Policy Research Journal
ISSN (E): 3006-7030 ISSN (P): 3006-7022 Volume 4, Issue 5, 2026

COMPUTER VISION-BASED SURFACE CRACK DETECTION IN CONCRETE PAVEMENTS: KEY METHODS, ARCHITECTURES, AND CHALLENGES

Dr. M. Adil Khan[1], Tabish Qureshi[2], Tariq Shah[*3], Saad Hanif[4], Shumaila Hussain[5]

[1]*Resident Engineer, NESPAK*
[2]*Upsource by Solutions (Solutions by STC) Masters of Computer Science - University of Karachi*
[*3]*CECOS University of IT and Emerging Sciences*
[4]*Zachry Department of Civil and Environmental Engineering, Texas A&M University, USA*
[5]*Assistant Professor SBKWU Quetta*

[1]adee.uol@gmail.com, [2]tabish.qureshi82@gmail.com, [*3]tariqshah13813@gmail.com,
[4]saadhanif107@tamu.edu, [5]shumailahussain70@gmail.com

DOI: http://doi.org/10.5281/zenodo.20133678

Keywords
Concrete pavement cracks, computer vision, deep learning, semantic segmentation, transformer networks

Article History
Received: 15 March 2026
Accepted: 24 April 2026
Published: 12 May 2026

Corresponding Author: *
Tariq Shah

Abstract

Concrete pavement crack detection is a critical component of infrastructure maintenance and structural health monitoring because surface cracks directly influence pavement durability, safety, and service life. Traditional manual inspection methods are labor-intensive, subjective, time-consuming, and inefficient for large-scale transportation networks. Consequently, computer vision-based crack detection techniques have emerged as a promising alternative for automated pavement assessment. This review critically examines recent advances in computer vision-based surface crack detection in concrete pavements, focusing on preprocessing methods, segmentation strategies, convolutional neural network architectures, transformer-based models, and multimodal sensing approaches. The study highlights the transition from conventional image processing and handcrafted-feature machine learning techniques toward deep learning-based semantic segmentation frameworks capable of achieving high pixel-level accuracy. Widely used architectures such as U-Net, DeepLabv3+, SegNet, ResNet, EfficientNet, MobileNet, and transformer hybrids including SwinUNet and CrackFormer are comparatively analyzed in terms of accuracy, computational efficiency, robustness, and deployment suitability. The review further discusses benchmark datasets, evaluation metrics, real-time implementation challenges, dataset imbalance, environmental variability, and the limitations of current systems in practical pavement management applications. Recent developments in attention mechanisms, RGB–infrared fusion, stereo vision, and lightweight embedded models are also explored. Finally, the paper identifies future research directions emphasizing multimodal datasets, efficient transformer architectures, domain adaptation, and integration of crack detection with automated structural condition assessment systems.

INTRODUCTION

Transportation infrastructure plays a fundamental role in economic development, urban mobility, and public safety. Concrete pavements are widely used in highways, bridges, airport runways, parking structures, and industrial facilities because of their durability, strength, and long service life. However, pavement surfaces are continuously subjected to traffic loading, environmental exposure, temperature fluctuations, moisture infiltration, freeze–thaw cycles, and material aging, which eventually lead to the formation of surface cracks and other forms of distress. Crack propagation reduces pavement performance, accelerates structural deterioration, increases maintenance costs, and may compromise user safety if not detected and repaired at an early stage (Cao et al., 2020; Golding et al., 2022). Therefore, accurate and timely crack detection is essential for effective pavement management and infrastructure maintenance planning.

Traditionally, pavement crack inspection has relied on manual surveys conducted by trained inspectors. Although manual inspection remains widely practiced, it suffers from several limitations, including subjectivity, inconsistency between inspectors, high labor requirements, safety risks during roadway inspection, and low efficiency for large-scale transportation networks (Cao et al., 2020; Zhang & Zhang, 2023). Manual methods are particularly challenging when cracks are thin, irregular, low-contrast, or located under poor illumination conditions. As transportation agencies increasingly require continuous and large-scale infrastructure monitoring, automated inspection systems have become a major research focus in structural health monitoring and intelligent transportation systems.

Computer vision-based crack detection has emerged as a promising alternative capable of replacing slow and subjective manual surveys with scalable, objective, and real-time inspection systems (Golding et al., 2022; Chen et al., 2023). Early computer vision methods primarily depended on conventional image processing techniques such as thresholding, edge detection, filtering, morphological operations, histogram analysis, and handcrafted feature extraction. These techniques attempted to distinguish cracks from the pavement background by exploiting differences in pixel intensity, texture, or geometry (Arafin et al., 2023; Cao et al., 2020). Although classical methods achieved moderate success under controlled imaging conditions, their performance significantly deteriorated in the presence of shadows, stains, uneven lighting, rough textures, and environmental noise (Huyan et al., 2022; Fan et al., 2022). Variations in pavement material, camera quality, and crack morphology further reduced the generalization capability of traditional approaches.

The rapid development of machine learning and deep learning has fundamentally transformed pavement crack detection research over the past decade. Convolutional neural networks (CNNs) have become the dominant paradigm because of their ability to automatically learn hierarchical feature representations directly from raw images without relying on handcrafted descriptors (Golding et al., 2022; Chen et al., 2023). Deep learning models can capture complex crack patterns, texture variations, and contextual information, enabling significantly higher accuracy and robustness than traditional image processing methods. Research has progressively evolved from image-level crack classification toward pixel-level semantic segmentation capable of accurately delineating crack boundaries and geometry (Ren et al., 2020; Sun et al., 2022).

Among deep learning approaches, encoder–decoder architectures such as U-Net, SegNet, DeepLabv3+, and fully convolutional networks have demonstrated remarkable performance in pavement crack segmentation tasks (Chen et al., 2020; Ren et al., 2020; Sun et al., 2022). These architectures use downsampling operations to extract high-level semantic features and upsampling paths to reconstruct detailed crack masks at pixel level. Skip connections and multi-scale feature fusion mechanisms further improve the detection of fine and discontinuous cracks. U-Net variants employing pre-trained backbones such as VGG16, ResNet34, ResNet50, and EfficientNetB3 have reported F1 scores exceeding

95% on several public crack datasets (Huyan et al., 2022; Nyathi et al., 2024; Fan et al., 2022). Similarly, DeepLabv3+ models enhanced with multi-scale attention mechanisms have achieved state-of-the-art segmentation performance by dynamically weighting low-level and high-level crack features (Sun et al., 2022).

Recent studies increasingly focus on lightweight and real-time architectures designed for deployment on embedded devices and vehicle-mounted inspection systems. Models such as ECSNet and MANet utilize depthwise separable convolutions, mobile backbones, and compact feature extraction modules to reduce computational complexity while maintaining competitive segmentation accuracy (Zhang et al., 2023; Chen et al., 2023). These lightweight frameworks are particularly important for practical implementation in autonomous pavement inspection vehicles, drones, and edge computing platforms where memory and processing resources are limited.

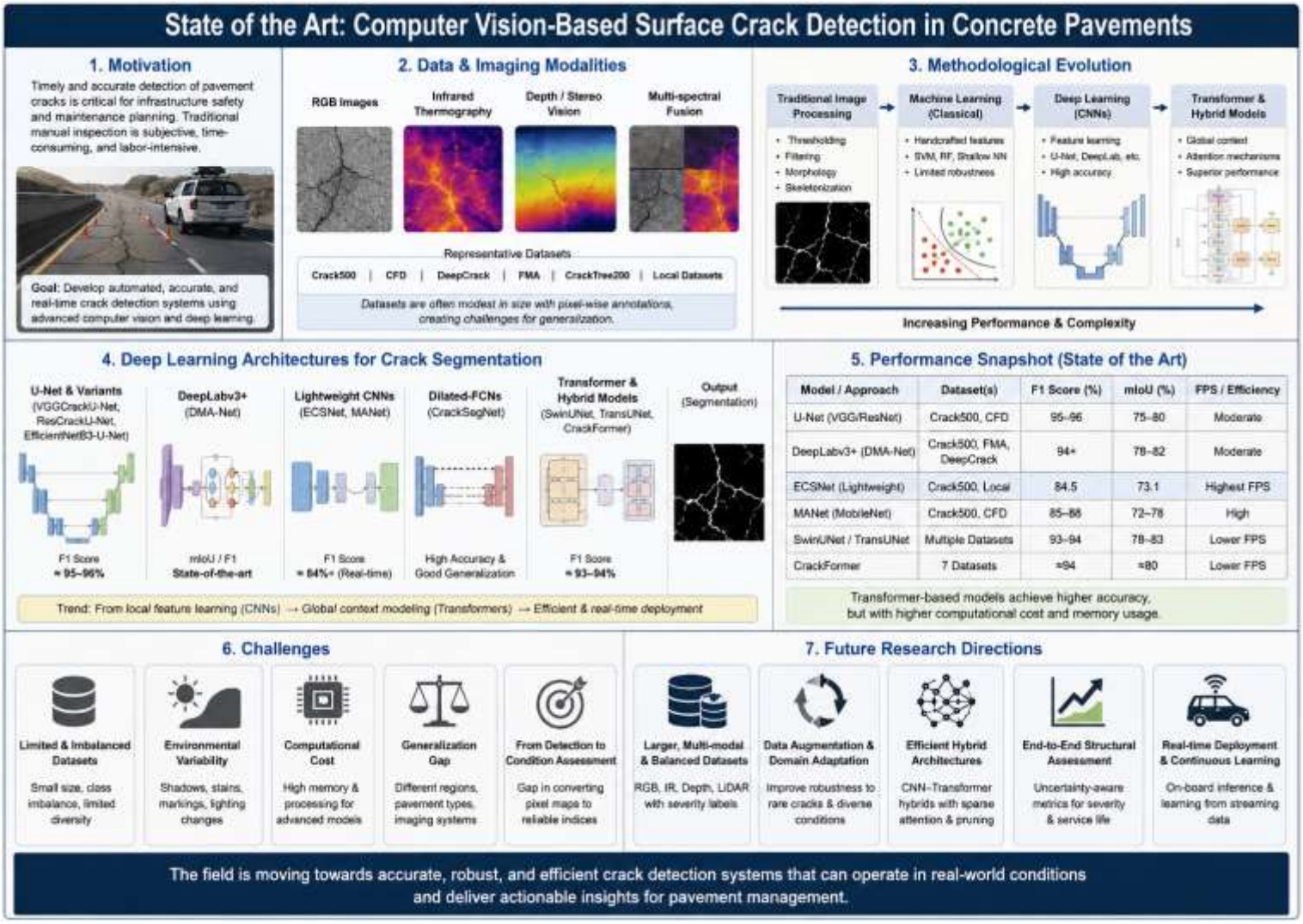

Figure 1 State-of-the-art overview of computer vision-based surface crack detection in concrete pavements, illustrating the evolution from traditional image processing techniques to deep learning and transformer-based architectures, along with imaging modalities, benchmark datasets, performance trends, major challenges, and future research directions in automated pavement inspection systems.

Another major development in computer vision-based crack detection is the integration of transformer architectures and attention mechanisms. Transformer-based models such as SwinUNet, TransUNet, MTUNet, and CrackFormer have demonstrated strong capability in modeling long-range dependencies and global contextual relationships within pavement images (Zhang & Zhang, 2023; Xiao et al., 2023; Guo et al., 2023). Unlike conventional CNNs, transformers employ self-attention mechanisms that allow the network to capture both local crack details and broader spatial patterns simultaneously. This capability is particularly valuable for detecting thin, tortuous, and discontinuous cracks embedded within complex pavement backgrounds. Hybrid CNN–transformer frameworks have shown improved segmentation accuracy and convergence stability, although they often require higher computational resources and memory consumption (Guo et al., 2023; Xiao et al., 2023).

In addition to advances in network architectures, recent research has explored multimodal sensing techniques to improve crack detection robustness under challenging environmental conditions. RGB images are commonly used because of their low cost and ease of acquisition, but visible-light imaging alone may struggle under low contrast or shadowed environments. Consequently, studies have investigated the fusion of visible images with infrared thermography, stereo vision, depth sensing, and laser-based measurements to enhance crack visibility and geometric estimation (Liu et al., 2022; Jinchao et al., 2021). Infrared imaging exploits temperature differences between cracked and intact pavement surfaces, while stereo vision systems enable three-dimensional measurement of crack width, depth, and pothole volume. Such multimodal approaches improve segmentation stability and provide richer information for structural condition assessment.

Despite substantial progress, several important challenges remain unresolved. Existing datasets are often relatively small, imbalanced, and limited to specific environmental conditions, restricting the generalization capability of deep learning models across different pavement types and geographical regions (Chen et al., 2020; Meftah et al., 2024). Environmental variability caused by lighting changes, shadows, stains, road markings, and rough textures continues to generate false detections and segmentation errors (Fan et al., 2022; Zhang & Zhang, 2023). Furthermore, many high-performing transformer-based models require significant computational resources, limiting their practical deployment in real-time pavement monitoring systems (Chen et al., 2023; Guo et al., 2023). Another critical issue is the lack of standardized benchmark datasets and evaluation protocols, which complicates fair comparison between competing methods.

Given the rapid expansion of research in this field, a comprehensive review of current computer vision-based pavement crack detection techniques is necessary to consolidate recent developments, identify research gaps, and guide future investigations. This review therefore critically examines existing methods for concrete pavement crack detection, focusing on preprocessing strategies, semantic segmentation techniques, CNN architectures, transformer-based frameworks, performance evaluation metrics, multimodal sensing systems, and deployment challenges. The paper also discusses emerging trends and future research opportunities aimed at improving accuracy, efficiency, robustness, and practical integration within intelligent pavement management systems.

Scope of Methods and Datasets

Existing work spans classification, detection, and semantic segmentation using 2D RGB images, multispectral data (infrared), and 3D/stereo imaging, often with transfer learning on large generic backbones (Golding et al., 2022; Zhang et al., 2023; Chen et al., 2020; Ren et al., 2020; Chen et al., 2023; Sun et al., 2022; Nyathi et al., 2024; Zhang & Zhang, 2023; Iraniparast et al., 2023; Huyan et al., 2022)[18–20]. Crack datasets include general concrete surfaces and specific pavement sets such as Crack500, CFD, DeepCrack, CrackTree200, FMA, and locally collected smartphone or monitoring-vehicle images (Zhang

et al., 2023; Chen et al., 2020; Ren et al., 2020; Chen et al., 2023; Sun et al., 2022; Zhang & Zhang, 2023; Fan et al., 2025; Huyan et al., 2022; Fan et al., 2022; Jinchao et al., 2021).

Automatic vision-based crack detection aims to replace subjective, slow manual surveys with scalable, objective inspection for pavements and other concrete infrastructures (Golding et al., 2022; Cao et al., 2020; Zhang & Zhang, 2023). Recent work focuses on pixel-level segmentation, high accuracy under noise, and real-time performance on embedded platforms (Zhang et al., 2023; Chen et al., 2023; Sun et al., 2022; Huyan et al., 2022; Fan et al., 2022). Deep learning, especially CNNs and emerging transformer-based models, dominates current research on concrete pavement crack detection (Cao et al., 2020; Chen et al., 2023; Sun et al., 2022; Zhang & Zhang, 2023; Xiao et al., 2023; Guo et al., 2023).

Table 1 Key pavement crack datasets and imaging modalities

Dataset / source type	Surface & scale	Acquisition and imaging notes	Citations
Crack500, CFD, DeepCrack, FMA	Pavement cracks, pixel-wise labels	High-resolution RGB; public benchmarks for segmentation and comparison of DL models	(Zhang et al., 2023; Chen et al., 2023; Sun et al., 2022; Fan et al., 2025; Fan et al., 2022)
Smartphone / action camera / monitoring systems	Asphalt and concrete pavements	On-vehicle or handheld imaging under diverse lighting and texture	(Zhang et al., 2023; Ren et al., 2020; Zhang & Zhang, 2023; Huyan et al., 2022; Jinchao et al., 2021)
Multi-view stereo / color–depth datasets	Asphalt distress (cracks, potholes)	Stereo vision with depth maps enabling 3D crack and pothole measurement	(Jinchao et al., 2021)
Visible + infrared thermography	Asphalt pavement	Combined RGB–IR fusion to exploit temperature contrasts at crack locations	(Liu et al., 2022)

Datasets are often modest in size (hundreds to a few thousand labeled images) with pixel-wise annotations, which constrains model generalization and motivates transfer learning and data fusion [1–3] (Ren et al., 2020; Chen et al., 2023; Sun et al., 2022; Zhang & Zhang, 2023; Iraniparast et al., 2023; Huyan et al., 2022).

Preprocessing and Classical Image Processing

Pre-deep-learning pipelines relied on thresholding, filtering, morphological operations, skeletonization and related techniques to enhance crack contrast and extract handcrafted features (Golding et al., 2022; Arafin et al., 2023; Cao et al., 2020; Zhang & Zhang, 2023; Huyan et al., 2022). These approaches remain relevant as preprocessing modules or post-processing refinements, but suffer under noise, shadows, and low-contrast textures (Cao et al., 2020; Zhang & Zhang, 2023; Huyan et al., 2022; Kang et al., 2020; Fan et al., 2022).

Image preprocessing for deep models includes grayscale conversion, thresholding, edge detection, wavelet-based multiresolution analysis, and color–depth fusion (Golding et al., 2022; Zhang & Zhang, 2023; Iraniparast et al., 2023; Jinchao et al., 2021; Liu et al., 2022). Golding et al. showed that grayscale inputs to VGG16 achieved F1 scores comparable to RGB, suggesting that CNN-based crack detection does not rely strongly on color information, whereas thresholded and edge-detected inputs reduced performance (Golding et al., 2022). In contrast, a YOLOv8-based detector benefitted from retaining RGB color-based contrast and texture, with RGB outperforming grayscale and binarized inputs across multiple

datasets (Fan et al., 2025). Binarization generally degraded accuracy unless applied to a balanced dataset (Fan et al., 2025).

Wavelet-based multiresolution analysis has been used after deep classification to segment crack pixels, achieving F1 ≈ 95% and demonstrating stable segmentation across metrics (Iraniparast et al., 2023). When color and depth are jointly available, fusing modalities—such as overlapping color and depth or combining visible and infrared images—improves robustness under variable backgrounds and lighting (Jinchao et al., 2021; Liu et al., 2022). Fusion images showed more stable performance than visible-only inputs, especially when cracks resemble their background (Liu et al., 2022). Overall, preprocessing choices interact strongly with sensor type and network architecture; preserving textural richness tends to improve learning-based detectors, while aggressive binarization is detrimental except in specific balanced settings (Golding et al., 2022; Fan et al., 2025; Iraniparast et al., 2023; Jinchao et al., 2021; Liu et al., 2022).

Segmentation Approaches and Methodology Comparison

Pixel-level crack detection is predominantly treated as a semantic segmentation problem, often using encoder–decoder CNNs or hybrid CNN–transformer models (Zhang et al., 2023; Chen et al., 2020; Ren et al., 2020; Chen et al., 2023; Sun et al., 2022; Zhang & Zhang, 2023; Huyan et al., 2022; Fan et al., 2022). Architectures differ in depth, receptive field design, attention mechanisms, and multi-scale feature fusion.

Table 2 Major segmentation architectures and indicative performance

Approach / model family	Key idea and design focus	Representative reported performance	Citations
U-Net and variants (VGGCrackU-net, ResCrackU-net, EfficientNetB3-U-Net, ResNet backbones)	Encoder–decoder with skip connections; multi-scale context via down/up-sampling	F1 up to ≈95–96% for pavement cracks; outperform FCN and PSPNet on asphalt pixels	(Chen et al., 2020; Ren et al., 2020; Nyathi et al., 2024; Huyan et al., 2022; Fan et al., 2022)
DeepLabv3+ with multi-scale attention (DMA-Net)	Atrous spatial pyramid pooling plus attention in decoder; dynamic weighting of multi-scale features	State-of-the-art mIoU and F1 on Crack500, DeepCrack, FMA, and smartphone images	(Sun et al., 2022)
Lightweight real-time CNN (ECSNet)	Small kernels, parallel pooling–conv paths for accelerated segmentation	F1 ≈84.5%, IoU ≈73.1 with highest FPS among compared models	(Zhang et al., 2023)
Multiscale attention MANet	MobileNet encoder with depthwise separable convs and hybrid attention	mIoU ≈0.72–0.78 on Crack500 and CFD; robust on local datasets	(Chen et al., 2023)
CrackSegNet and similar dilated-FCNs	FCN with dilated conv, spatial pyramid pooling, skip connections	Higher accuracy and generalization than conventional image processing methods	(Ren et al., 2020)
Transformer–CNN hybrids (Swin-based encoder + attention	Long-range self-attention with local	Transformer models show higher accuracy but greater memory cost; SwinUNet	(Zhang & Zhang, 2023; Xiao et al.,

decoder; TransUNet, SwinUNet, MTUNet)	decoding; focus on thin and complex cracks	often best among nine tested models	2023; Guo et al., 2023)

U-Net-based models remain a strong baseline, with backbones such as VGG19, ResNet34/50 and EfficientNetB3 enabling high F1 scores for both cracks and spalling (Chen et al., 2020; Ren et al., 2020; Nyathi et al., 2024; Fan et al., 2022). DeepLabv3+ enhancements with multi-scale attention (DMA-Net) provide state-of-the-art results on several benchmarks, demonstrating the value of dynamic weighting between low- and high-level features (Sun et al., 2022).

Lightweight architectures such as ECSNet and MANet introduce small kernels, depthwise separable convolutions, and mobile backbones to achieve real-time segmentation while keeping accuracy competitive (Zhang et al., 2023; Chen et al., 2023). These models trade some absolute performance for efficiency, but are critical for deployment on inspection vehicles or embedded systems.

Transformer-based methods, including Swin Transformer encoders with attention-rich decoders, exploit global context to capture long and tortuous cracks in noisy backgrounds (Zhang & Zhang, 2023; Xiao et al., 2023; Guo et al., 2023). Comparative experiments indicate that transformer variants tend to converge more easily and yield higher accuracy than pure CNNs, though at the cost of higher memory usage and slower inference (Zhang & Zhang, 2023; Xiao et al., 2023). CrackFormer's hybrid-window self-attention and weighted multi-head design attain precision, recall, and F1 near 0.94 across seven datasets, underlining the benefit of local–global feature fusion in challenging pavement scenes (Xiao et al., 2023).

Machine Learning and Deep Learning Paradigms

Early machine learning approaches combined handcrafted features derived from thresholding, edge detection, and morphology with classifiers such as support vector machines or shallow neural networks (Arafin et al., 2023; Cao et al., 2020). These methods struggle with generalization across varying textures and illumination.

Deep learning has transformed crack detection, with three main paradigms: (i) image-level classification (crack vs. non-crack or defect type), (ii) patch-level or object detection, and (iii) pixel-level segmentation [1–3][5–7] (Nyathi et al., 2024)[13–16] (Fan et al., 2022).

Classification networks employing VGG, ResNet, Inception, MobileNet, and Xception backbones achieve very high accuracy (often above 95–99%) in distinguishing cracked from intact concrete (Golding et al., 2022; Meftah et al., 2024; Nyathi et al., 2024; Iraniparast et al., 2023; Fan et al., 2022). For instance, a study combining a Random Forest classifier with MobileNet/InceptionV3/Xception reached test accuracy and F1 around 99.9% on 30,000+ pavement images, highlighting that binary crack recognition is largely solved at image level under controlled datasets (Meftah et al., 2024). However, such models do not directly provide crack geometry.

Object detection frameworks, including Faster R-CNN and YOLO variants, localize crack regions with bounding boxes, supporting region-of-interest extraction for subsequent segmentation or measurement (Meftah et al., 2024; Fan et al., 2025; Kang et al., 2020; Fan et al., 2022). Faster R-CNN combined with a modified tubularity flow field segmentation and distance transform achieved 95% average precision and IoU ≈83% for pixel-level segmentation, while enabling crack length and thickness estimation with ≈93% accuracy (Kang et al., 2020). YOLOv8-based detectors have been used to study the effects of preprocessing and dataset balance on detection performance, showing that RGB inputs and balanced class distributions significantly improve results (Fan et al., 2025).

Semantic segmentation models, as discussed earlier, provide the most detailed representation, enabling direct use for structural assessment and distress quantification (Zhang et al., 2023; Chen et

al., 2020; Ren et al., 2020)[8–11] (Huyan et al., 2022)[18–20]. Some pipelines integrate classification, segmentation, and measurement stages: for example, a custom CNN plus U-Net plus laser-based calibration achieved 99.2% and 96.5% accuracy for classification and segmentation respectively, while measuring crack width with a mean absolute error of 0.16 mm (Nyathi et al., 2024).

CNN Architectures

Many works adopt pre-trained CNNs as encoders or classifiers, leveraging transfer learning from ImageNet to compensate for limited crack datasets [1–3][6–8] (Nyathi et al., 2024; Iraniparast et al., 2023; Huyan et al., 2022)[18–20]. Popular choices include VGG16/19, ResNet families, InceptionV3, EfficientNetB3, MobileNet, and custom residual networks such as Parallel ResNet [1–3][6–8] (Nyathi et al., 2024)[13–15] (Fan et al., 2022; Liu et al., 2022).

VGG-type encoders are frequently used for SegNet-like and U-Net-like models (e.g., PCSN, VGGCrackU-net), enabling semantic segmentation with pixel-wise supervision [1–3] (Huyan et al., 2022). Residual architectures, including ResNet34/50 backbones and bespoke Parallel ResNet configurations, improve gradient flow and allow deeper networks that handle complex crack patterns with better robustness to noise (Chen et al., 2020; Ren et al., 2020; Huyan et al., 2022; Fan et al., 2022). Parallel ResNet in particular uses parallel residual branches and morphological post-processing to suppress noise and extract crack skeletons, achieving F1 ≈93–96% on standard datasets and enabling measurement of crack length, width, and area (Fan et al., 2022).

Mobile-oriented backbones (MobileNet, depthwise separable convolutions) are designed for reduced parameter counts and inference time, as in MANet and ECSNet, while preserving adequate representational capacity (Zhang et al., 2023; Chen et al., 2023). These designs are particularly relevant for pavement survey vehicles and autonomous platforms, where computational resources may be limited (Zhang et al., 2023; Meftah et al., 2024; Chen et al., 2023; Jinchao et al., 2021).

Transformer-Based Models and Attention Mechanisms

Transformer-based architectures are emerging as a strong alternative or complement to CNNs in pavement crack segmentation (Chen et al., 2023; Sun et al., 2022; Zhang & Zhang, 2023; Xiao et al., 2023; Guo et al., 2023). Their core advantage is the ability to model long-range dependencies and global context, which is crucial for thin, meandering cracks amidst complex backgrounds.

Swin Transformer encoders, when combined with attention-equipped decoders such as UperNet or UNet-like structures, have been shown to outperform CNN baselines in terms of segmentation accuracy and convergence stability (Zhang & Zhang, 2023; Guo et al., 2023). Transformer-based models, including TransUNet, SwinUNet, and MTUNet, generally reach higher accuracy but require more memory and exhibit lower processing speed than CNN counterparts (Zhang & Zhang, 2023).

Hybrid-window attention, as in CrackFormer, uses dense local windows and sparse global windows to simultaneously capture fine details and large-scale patterns, further refined by weighted multi-head self-attention that emphasizes the most informative attention heads (Xiao et al., 2023). Multi-scale attention modules added to DeepLabv3+ or MobileNet encoders selectively weight features from different layers and channels, improving aggregation of crack cues across scales and enhancing the delineation of fine crack structures (Chen et al., 2023; Sun et al., 2022).

Overall, attention mechanisms—whether embedded in CNNs or transformers—consistently improve pixel-level crack segmentation, particularly under noisy conditions and for small or low-contrast cracks (Chen et al., 2023; Sun et al., 2022; Zhang & Zhang, 2023; Xiao et al., 2023; Guo et al., 2023).

Performance Evaluation Metrics and Comparative Insights

Performance is commonly assessed using accuracy, precision, recall, F1 score, and intersection over union (IoU) or mean IoU, with some studies additionally reporting latency (FPS), training time, and computational complexity.

Classification models often report accuracies above 90–99%, reflecting the relative simplicity of image-level crack vs. non-crack discrimination in curated datasets (Golding et al., 2022; Meftah et al., 2024; Nyathi et al., 2024; Iraniparast et al., 2023; Fan et al., 2022). Segmentation metrics are more challenging; state-of-the-art systems typically achieve F1 scores in the mid-80s to mid-90s and IoU or mIoU values around 0.7–0.8 on standard pavement datasets (Zhang et al., 2023; Chen et al., 2020; Ren et al., 2020; Chen et al., 2023; Sun et al., 2022; Zhang & Zhang, 2023; Huyan et al., 2022; Xiao et al., 2023)[18–20].

ECSNet illustrates explicit accuracy–efficiency trade-offs: while not achieving the absolute top F1 score (≈84.5%), it offers the highest FPS and lowest training time among compared segmentation models, making it attractive for real-time applications (Zhang et al., 2023). DMA-Net and attention-based transforms deliver higher mF1 and mRecall, but at greater computational cost (Sun et al., 2022; Zhang & Zhang, 2023; Xiao et al., 2023; Guo et al., 2023). Studies that jointly evaluate accuracy, parameter count, and inference time (e.g., FPN on visible/IR fusion, mobile attention networks) highlight that optimal choices depend on deployment constraints, not solely on peak accuracy (Zhang et al., 2023; Chen et al., 2023; Liu et al., 2022).

Challenges, Limitations, and Open Issues

Despite substantial advances, several challenges persist. Many datasets are relatively small, imbalanced, or limited to specific crack types and environmental conditions, which restricts generalization to unseen pavements (Chen et al., 2020; Meftah et al., 2024)[10–13][18–20]. Dataset imbalance can bias detectors toward dominant crack categories; an analysis with YOLOv8 showed that balanced datasets yield more stable and generalized predictions than imbalanced ones, and that dataset balance exerts a stronger influence on performance than preprocessing alone (Fan et al., 2025).

Environmental variability remains problematic: shadows, oil stains, markings, rough textures, and lighting changes induce false positives and false negatives, even for advanced networks (Cao et al., 2020; Zhang & Zhang, 2023; Huyan et al., 2022; Xiao et al., 2023; Fan et al., 2022; Liu et al., 2022). Some methods explicitly incorporate complex backgrounds and disturbances in dataset design, yet generalization across regions, pavement materials, and imaging systems is still only partially addressed (Meftah et al., 2024; Zhang & Zhang, 2023; Huyan et al., 2022; Jinchao et al., 2021).

Another limitation is the gap between pixel-level segmentation performance and structural assessment needs. While several works perform crack quantification (length, width, area) and even volume estimation for potholes using 3D stereo or calibrated laser/IR setups, robust, standardized pipelines that convert segmentation maps into reliable condition indices for network-level pavement management are still emerging (Nyathi et al., 2024; Kang et al., 2020; Fan et al., 2022; Jinchao et al., 2021).

Computational efficiency is an ongoing concern. High-performing transformer-based and deep attention networks demand significant memory and processing power, which can be incompatible with low-cost, vehicle-mounted, or edge devices (Zhang et al., 2023; Chen et al., 2023; Zhang & Zhang, 2023; Xiao et al., 2023; Guo et al., 2023). Lightweight models reduce complexity but may underperform in highly noisy or rare-case scenarios.

Finally, a lack of unified benchmarks and standardized evaluation protocols across pavement datasets hampers direct comparison of methods. Although some works test multiple networks on shared public datasets, variation in preprocessing, training–testing splits, and tolerance thresholds (e.g., 0-pixel vs. relaxed matching) complicates meta-analyses (Zhang et al., 2023; Chen et al., 2020; Cao et al., 2020; Chen et al., 2023; Sun et

al., 2022; Zhang & Zhang, 2023; Huyan et al., 2022; Xiao et al., 2023; Fan et al., 2022).

Future Research Directions

Several promising directions emerge from current trends. First, constructing larger, more diverse, and better-balanced benchmark datasets for concrete pavements, possibly via multi-agency collaborations, would improve generalization and support fairer comparisons (Chen et al., 2020; Meftah et al., 2024)[10–13][18–20]. Including multi-modal data (RGB, depth, infrared, LiDAR) and standardized pixel-wise labels with crack severity metadata would enable models that jointly estimate presence, geometry, and severity (Jinchao et al., 2021; Liu et al., 2022).

Second, integrating advanced data augmentation, class-rebalancing, and domain adaptation techniques could alleviate dataset dependence, particularly for rare crack types or extreme illumination conditions [11–13] (Xiao et al., 2023; Guo et al., 2023). YOLOv8-based findings suggest that addressing class imbalance is as important as refining preprocessing (Fan et al., 2025).

Third, further exploration of efficient transformer variants and attention mechanisms tailored to pavement imagery may yield architectures that approach transformer-level accuracy with CNN-level efficiency. Hybrid CNN–transformer models with hierarchical tokenization, sparse attention, and dynamic pruning represent one pathway (Chen et al., 2023; Sun et al., 2022; Zhang & Zhang, 2023; Xiao et al., 2023; Guo et al., 2023).

Fourth, tighter coupling between segmentation and structural assessment is needed. Future systems should propagate uncertainty from detection through to indices such as crack density, severity, and remaining service life, potentially using probabilistic deep learning and physics-informed post-processing (Nyathi et al., 2024; Kang et al., 2020; Fan et al., 2022; Jinchao et al., 2021). Finally, deployment-focused research involving real-time vehicle-based inspection systems, continuous learning from streaming data, and cross-region validation will be crucial to transition from laboratory models to operational pavement management tools (Zhang et al., 2023; Meftah et al., 2024; Chen et al., 2023; Jinchao et al., 2021).

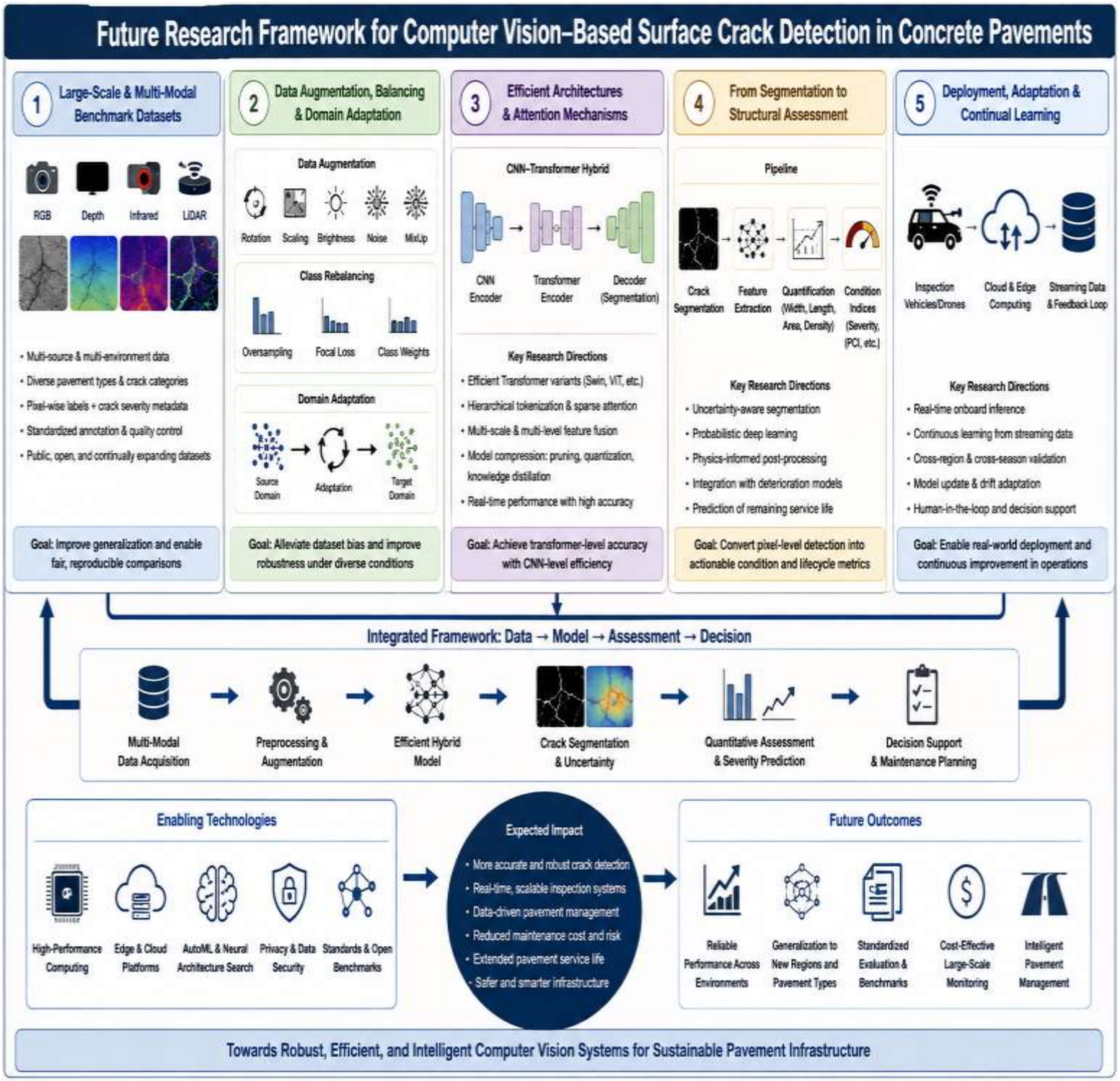

Figure 2 Proposed future research framework for computer vision-based surface crack detection in concrete pavements, highlighting integrated directions including large-scale multimodal datasets, data augmentation and domain adaptation, efficient CNN–transformer hybrid architectures, uncertainty-aware structural assessment, and real-time deployment systems for intelligent pavement management applications.

Conclusion

Recent research on computer vision-based surface crack detection in concrete pavements shows a clear evolution from classical image processing and handcrafted-feature machine learning toward deep CNNs and, increasingly, transformer-based architectures. High-performing segmentation networks, particularly U-Net variants, DeepLabv3+ with attention, and transformer–CNN hybrids, achieve strong F1 and IoU scores on public pavement datasets and real-world images, while lightweight models such as ECSNet and MANet address real-time requirements. Multi-modal sensing (stereo, depth, infrared) and fusion strategies further enhance robustness for crack segmentation and quantification.

Key limitations remain in dataset size and diversity, class imbalance, environmental robustness, computational cost, and the translation of pixel-level predictions into actionable structural health metrics. Future work that combines richer datasets, balanced and adaptive learning strategies, efficient attention-based architectures, and end-to-end pipelines for condition assessment will be central to maturing automated pavement crack detection into a widely adopted component of modern pavement management and infrastructure health monitoring.

References

Arafin, P., Billah, A., & Issa, A. (2023). Deep learning-based concrete defects classification and detection using semantic segmentation. *Structural Health Monitoring, 23*, 383 - 409. https://doi.org/10.1177/14759217231168212

Cao, W., Liu, Q., & He, Z. (2020). Review of Pavement Defect Detection Methods. *IEEE Access, 8*, 14531-14544. https://doi.org/10.1109/access.2020.2966881

Chen, J., Wen, Y., Nanehkaran, Y., Zhang, D., & Zeb, A. (2023). Multiscale Attention Networks for Pavement Defect Detection. *IEEE Transactions on Instrumentation and Measurement, 72*, 1-12. https://doi.org/10.1109/tim.2023.3298391

Chen, T., Cai, Z., Zhao, X., Chen, C., Liang, X., Zou, T., & Wang, P. (2020). Pavement crack detection and recognition using the architecture of segNet. *J. Ind. Inf. Integr., 18*, 100144. https://doi.org/10.1016/j.jii.2020.100144

Fan, Z., Lin, H., Li, C., Su, J., Bruno, S., & Loprencipe, G. (2022). Use of Parallel ResNet for High-Performance Pavement Crack Detection and Measurement. *Sustainability*. https://doi.org/10.3390/su14031825

Fan, L., Tang, S., Ariffin, M., Ismail, M., & Wang, X. (2025). Impact of Image Preprocessing and Crack Type Distribution on YOLOv8-Based Road Crack Detection. *Sensors (Basel, Switzerland), 25*. https://doi.org/10.3390/s25072180

Golding, V., Gharineiat, Z., Munawar, H., & Ullah, F. (2022). Crack Detection in Concrete Structures Using Deep Learning. *Sustainability*. https://doi.org/10.3390/su14138117

Guo, F., Liu, J., Lv, C., & Yu, H. (2023). A novel transformer-based network with attention mechanism for automatic pavement crack detection. *Construction and Building Materials*. https://doi.org/10.1016/j.conbuildmat.2023.131852

Huyan, J., , T., Li, W., Yang, H., & Xu, Z. (2022). Pixelwise asphalt concrete pavement crack detection via deep learning-based semantic segmentation method. *Structural Control and Health Monitoring, 29*. https://doi.org/10.1002/stc.2974

Iraniparast, M., Ranjbar, S., Rahai, M., & Nejad, M. (2023). Surface concrete cracks detection and segmentation using transfer learning and multi-resolution image processing. *Structures*. https://doi.org/10.1016/j.istruc.2023.05.062

Jinchao, G., Yang, X., Ling, D., Cheng, X., Lee, V., & Jin, C. (2021). Automated pixel-level pavement distress detection based on stereo vision and deep learning. *Automation in Construction, 129*, 103788. https://doi.org/10.1016/j.autcon.2021.103788

Kang, D., Benipal, S., Gopal, D., & Cha, Y. (2020). Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning. *Automation in Construction, 118*, 103291. https://doi.org/10.1016/j.autcon.2020.103291

Liu, F., Liu, J., & Wang, L. (2022). Asphalt Pavement Crack Detection Based on Convolutional Neural Network and Infrared Thermography. *IEEE Transactions on Intelligent Transportation Systems, 23*, 22145-22155. https://doi.org/10.1109/tits.2022.3142393

Meftah, I., Hu, J., Asham, M., Meftah, A., Zhen, L., & Wu, R. (2024). Visual Detection of Road Cracks for Autonomous Vehicles Based on Deep Learning. *Sensors (Basel, Switzerland), 24*. https://doi.org/10.3390/s24051647

Nyathi, M., Bai, J., & Wilson, I. (2024). Deep Learning for Concrete Crack Detection and Measurement. *Metrology*. https://doi.org/10.3390/metrology4010005

Ren, Y., Huang, J., Hong, Z., Lu, W., Yin, J., Zou, L., & Shen, X. (2020). Image-based concrete crack detection in tunnels using deep fully convolutional networks. *Construction and Building Materials, 234*, 117367. https://doi.org/10.1016/j.conbuildmat.2019.117367

Sun, X., Xie, Y., Jiang, L., Cao, Y., & Liu, B. (2022). DMA-Net: DeepLab With Multi-Scale Attention for Pavement Crack Segmentation. *IEEE Transactions on Intelligent Transportation Systems, 23*, 18392-18403. https://doi.org/10.1109/tits.2022.3158670

Xiao, S., Shang, K., Lin, K., Wu, Q., Gu, H., & Zhang, Z. (2023). Pavement crack detection with hybrid-window attentive vision transformers. *Int. J. Appl. Earth Obs. Geoinformation, 116*, 103172. https://doi.org/10.1016/j.jag.2022.103172

Zhang, Y., & Zhang, L. (2023). Detection of Pavement Cracks by Deep Learning Models of Transformer and UNet. *IEEE Transactions on Intelligent Transportation Systems, 25*, 15791-15808. https://doi.org/10.1109/tits.2024.3420763

Zhang, T., Wang, D., & Lu, Y. (2023). ECSNet: An Accelerated Real-Time Image Segmentation CNN Architecture for Pavement Crack Detection. *IEEE Transactions on Intelligent Transportation Systems, 24*, 15105-15112. https://doi.org/10.1109/tits.2023.3300312