fully autonomous applications without human supervision.

3.4.2. Large language models

In computer vision-based structural health monitoring, integrating LLMs can enhance human–machine interaction and data interpretation, offering a promising approach to better understand and manage complex visual inspection data. In [28], a cascaded crack detection strategy using multi-modal LLMs was proposed to enhance zero-shot fatigue crack detection in steel bridges. The study systematically evaluated five LLMs: Claude, GPT-4o, GPT-4o mini, Grok, and Gemini. The input to the crack detection system based on LLMs consisted of images of steel bridge components that may contain fatigue cracks, along with textual prompts (such as instructions or damage labels) to guide the detection process. The output is a classification result for each image or image patch, indicating the presence or absence of fatigue cracks. Among the evaluated models, GPT-4o mini consistently achieved the best performance in image-level crack classification tasks. The advantage of using LLMs lies in its ability to perform zero-shot learning, enabling direct predictions without the need for task-specific training. However, its prediction accuracy is generally lower compared to dedicated classification models (e.g., ResNet). If resources are available, fine-tuning an open-source multi-modal LLMs (e.g., Llama-3-8B) with domain-specific data is recommended. Even a small set of relevant images can significantly enhance prediction accuracy [116]. Fig. 14 illustrates the application of LLMs in damage inspection.

In [117], the study presented SDA-Chat, a novel multi-modal LLM-based framework designed for rapid post-earthquake structural damage assessment. By integrating visual encoders, a query transformer, and LLMs (e.g., LLaMA3), the system can interpret structural damage images and respond with professional textual evaluations. SDA-Chat was trained on 8,195 annotated image–text pairs and supports seven structural assessment tasks, such as damage type, collapse level, and material classification. Experimental results demonstrate that SDA-Chat achieves an accuracy of up to 76.11 % with an inference speed of 435 tokens per second. A recent study explored the use of LLMs, such as ChatGPT, to enhance automated post-disaster building damage assessment from ground-level images [118]. To integrate visual and textual information, a vision-language model called CLIP is used, which employs a dual-encoder architecture to separately encode images and text before fusing them in a shared embedding space through late fusion. Experiments on a curated dataset of hurricane-affected buildings show that combining image and LLM-generated captions improves classification performance by approximately 4 % for using image alone and 17 % for using text alone. These results indicate that the generated textual descriptions provide complementary information beyond the visual content, even though the model was not explicitly trained for damage assessment tasks. SDIGLM is a novel LLMs for structural damage identification, built on the VisualGLM-6B architecture [119]. It integrates a U-Net-based semantic segmentation module with a multi-modal Chain-of-Thought reasoning framework to deliver both precise classification and detailed natural language descriptions of damage types such as cracks, holes, and corrosion. Trained on a curated dataset of 11,722 image-text pairs and fine-tuned with LoRA, SDIGLM surpasses general-purpose models like GPT-4o and GLM-4v, achieving 95.24 % accuracy across diverse structural scenarios.

Another potential application of LLMs is the enhancement of contextual object detection capabilities. Unlike conventional approaches such as YOLO, which rely solely on visual features to identify objects, LLM-enhanced methods incorporate contextual cues, such as surrounding objects and scene settings into the detection process. A representative example can be found in [120]. In the context of structural health monitoring, this capability allows for precise detection of critical structural features such as cracks, joints, and deformations. Inspectors can issue natural language queries like "highlight cracks" or "detect corrosion," and the system responds with accurate and intuitive results.

3.5. Discussions

3.5.1. Deployability

Table 5 compares generative AI models across two critical deployability dimensions, including real-time inference capability and field deployment readiness. Real-time inference remains a key limitation, as

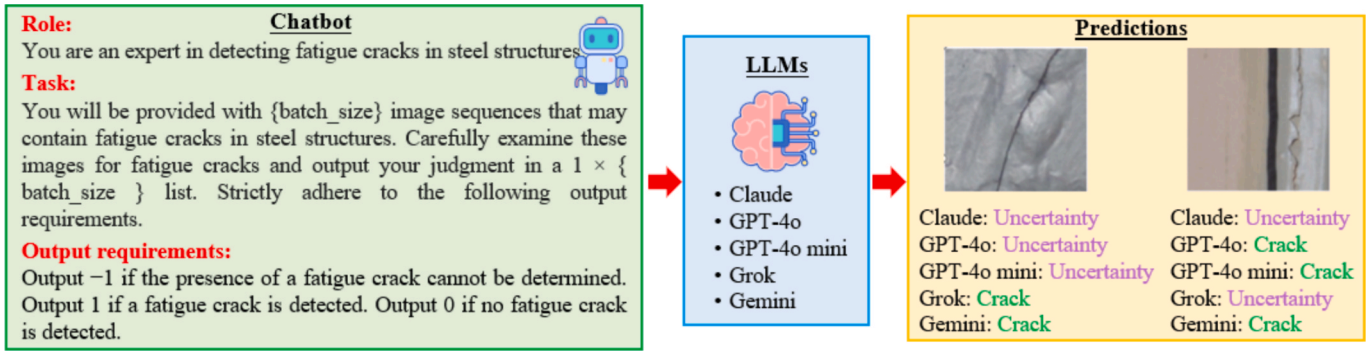


Fig. 14. Illustration of zero-shot crack identification using LLMs [28]. Note: −1 represents "Uncertainty", 1 represents "crack", and 0 represents "no-crack".

Table 5
Deployability of generative AI models.

Models		Model size	Real-time inference	Field deployment
DCGANs [126]		6–10 M	Yes	Yes
Conditional GANs [102]		~105.3 M	No	Yes
Diffusion models [39]		387 M	No	Yes
Stable diffusion [127]		~ 860 M	No	Yes
CLIP [128]		428 M	No	Yes
LLMs	GPT-4o [129]	~200B	No	No
	GPT-4o mini [130]	8B	No	No
	Claude 3.5 [130]	~175B	No	No

none of the listed models support it, which is a critical requirement for applications like UAV-based inspection and robotics. Meeting real-time constraints typically requires lightweight architecture such as TinyYOLO (< 20 M parameters) [121]. While DCGAN can operate in real time, speed is generally less critical for image generation and augmentation, where image quality and realism are prioritized. Smaller models under 500 M parameters, such as conditional generators and CLIP, are more suitable for field deployment on edge devices but still require latency optimization. Field deployment refers to models capable of running directly on edge devices at the data source, without relying on external computing resources such as cloud servers or transmitting the data to external systems for processing. In contrast, larger models, particularly LLMs with over 8 billion parameters, are designed for high-performance computing and are impractical for real-time or embedded applications. For instance, fine-tuning an 8B model with parameter-efficient methods like LoRA still requires at least one GPU with 12 GB of memory [122], and training times can reach 14 h [123]. Even inference at FP16 precision requires approximately 12 GB of memory [122]. As a result, deployment on resource-constrained platforms like UAVs equipped with Raspberry Pi or Jetson Nano remains unfeasible [124]. Model distillation offers a promising solution by compressing large models into smaller ones (for example, under 1 billion parameters), substantially reducing memory and compute requirements [125]. However, this often comes at the cost of reduced accuracy and warrants further research. In summary, small models that can run on edge devices are suitable for field deployment, although some may not achieve real-time inference. In contrast, large models are typically utilized by uploading data to external computational resources, which limits their ability to field deployment and support real-time inference.

3.5.2. Generalizability

Although generative AI models can generate additional images and expand the training set, thereby improving the prediction accuracy of downstream tasks such as damage detection and segmentation, this improvement is often limited to the specific task or domain [131]. The transferability of such models to other areas remains underexplored. For example, a model trained on an augmented concrete crack dataset may exhibit reduced accuracy when applied to asphalt pavement crack identification. Furthermore, overfitting synthetic characteristics such as uniform lighting, clean backgrounds, and idealized crack patterns can significantly degrade model performance in real-world. To address this limitation, several strategies can be conducted: (1) cross-dataset evaluation has been proposed as a robust validation strategy. For instance, a model trained on synthetic concrete crack images can be validated on real-world datasets such as pavement cracks or steel bridge cracks. This approach provides a more reliable measure of generalization performance under varying conditions [132]. (2) Other strategies include domain randomization techniques, which introduce variations in image properties such as lighting conditions and noise. This helps the model focus on task-relevant features rather than relying on domain-specific artifacts [15]. Without these strategies, models may excel on synthetic benchmarks but fail in real-world settings with surface variation, shadows, occlusion, and debris. (3) Transfer learning can be used to adapt pre-trained models to new domains by fine-tuning them on a small set of labeled real-world data. This approach has demonstrated potential in enhancing model generalization across various material types and surface conditions [133].

3.5.3. Practical application of generative AI

Generative AI has shifted from theory to a practical tool with tangible benefits for infrastructure maintenance. One of the most impactful applications lies in automated data augmentation. Civil infrastructure systems, such as pavements, bridges, and build structures often suffer from sparse and imbalanced datasets. Generative AI can produce high-resolution synthetic images of surface damage, including cracks, potholes, spalling, and exposed rebars, allowing deep learning models to be trained with greater variety and volume. Generative models such as DCGAN, WGAN-GP, and StyleGAN effectively address this bottleneck by synthesizing diverse and realistic defect images, enriching datasets and improving model generalization [18–20]. For example, models trained with synthetic data have shown higher accuracy in crack detection tasks. The impact spans various model types, including VGG16, YOLOv4, and U-Net [16,18,71]. Additionally, the ability to simulate underrepresented defect types allows for more balanced training data, which helps reduce model bias and enhances detection robustness. In addition to data augmentation, generative AI models are also increasingly used for image restoration. Damage inspection images are often affected by noise, blurriness, low resolution, or poor lighting conditions, especially when captured by UAVs. GAN-based restoration models can reconstruct missing or corrupted regions, enhance resolution, and improve overall image quality. This helps preserve critical structural details, ensuring more accurate downstream analysis such as defect detection or segmentation. Generative AI also plays a valuable role in semantic segmentation of cracks by enabling style translation from RGB images to binary masks. This capability is particularly useful in situations where pixel-level annotations are scarce or labor-intensive to produce. Models like pix2pix can generate accurate segmentation masks from raw RGB inputs. Recent advances in multi-modal AI, especially vision-language models like CLIP and LLMs, offer practical tools for structural health monitoring. These models align visual data with semantic understanding, enabling tasks such as question answering, component identification, and defect classification. CLIP-based systems have achieved high accuracy in identifying bridge damage, while frameworks like Bridge-CLIP allow image-only classification after training with text-image pairs. In UAV-assisted structural health monitoring, few-shot CLIP models detect defects with minimal labeled data, enabling semi-autonomous inspection. LLMs like GPT-4o mini further support zero-shot crack detection using multi-modal inputs. Together, these tools reduce labeling demands, speed up assessment, and enhance collaboration between AI systems and human inspectors. Based on the above discussion, it is feasible to build an automated pipeline that integrates robotic-based data collection with generative AI-driven dataset augmentation and image restoration. This would enable the creation of a high-quality data platform to support damage inspection through segmentation. Furthermore, multi-modal generative AI models can align visual data with textual information, enabling more comprehensive interpretation of structural damage and supporting intelligent interfaces such as chatbots for human-AI interaction. The proposed pipeline enables practical, real-world implementation of generative AI models in structural health monitoring by integrating robotic data collection, synthetic data generation, image restoration, semantic analysis, and human-AI interaction into a unified system, as shown in Fig. 15.

3.5.4. Paradigm shift enabled by generative AI

The rise of generative AI is driving a transformative shift in structural health monitoring, not only by enhancing performance (e.g., higher IoU scores), but also by fundamentally reshaping methodologies and workflows. Traditionally, computer vision-based structural health

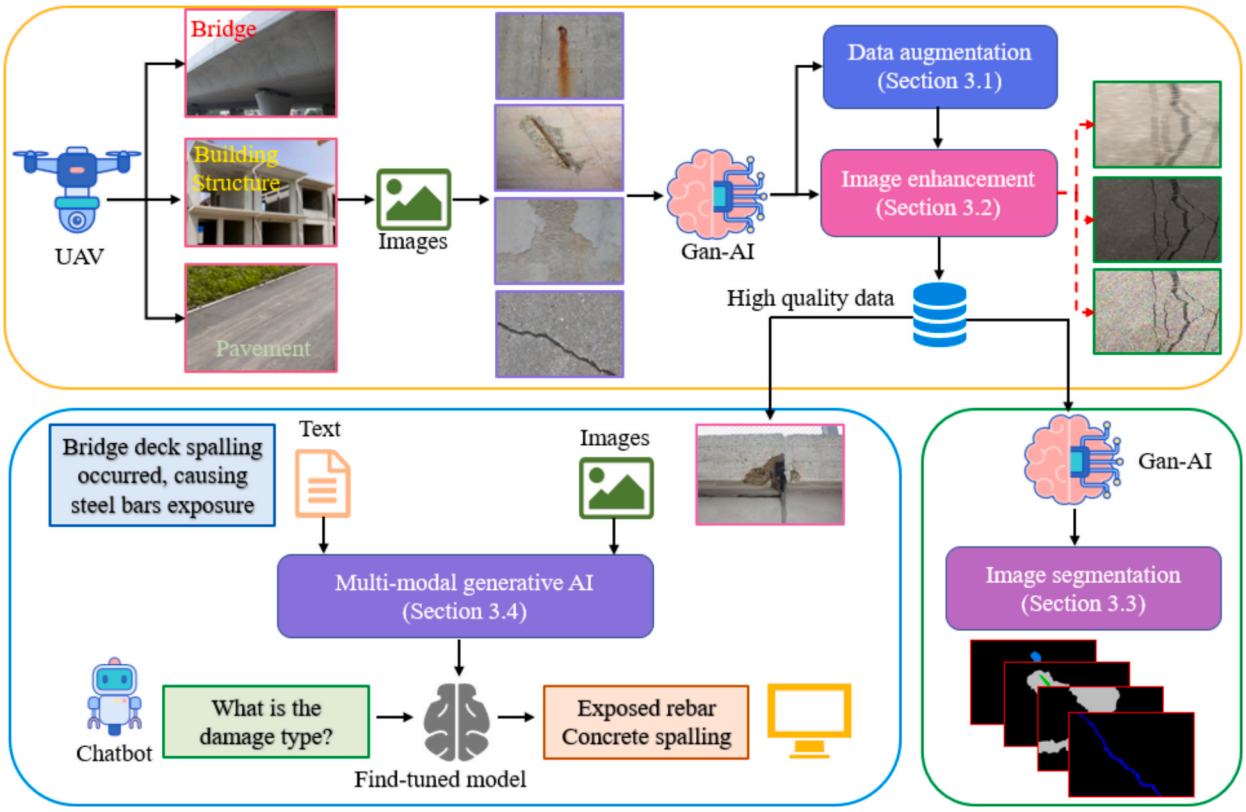


Fig. 15. Practical implementation of generative AI models in structural health monitoring.

monitoring has depended on manually curated datasets, expert-designed feature extraction, and offline post-processing. In contrast, generative AI enables automated and interactive workflows that redefine the processes of detecting, analyzing, and interpreting structural damage. For instance, the generation of diverse and high-resolution synthetic damage images reduces reliance on costly and time-consuming field data collection, accelerating model development. This shift moves structural health monitoring from a data-scarce to a data-abundant paradigm. Advanced generative AI models, such as Stable Diffusion, offer promising capabilities for precise control over image content and background. Although their use in structural health monitoring is still emerging, they hold significant potential for future research. These models facilitate accurate and controllable generation of structural damage, aligned with specific expectations or predefined conditions. Furthermore, multi-modal models such as CLIP and SDIGLM enable intuitive interaction between human inspectors and AI systems through natural language, allowing users to ask questions, receive insights, and generate inspection reports directly from visual data. By integrating visual-language models with LLMs, generative AI facilitates the development of more interactive, interpretable, and collaborative systems. These models go beyond traditional classification and segmentation tasks to support advanced capabilities such as visual question answering, automated damage description, and AI-assisted decision-making. Such features not only improve user experience but also lower the technical barrier for field inspectors, enabling non-experts to engage with complex AI systems through simple prompts or queries. A critical avenue for future research lies in the development of lightweight and computationally efficient models that can be seamlessly deployed on edge devices. Achieving real-time inference under resource-constrained environments is imperative for enabling autonomous, on-site structural damage assessment and effective human–computer interaction.

4. Conclusions

This study highlights the transformative role of generative AI in advancing infrastructure maintenance by addressing critical challenges such as data scarcity and quality issues. By leveraging models like GANs, Diffusion models, and multi-modal AI architectures, generative AI enhances data augmentation, image restoration, damage identification capabilities, significantly improving the accuracy and reliability of infrastructure inspection. Generative AI is driving a transformative shift in structural health monitoring by enabling data-rich, interactive, and intelligent workflows. To sustain this transformation, future research should prioritize lightweight, deployable models capable of real-time performance on edge devices for practical, on-site applications. Based on the above investigation, the following conclusions can be drawn:

- As discussion in Section 3.1, data augmentation is a critical technique for improving the performance of AI models, especially in damage inspection tasks where datasets are limited or imbalanced. Generative AI models expand datasets, address class imbalance, and enhance accuracy and robustness in damage inspection. The generation of diverse, high-resolution synthetic damage images reduces dependence on costly and time-consuming field data collection, thereby accelerating model development and driving structural health monitoring toward a data-rich paradigm.
- Image restoration with conditional generative models improves degraded images, for more effective damage detection, as discussed in Section 3.2. Techniques such as denoising, super-resolution, low-light enhancement, overexposure correction, and deblurring address common issues like noise, poor resolution, and lighting inconsistencies. By restoring fine structural details and improving overall image clarity, these models significantly enhance the effectiveness of downstream detection and segmentation tasks, supporting robust analysis even under adverse conditions.

- Generative AI (e.g., GANs and diffusion models) achieves higher accuracy in image segmentation compared with CNN-based model (e.g., DeepLabV3 +) and transformer-based model (e.g., SegFormer), as discussed in Table 4. This improvement is largely attributed to their ability to model global context and preserve fine structural details, especially in challenging conditions such as noise, shadows, or fragmented annotations.
- The integration of LLMs and multi-modal systems enables more comprehensive defect identification (Section 3.4). By jointly processing visual and textual inputs, these models enable contextual reasoning, generate natural language explanations, and support interactive human-AI communication. This advancement improves the transparency and reliability of inspection results and contributes to a broader transformation of structural health monitoring workflows into interactive and explainable systems.

Despite advancements in generative AI models, remaining challenges and future research directions are summarized below:

- Infrastructure inspection is a specialized task requiring tailored datasets to train generative models effectively. Collecting defect images under varied conditions is time-consuming and resource-intensive, limiting the scalability of generative AI approaches in structural health monitoring.
- Research on advanced generative AI models such as Stable Diffusion, LLMs, and multi-modal systems remains limited due to high computational costs. Increased investment is crucial, as these models offer superior efficiency and adaptability across varied scenarios.
- Advanced generative AI models can be fine-tuned for specific tasks, allowing them to adapt to diverse contexts with minimal additional training. This capability represents a key direction for future research in this field. Fine-tuning reduces the need for extensive retraining, making these models more versatile for infrastructure maintenance.

CRediT authorship contribution statement

Shundi Duan: Writing – review & editing, Writing – original draft. **Xiao Tan:** Writing – review & editing. **Pengwei Guo:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Yurong Guo:** Writing – review & editing, Project administration, Funding acquisition. **Yi Bao:** Writing – review & editing.

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.

Acknowledgement

This research was funded by the National Natural Science Foundation of China [award numbers: 51878259].

Data availability

No data was used for the research described in the article.

References

[1] U.S. Infrastructure Grade: Explore the Categories. 2021 URL: https://infrastructurereportcard.org/infrastructure-categories/.
[2] Z. Hao, C. Lu, Z. Li, Highly accurate and automatic semantic segmentation of multiple cracks in engineered cementitious composites (ECC) under dual pre-modification deep-learning strategy, Cem. Concr. Res. 165 (2023) 107066, https://doi.org/10.1016/j.cemconres.2022.107066.
[3] L. Deng, H. Yuan, L. Long, P.-J. Chun, W. Chen, H. Chu, Cascade refinement extraction network with active boundary loss for segmentation of concrete cracks from high-resolution images, Autom. Constr. 162 (2024) 105410, https://doi.org/10.1016/j.autcon.2024.105410.
[4] J. Wang, T. Ueda, A review study on unmanned aerial vehicle and mobile robot technologies on damage inspection of reinforced concrete structures, Struct. Concr. 24 (1) (2023) 536–562, https://doi.org/10.1002/suco.202200846.
[5] G.-H. Gwon, J.-H. Lee, I.-H. Kim, S.-C. Baek, H.-J. Jung, Image-to-image translation-based structural damage data augmentation for infrastructure inspection using unmanned aerial vehicle, Drones 7 (11) (2023) 666, https://doi.org/10.3390/drones7110666.
[6] R. Fu, M. Cao, D. Novák, X. Qian, N.F. Alkayem, Extended efficient convolutional neural network for concrete crack detection with illustrated merits, Autom. Constr. 156 (2023) 105098, https://doi.org/10.1016/j.autcon.2023.105098.
[7] Q. Qiu, D. Lau, Real-time detection of cracks in tiled sidewalks using YOLO-based method applied to unmanned aerial vehicle (UAV) images, Autom. Constr. 147 (2023) 104745, https://doi.org/10.1016/j.autcon.2023.104745.
[8] P. Guo, X. Meng, W. Meng, Y. Bao, Monitoring and automatic characterization of cracks in strain-hardening cementitious composite (SHCC) through intelligent interpretation of photos, Compos. B Eng. 242 (2022) 110096, https://doi.org/10.1016/j.compositesb.2022.110096.
[9] H. Chu, L. Deng, H. Yuan, L. Long, J. Guo, A transformer and self-cascade operation-based architecture for segmenting high-resolution bridge cracks, Autom. Constr. 158 (2024) 105194, https://doi.org/10.1016/j.autcon.2023.105194.
[10] Z. Liu, Y. Cao, Y. Wang, W. Wang, Computer vision-based concrete crack detection using U-net fully convolutional networks, Autom. Constr. 104 (2019) 129–139, https://doi.org/10.1016/j.autcon.2019.04.005.
[11] H. Chu, P.J. Chun, Fine-grained crack segmentation for high-resolution images via a multiscale cascaded network, Comput. Aided Civ. Inf. Eng. 39 (4) (2024) 575–594, https://doi.org/10.1111/mice.13111.
[12] L. Xu, H. Liu, B. Xiao, X. Luo, Z. Zhu, A systematic review and evaluation of synthetic simulated data generation strategies for deep learning applications in construction, Adv. Eng. Inf. 62 (2024) 102699, https://doi.org/10.1016/j.aei.2024.102699.
[13] P. Guo, W. Meng, Y. Bao, Intelligent characterization of complex cracks in strain-hardening cementitious composites based on generative computer vision, Constr. Build. Mater. 411 (2024) 134812, https://doi.org/10.1016/j.conbuildmat.2023.134812.
[14] L. Huang, G. Fan, J. Li, H. Hao, Deep learning for automated multiclass surface damage detection in bridge inspections, Autom. Constr. 166 (2024) 105601, https://doi.org/10.1016/j.autcon.2024.105601.
[15] P. Guo, X. Meng, W. Meng, Y. Bao, Automatic assessment of concrete cracks in low-light, overexposed, and blurred images restored using a generative AI approach, Autom. Constr. 168 (2024) 105787, https://doi.org/10.1016/j.autcon.2024.105787.
[16] H. Shin, Y. Ahn, S. Tae, H. Gil, M. Song, S. Lee, Enhancement of multi-class structural defect recognition using generative adversarial network, Sustainability 13 (22) (2021) 12682, https://doi.org/10.3390/su132212682.
[17] 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. Intel. 131 (2024) 107767, https://doi.org/10.1016/j.engappai.2023.107767.
[18] J. Zhong, J. Huyan, W. Zhang, H. Cheng, J. Zhang, Z. Tong, X. Jiang, B. Huang, A deeper generative adversarial network for grooved cement concrete pavement crack detection, Eng. Appl. Artif. Intel. 119 (2023) 105808, https://doi.org/10.1016/j.engappai.2022.105808.
[19] 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. Intel. 104 (2021) 104376, https://doi.org/10.1016/j.engappai.2021.104376.
[20] H. Ma, Z. Wang, H. Gao, Z. Shen, H. Zhang, X. Hu, C. Li, G. Xiong, Parallel systems for the bridge inspection, IEEE J. Radio Frequ. Identific. 6 (2022) 783–786, https://doi.org/10.1109/JRFID.2022.3212598.
[21] C. Xiang, W. Wang, L. Deng, P. Shi, X. Kong, Crack detection algorithm for concrete structures based on super-resolution reconstruction and segmentation network, Autom. Constr. 140 (2022) 104346, https://doi.org/10.1016/j.autcon.2022.104346.
[22] Z. Hao, C. Lu, B. Dong, V.C. Li, 3D crack recognition in Engineered Cementitious Composites (ECC) based on super-resolution reconstruction and semantic segmentation of X-ray Computed Microtomography, Compos. B Eng. 285 (2024) 111730, https://doi.org/10.1016/j.compositesb.2024.111730.
[23] G. Chen, S. Teng, M. Lin, X. Yang, X. Sun, Crack detection based on generative adversarial networks and deep learning, KSCE J. Civ. Eng. 26 (4) (2022) 1803–1816, https://doi.org/10.1007/s12205-022-0518-2.
[24] Y. Liu, J.K. Yeoh, D.K. Chua, Deep learning–based enhancement of motion blurred UAV concrete crack images, J. Comput. Civ. Eng. 34 (5) (2020) 04020028, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000907.
[25] B. Xu, C. Liu, Pavement crack detection algorithm based on generative adversarial network and convolutional neural network under small samples, Measurement 196 (2022) 111219, https://doi.org/10.1016/j.measurement.2022.111219.
[26] S. Shim, Self-training approach for crack detection using synthesized crack images based on conditional generative adversarial network, Comput. Aided Civ. Inf. Eng. 39 (7) (2024) 1019–1041, https://doi.org/10.1111/mice.13119.
[27] T. Kunlamai, T. Yamane, M. Suganuma, P.J. Chun, T. Okatani, Improving visual question answering for bridge inspection by pre-training with external data of

image–text pairs, Compute-Aided Civil Infrastruct. Eng. 39 (3) (2024) 345–361, https://doi.org/10.1111/mice.13086.
[28] X. Wang, Q. Yue, X. Liu, Crack image classification and information extraction in steel bridges using multimodal large language models, Autom. Constr. 171 (2025) 105995, https://doi.org/10.1016/j.autcon.2025.105995.
[29] C. Che, Q. Lin, X. Zhao, J. Huang, L. Yu, Enhancing Multimodal Understanding with CLIP-Based Image-to-Text Transformation, in: Proceedings of the 2023 6th International Conference on Big Data Technologies, 2023, pp. 414–418, https://doi.org/10.1145/3627377.3627442.
[30] Y. Shi, L. Cui, Z. Qi, F. Meng, Z. Chen, Automatic road crack detection using random structured forests, IEEE Trans. Intell. Transp. Syst. 17 (12) (2016) 3434–3445, https://doi.org/10.1109/TITS.2016.2552248.
[31] Q. Zou, Y. Cao, Q. Li, Q. Mao, S. Wang, CrackTree: Automatic crack detection from pavement images, Pattern Recogn. Lett. 33 (3) (2012) 227–238, https://doi.org/10.1016/j.patrec.2011.11.004.
[32] M. Eisenbach, R. Stricker, D. Seichter, K. Amende, K. Debes, M. Sesselmann, D. Ebersbach, U. Stoeckert, H.-M. Gross, How to get pavement distress detection ready for deep learning? a systematic approach, Int. Joint Conf. Neural Netw. 2017 (2017) 2039–2047, https://doi.org/10.1109/IJCNN.2017.7966101.
[33] J. An, S. Cho, Variational autoencoder based anomaly detection using reconstruction probability, Specl. Lecture IE 2 (1) (2015) 1–18.
[34] F. Luleci, F.N. Catbas, O. Avci, A literature review: Generative adversarial networks for civil structural health monitoring, Front. Built Environ. 8 (2022) 1027379, https://doi.org/10.3389/fbuil.2022.1027379.
[35] G. Yu, X. Zhou, X. Chen, VDCrackGAN: a generative adversarial network with transformer for pavement crack data augmentation, Appl. Sci. 14 (17) (2024) 7907, https://doi.org/10.3390/app14177907.
[36] T. Jin, X.-W. Ye, Z. Li, Establishment and evaluation of conditional GAN-based image dataset for semantic segmentation of structural cracks, Eng. Struct. 285 (2023) 116058, https://doi.org/10.1016/j.engstruct.2023.116058.
[37] C.C. Rakowski, T. Bourlai, On enhancing crack semantic segmentation using StyleGAN and Brownian bridge diffusion, IEEE Access (2024), https://doi.org/10.1109/ACCESS.2024.3368376.
[38] 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) (2022) 9092, https://doi.org/10.3390/s22239092.
[39] H. Zhang, N. Chen, M. Li, S. Mao, The crack diffusion model: an innovative diffusion-based method for pavement crack detection, Remote Sens. (Basel) 16 (6) (2024) 986, https://doi.org/10.3390/rs16060986.
[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, IEEE/CVF Conf. Comput. Vision Pattern Recogn. (2022) 10684–10695, https://doi.org/10.48550/arXiv.2112.10752.
[41] Peng, X., Koch, J., and Mackay, W.E., Designprompt: Using multimodal interaction for design exploration with generative ai. Proceedings of the 2024 ACM Designing Interactive Systems Conference, 2024: p. 804-818 Doi: 10.1145/3643834.3661588.
[42] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, Int Conf Machine Learning (2021) 8821–8831, https://doi.org/10.48550/arXiv.2102.12092.
[43] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E.L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, Photorealistic text-to-image diffusion models with deep language understanding, Adv. Neural Inf. Proces. Syst. 35 (2022) 36479–36494, https://doi.org/10.48550/arXiv.2205.11487.
[44] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, Make-a-video: Text-to-video generation without text-video data, ArXiv Preprint (2022), https://doi.org/10.48550/arXiv.2209.14792.
[45] H. Pu, X. Yang, J. Li, R. Guo, AutoRepo: a general framework for multimodal LLM-based automated construction reporting, Expert Syst. Appl. 255 (2024) 124601, https://doi.org/10.1016/j.eswa.2024.124601.
[46] Topcu, O.K., An AI-Assisted Bridge Inspection System: Ontology-Based Visual Question-Answering Methodology Using Large Language Models. *Southern Illinois University at Edwardsville*, 2023 URL: https://www.proquest.com/openview/6259c78573d93823d7b8a8846b6aaa93/1?pq-origsite=gscholar&cbl=18750&diss=y.
[47] S. Choi, S. Yoon, GPT-based data-driven urban building energy modeling (GPT-UBEM): Concept, methodology, and case studies, Energ. Buildings 325 (2024) 115042, https://doi.org/10.1016/j.enbuild.2024.115042.
[48] F. Xu, T. Nguyen, J. Du, Augmented reality for maintenance tasks with ChatGPT for automated text-to-action, J. Constr. Eng. Manag. 150 (4) (2024) 04024015, https://doi.org/10.1061/JCEMD4.COENG-14142.
[49] F. Panella, A. Lipani, J. Boehm, Semantic segmentation of cracks: Data challenges and architecture, Autom. Constr. 135 (2022) 104110, https://doi.org/10.1016/j.autcon.2021.104110.
[50] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 1–48, https://doi.org/10.1186/s40537-019-0197-0.
[51] J. Yu, Y. Weng, J. Yu, W. Chen, S. Lu, K. Yu, Generative AI for performance-based design of engineered cementitious composite, Compos. B Eng. 266 (2023) 110993, https://doi.org/10.1016/j.compositesb.2023.110993.
[52] E. Vrochidou, G.K. Sidiropoulos, A.G. Ouzounis, I. Tsimperidis, V. Kalpakis, A. Stamkos, G.A. Papakostas, Utilizing Generative AI for Crack detection in the marble industry, Eng. Res. Express (2025), https://doi.org/10.1088/2631-8695/adaca7.
[53] H. Lyu, N. Sha, S. Qin, M. Yan, Y. Xie, R. Wang, Adv. Neural Inf. Proces. Syst. (2019) 32. https://par.nsf.gov/biblio/10195511.
[54] P. Liu, H. Qi, J. Liu, L. Feng, D. Li, J. Guo, Automated clash resolution for reinforcement steel design in precast concrete wall panels via generative adversarial network and reinforcement learning, Adv. Eng. Inf. 58 (2023) 102131, https://doi.org/10.1016/j.aei.2024.102791.
[55] J. Li, X. Wang, J. Li, J. Zhang, G. Ma, A generative adversarial learning strategy for spatial inspection of compaction quality, Adv. Eng. Inf. 62 (2024) 102791, https://doi.org/10.1016/j.aei.2024.102791.
[56] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., Generative adversarial nets. ArXiv:1406.2661, 2014. **27** Doi: 10.48550/arXiv.1406.2661.
[57] E. Branikas, P. Murray, G. West, A novel data augmentation method for improved visual crack detection using generative adversarial networks, IEEE Access 11 (2023) 22051–22059, https://doi.org/10.1109/ACCESS.2023.3251988.
[58] N. Zhao, Y. Song, A. Yang, K. Lv, H. Jiang, C. Dong, Accurate classification of tunnel lining cracks using lightweight shuffleNetV2-1.0-SE model with DCGAN-based data augmentation and transfer learning, Appl. Sci. 14 (10) (2024) 4142, https://doi.org/10.3390/app14104142.
[59] K. Hacıefendioğlu, A.C. Altunışık, T. Abdioğlu, Deep learning-based automated detection of cracks in historical masonry structures, Buildings 13 (12) (2023) 3113, https://doi.org/10.3390/buildings13123113.
[60] X. Zhou, M. Li, Y. Liu, W. Yu, M. Elchalakani, Cross-domain damage identification of bridge based on generative adversarial and deep adaptation networks, Structures 64 (2024) 106540, https://doi.org/10.1016/j.istruc.2024.106540.
[61] N. Syamala, C.A. Kumar, V.P. Kumar, A. Vamshi, M. Gayathri, L.Y. Chowdary, Crack detection and structural assessment in bridges using generative adversarial networks, Int. Conf. Machine Learn. Autonom. Syst. (ICMLAS) 2025 (2025) 867–872, https://doi.org/10.1109/ICMLAS64557.2025.10967778.
[62] E. El-Din Hemdan, M.J.I.J.o.P.R. Al-Atroush, A review study of intelligent road crack detection: algorithms and systems, Int. J. Pavement Res. Technol. (2025) 1–31, https://doi.org/10.1007/s42947-025-00556-x.
[63] S. Zhu, T. Xiang, M. Yang, Y. Li, Application of deep convolutional generative adversarial network to identification of bridge structural damage, Int. J. Struct. Stab. Dyn. 25 (10) (2025) 2550098, https://doi.org/10.1142/S0219455425500981.
[64] S. Gao, C. Wan, Z. Zhou, J. Hou, L. Xie, S. Xue, Enhanced data imputation framework for bridge health monitoring using Wasserstein generative adversarial networks with gradient penalty, Structures 57 (2023) 105277, https://doi.org/10.1016/j.istruc.2023.105277.
[65] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, ArXiv:1704.00028 30 (2017), https://doi.org/10.48550/arXiv.1704.00028.
[66] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410, https://doi.org/10.1109/CVPR.2019.00453.
[67] Y. Que, Y. Dai, X. Ji, A.K. Leung, Z. Chen, Z. Jiang, Y. Tang, Automatic classification of asphalt pavement cracks using a novel integrated generative adversarial networks and improved VGG model, Eng. Struct. 277 (2023) 115406, https://doi.org/10.1016/j.engstruct.2022.115406.
[68] B. Padhi, M. Reza, M.S. Shams, A.N. Sai, Concrete crack detection using deep convolutional generative adversarial network, Int. Adv. Comput. Conf. (2022) 147–161, https://doi.org/10.1007/978-3-031-35641-4_11.
[69] 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. Transp. Syst. 24 (4) (2023) 4474–4481, https://doi.org/10.1109/TITS.2023.3236247.
[70] K. Dunphy, M.N. Fekri, K. Grolinger, A. Sadhu, Data augmentation for deep-learning-based multiclass structural damage detection using limited information, Sensors 22 (16) (2022) 6193, https://doi.org/10.3390/s22166193.
[71] F. Guo, Q. Cui, H. Zhang, Y. Wang, H. Zhang, X. Zhu, J. Chen, A new deep learning-based approach for concrete crack identification and damage assessment, Adv. Struct. Eng. (2024), https://doi.org/10.1177/13694332241266535.
[72] J. Kim, J. Seon, S. Kim, Y. Sun, S. Lee, J. Kim, B. Hwang, J. Kim, Generative AI-driven data augmentation for crack detection in physical structures, Electronics 13 (19) (2024) 3905, https://doi.org/10.3390/electronics13193905.
[73] S. Li, S. Li, H. Li, Z. Zhou, Data enhancement and feature extraction optimization in tunnel surface defect detection: combining DCGAN-RC and Repvit-YOLO methods, Eng. Fail. Anal. (2025) 109715, https://doi.org/10.1016/j.engfailanal.2025.109715.
[74] X. Zhang, J. Ding, Y. Wang, K. Wang, Cogeneration method for crack images and masks, Autom. Constr. 171 (2025) 105985, https://doi.org/10.1016/j.autcon.2025.105985.
[75] W. Choi, J. Heo, C. Ahn, Development of road surface detection algorithm using cyclegan-augmented dataset, Sensors 21 (22) (2021) 7769, https://doi.org/10.3390/s21227769.
[76] W. Kong, Y. Li, J. Hu, S. Luo, J. Rui, A Shadow-robust pavement damage detection framework based on RACycle-GAN and DDE-YOLOv8, IEEE Trans. Intell. Transp. Syst. (2025), https://doi.org/10.1109/TITS.2025.3556941.
[77] 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) (2022) 14304, https://doi.org/10.3390/su142114304.
[78] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, IEEE Int. Conf. Comput. Vision (2017) 2223–2232, https://doi.org/10.48550/arXiv.1703.10593.

[79] S. Li, X. Zhao, High-resolution concrete damage image synthesis using conditional generative adversarial network, Autom. Constr. 147 (2023) 104739, https://doi.org/10.1016/j.autcon.2022.104739.
[80] Li, Z., Xiao, X., Xie, J., Fan, Y., Wang, W., Chen, G., Zhang, L., and Wang, T., Cycle-YOLO: A Efficient and Robust Framework for Pavement Damage Detection. *ArXiv preprint arXiv:.17905*, 2024 Doi: 10.48550/arXiv.2405.17905.
[81] B. Li, H. Guo, Z. Wang, Data augmentation using CycleGAN-based methods for automatic bridge crack detection, Structures 62 (2024) 106321, https://doi.org/10.1016/j.istruc.2024.106321.
[82] F. Ni, Z. He, S. Jiang, W. Wang, J. Zhang, A generative adversarial learning strategy for enhanced lightweight crack delineation networks, Adv. Eng. Inf. 52 (2022) 101575, https://doi.org/10.1016/j.aei.2022.101575.
[83] S. Shim, J. Kim, S.-W. Lee, G.-C. Cho, Road damage detection using super-resolution and semi-supervised learning with generative adversarial network, Autom. Constr. 135 (2022) 104139, https://doi.org/10.1016/j.autcon.2022.104139.
[84] K. Han, V.S. Sheng, Y. Song, Y. Liu, C. Qiu, S. Ma, Z. Liu, Deep semi-supervised learning for medical image segmentation: a review, Expert Syst. Appl. (2024) 123052, https://doi.org/10.1016/j.eswa.2023.123052.
[85] X. Fan, M. Khishe, A. Alqahtani, S. Alsubai, A. Alanazi, M.M. Zaidi, A dual adaptive semi-supervised attentional residual network framework for urban sound classification, Adv. Eng. Inf. 62 (2024) 102761, https://doi.org/10.1016/j.aei.2024.102761.
[86] G. Zhang, Y. Pan, L. Zhang, Semi-supervised learning with GAN for automatic defect detection from images, Autom. Constr. 128 (2021) 103764, https://doi.org/10.1016/j.autcon.2021.103764.
[87] K.-L. Lu, G.-R. Luo, M. Zhang, J.-F. Qi, C.-Y. Huang, Comparison of deep learning methods and a transfer-learning semi-supervised GAN combined framework for pavement crack image identification, Cecnet (2022) 499–508, https://doi.org/10.3233/FAIA220571.
[88] R. Zhu, G. Niu, Z. Qu, P. Wang, R. Zhang, D. Fang, In-situ quantitative tracking of micro-crack evolution behavior inside CMCs under load at high temperature: a deep learning method, Acta Mater. 255 (2023) 119073, https://doi.org/10.1016/j.actamat.2023.119073.
[89] Yasuno, T., Nakajima, M., Sekiguchi, T., Noda, K., Aoyanagi, K., and Kato, S., Synthetic image augmentation for damage region segmentation using conditional GAN with structure edge. *ArXiv preprint arXiv:.08628*, 2020 Doi: 10.48550/arXiv.2005.08628.
[90] S.-B. Shim, A study on generation quality comparison of concrete damage image using stable diffusion base models, Korea Instit. Struct. Mainten. Inspect. 28 (4) (2024) 55–61, https://doi.org/10.11112/jksmi.2024.28.4.55.
[91] S. Shim, Self-supervised domain adaptive approach for extrapolated crack segmentation with fine-tuned inpainting generative model, Comput. Aided Civ. Inf. Eng. (2025), https://doi.org/10.1111/mice.13517.
[92] S. Li, Y. Le, X. Zhao, Style-controlled image synthesis of concrete damages based on fusion of convolutional encoder and attention-enhanced conditional generative adversarial network, J. Comput. Civ. Eng. 38 (6) (2024) 04024032, https://doi.org/10.1061/JCCEE5.CPENG-6007.
[93] S. Izadi, D. Sutton, G. Hamarneh, Image denoising in the deep learning era, Artif. Intell. Rev. 56 (7) (2023) 5929–5974, https://doi.org/10.1007/s10462-022-10305-2.
[94] P. Guo, X. Meng, W. Meng, Y. Bao, Automatic assessment of concrete cracks with low-light, overexposed, and blur images restored using a generative artificial intelligence (AI) approach, Autom. Constr. (2024), https://doi.org/10.1016/j.autcon.2024.105787.
[95] Y. Kondo, N. Ukita, Joint learning of blind super-resolution and crack segmentation for realistic degraded images, IEEE Trans. Instrum. Meas. (2024), https://doi.org/10.1109/TIM.2024.3374293.
[96] R. Sun, X. Li, S.-S. Law, L. Zhang, L. Hu, G. Liu, An improved EnlightenGAN shadow removal framework for images of cracked concrete, Mech. Syst. Sig. Process. 223 (2025) 111943, https://doi.org/10.1016/j.ymssp.2024.111943.
[97] W. Shen, D. Zeng, Y. Zhang, X. Tian, Z. Li, Image augmentation for nondestructive testing in engineering structures based on denoising diffusion probabilistic model, J. Build. Eng. 89 (2024) 109299, https://doi.org/10.1016/j.jobe.2024.109299.
[98] L. Chen, L. Zhou, L. Li, M. Luo, CrackDiffusion: crack inpainting with denoising diffusion models and crack segmentation perceptual score, Smart Mater. Struct. 32 (5) (2023) 054001, https://doi.org/10.1088/1361-665X/acc624.
[99] J. Wang, G. Xiao, H. Zhu, W. Li, J. Cui, Y. Wan, Z. Wang, Q. Sui, N-LoLiGan: Unsupervised low-light enhancement GAN with an N-Net for low-light tunnel images, Digital Signal Process. 143 (2023) 104259, https://doi.org/10.1016/j.dsp.2023.104259.
[100] M. Afifi, K.G. Derpanis, B. Ommer, M.S. Brown, Learning multi-scale photo exposure correction, IEEE/CVF Conf. Comput. Vision Patt. Recogn. (2021) 9157–9167, https://doi.org/10.48550/arXiv.2003.11596.
[101] W. Wang, C. Su, G. Han, Enhancement of motion blurred crack images based on conditional generative adversarial network, Arab. J. Sci. Eng. (2024) 1–17, https://doi.org/10.1007/s13369-024-09772-2.
[102] Y. Zhang, L. Zhang, A generative adversarial network approach for removing motion blur in the automatic detection of pavement cracks, Computer-Aided Civil Infrastr. Eng. (2024), https://doi.org/10.1111/mice.13231.
[103] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, J. Matas, Deblurgan: blind motion deblurring using conditional adversarial networks, IEEE Conf. Computer Vision Pattern Recognit. (2018) 8183–8192, https://doi.org/10.48550/arXiv.1711.07064.
[104] K. Sathya, D. Sangavi, P. Sridharshini, M. Manobharathi, G. Jayapriya, Improved image based super resolution and concrete crack prediction using pre-trained deep learning models, J. Soft Comput. Civil Eng. 4 (3) (2020) 40–51, https://doi.org/10.22115/scce.2020.229355.1219.
[105] J. Kim, S. Shim, S.-J. Kang, G.-C. Cho, Learning structure for concrete crack detection using robust super-resolution with generative adversarial network, Struct. Control Health Monit. 2023 (1) (2023) 8850290, https://doi.org/10.1155/2023/8850290.
[106] D. Oh, S. Jeong, S.-K. Bae, B. Kim, S. Cho, Training deep learning segmentation models using super-resolution crack images for detection of thin concrete cracks, J. Comput. Civ. Eng. 39 (4) (2025) 04025035, https://doi.org/10.1061/JCCEE5.CPENG-624.
[107] S. Cano-Ortiz, E. Sainz-Ortiz, L.L. Iglesias, P.M.R. del Árbol, D. Castro-Fresno, Enhancing pavement crack segmentation via semantic diffusion synthesis model for strategic road assessment, Results Eng. 23 (2024) 102745, https://doi.org/10.1016/j.rineng.2024.102745.
[108] C. Han, H. Yang, T. Ma, S. Wang, C. Zhao, Y. Yang, CrackDiffusion: a two-stage semantic segmentation framework for pavement crack combining unsupervised and supervised processes, Autom. Constr. 160 (2024) 105332, https://doi.org/10.1016/j.autcon.2024.105332.
[109] Z. Pan, S.L. Lau, X. Yang, N. Guo, X. Wang, Automatic pavement crack segmentation using a generative adversarial network (GAN)-based convolutional neural network, Results Eng. 19 (2023) 101267, https://doi.org/10.1016/j.rineng.2023.101267.
[110] K. Zhang, Y. Zhang, H.-D. Cheng, CrackGAN: Pavement crack detection using partially accurate ground truths based on generative adversarial learning, IEEE Trans. Intell. Transp. Syst. 22 (2) (2020) 1306–1319, https://doi.org/10.1109/TITS.2020.2990703.
[111] K.A. Nandini, S.S. Dharshan, T. Bandaragoda, Automated Crack Analysis and Reporting in Civil Infrastructure using Generative AI, in: IECON 2024-50th Annual Conference of the IEEE Industrial Electronics Society, 2024, pp. 1–6, https://doi.org/10.1109/IECON55916.2024.10905875.
[112] P. Liao, G. Nakano, BridgeCLIP: Automatic Bridge Inspection by Utilizing Vision-Language Model, Int. Conf. Pattern Recognit. (2025) 61–76, https://doi.org/10.1007/978-3-031-78447-7_5.
[113] T. Yamane, P.-J. Chun, J. Dang, T. Okatani, Deep learning-based bridge damage cause estimation from multiple images using visual question answering, Struct. Infrastruct. Eng. (2024) 1–14, https://doi.org/10.1080/15732479.2024.2355929.
[114] R.D.S. Lowande, H.E. Sevil, Feasibility of visual question answering (vqa) for post-disaster damage detection using aerial footage, Appl. Sci. 13 (8) (2023) 5079, https://doi.org/10.3390/app13085079.
[115] P.-J. Chun, H. Chu, K. Shitara, T. Yamane, Y. Maemura, Implementation of explanatory texts output for bridge damage in a bridge inspection web system, Adv. Eng. Softw. 195 (2024) 103706, https://doi.org/10.1016/j.advengsoft.2024.103706.
[116] K. Song, Y. Zhu, B. Liu, Q. Yan, A. Elgammal, X. Yang, Moma: Multimodal llm adapter for fast personalized image generation, Europ. Conf. Comput. Vision (2024) 117–132, https://doi.org/10.1007/978-3-031-73661-2_7.
[117] Y. Jiang, J. Wang, X. Shen, K. Dai, Large language model for post-earthquake structural damage assessment of buildings, Comput. Aided Civ. Inf. Eng. (2025), https://doi.org/10.1111/mice.70010.
[118] S. Jayati, E. Choi, H. Burton, S. Newsam, Leveraging Large Multimodal Models to Augment Image-based Building damage Assessment, in: Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, 2024, pp. 79–85, https://doi.org/10.1145/3687123.3698291.
[119] Zhang, Y., Wei, S., Huang, Y., Su, Y., Lu, S., and Li, H., SDIGLM: Leveraging Large Language Models and Multi-Modal Chain of Thought for Structural Damage Identification. *ArXiv preprint arXiv:2504.11477*, 2025 DOI: 10.48550/arXiv.2504.11477.
[120] Y. Zang, W. Li, J. Han, K. Zhou, C.C. Loy, Contextual object detection with multimodal large language models, Int. J. Comput. Vis. (2024) 1–19, https://doi.org/10.1007/s11263-024-02214-4.
[121] I. Khokhlov, E. Davydenko, I. Osokin, I. Ryakin, A. Babaev, V. Litvinenko, R. Gorbachev, Tiny-YOLO object detection supplemented with geometrical data, in: 2020 IEEE 91st Vehicular Technology Conference, 2020, pp. 1–5, https://doi.org/10.1109/VTC2020-Spring48590.2020.9128749.
[122] P. Ersoy, M. Erşahin, Optimal llm execution strategies for llama 3.1 language models across diverse hardware configurations: a comprehensive guide, Comput. Intellig. Mach. Learn. 5 (2024), https://doi.org/10.36647/CIML/05.01.A002.
[123] R. Zhang, J. Han, C. Liu, A. Zhou, P. Lu, Y. Qiao, H. Li, P. Gao, LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention, The Twelfth International Conference on Learning Representations (2024).
[124] A. Mukherjee, M. Sasidharan, M. Herrera, A.K. Parlikad, Unsupervised constrained discord detection in IoT-based online crane monitoring, Adv. Eng. Inf. 60 (2024) 102444, https://doi.org/10.1016/j.aei.2024.102444.
[125] J. Gou, B. Yu, S.J. Maybank, D. Tao, Knowledge distillation: a survey, Int. J. Comput. Vis. 129 (6) (2021) 1789–1819, https://doi.org/10.1007/s11263-021-01453-z.
[126] Y. Jin, H. Gao, X. Fan, H. Khan, Y. Chen, Defect identification of adhesive structure based on DCGAN and YOLOv5, IEEE Access 10 (2022) 79913–79924, https://doi.org/10.1109/ACCESS.2022.3193775.
[127] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R., Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv preprint arXiv:.01952*, 2023 Doi: 10.48550/arXiv.2307.01952.
[128] CLIP. URL: https://huggingface.co/openai/clip-vit-large-patch14.

[129] Li, B., Zhang, J., Fan, J., Xu, Y., Chen, C., Tang, N., and Luo, Y., Alpha-sql: Zero-shot text-to-sql using monte carlo tree search. *ArXiv preprint arXiv:.17248*, 2025 Doi: 10.48550/arXiv.2502.17248.
[130] Abacha, A.B., Yim, W.-w., Fu, Y., Sun, Z., Yetisgen, M., Xia, F., and Lin, T., Medec: A benchmark for medical error detection and correction in clinical notes. *ArXiv preprint arXiv:.19260*, 2024 Doi: 10.48550/arXiv.2412.19260.
[131] J. Xu, C. Yuan, J. Gu, J. Liu, J. An, Q. Kong, Innovative synthetic data augmentation for dam crack detection, segmentation, and quantification, Struct. Health Monit. 22 (4) (2023) 2402–2426, https://doi.org/10.1177/14759217221122318.
[132] P. Guo, Z. Xue, L.R. Long, S. Antani, Cross-dataset evaluation of deep learning networks for uterine cervix segmentation, Diagnostics 10 (1) (2020) 44, https://doi.org/10.3390/diagnostics10010044.
[133] D. Dais, I.E. Bal, E. Smyrou, V. Sarhosis, Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning, Autom. Constr. 125 (2021) 103606, https://doi.org/10.1016/j.autcon.2021.103606.