Scilight

Bulletin of Computational Intelligence
https://www.sciltp.com/journals/bci

Review

Check for updates

Application of Deep Learning for Pavement Monitoring: Movement towards Autonomous Future

Mohak Desai [1], Kaustav Chatterjee [2,*] and Joshua Li [2]

[1] Independent Researcher, Vadodara 390018, Gujarat, India
[2] School of Civil and Environmental Engineering, Oklahoma State University, Stillwater, OK 74078, USA
* Correspondence: kaustav.chatterjee10@okstate.edu

How To Cite: Desai, M.; Chatterjee, K.; Li, J. Application of Deep Learning for Pavement Monitoring: Movement towards Autonomous Future. *Bulletin of Computational Intelligence* **2026**, *2*(3), 235–251. https://doi.org/10.53941/bci.2026.100013

Received: 7 June 2026
Revised: 6 August 2026
Accepted: 10 August 2026
Published: 27 August 2026

Abstract: Pavements are essential elements of transportation networks and are instrumental to the growth and development of a country. To harness the full potential of this infrastructure, it is necessary to maintain it in proper structural and functional conditions. Conventional techniques, such as manual inspection can be used to determine the functional conditions of the pavement. However, these techniques have different shortcomings, including high monitoring costs, time consumption and requirement of traffic control during pavement surveying. These shortcomings can be mitigated by utilizing various Artificial Intelligence (AI) techniques such as machine learning and deep learning, for pavement surface inspection. This study aims to systematically review state-of-the-art deep learning techniques such as vision transformer model for pavement distress detection. This research contributes to the knowledge body by summarizing the application of vision transformer model for pavement distress detection. Moreover, this research provides a brief description of the application of other deep learning architectures including You Only Look Once (YOLO), and Convolutional Neural Networks (CNNs) for pavement distress detection. Deep learning techniques can autonomously detect various types of pavement distress including longitudinal and transverse cracks, rutting, faulting, patching, shoving, raveling and potholes from the pavement surface. The research was performed in three different steps namely: keyword selection, identification of databases, literature acquisition and summarizing the literature based on different deep learning models. The findings from the review indicate that YOLO and CNN were extensively employed by researchers, however in recent times, vision transformers gained popularity among researchers and pavement engineers. Overall, this study highlights the critical role played by different deep learning techniques in transforming pavement monitoring, leading to safer, more resilient, and sustainable transportation infrastructure.

Keywords: transformer; CNN; YOLO; pavement distress; deep learning

1. Introduction

Pavements are crucial components of the transportation infrastructure that directly affect safety, mobility, and economic development. Maintaining pavement's structural quality is essential to ensure a long service life and reduce maintenance costs. Conventionally, pavement condition monitoring relies on manual inspection and conventional surveying methods, which are labor and cost-intensive and cause traffic disruption. These limitations have driven the adoption of automatic and data-driven techniques for pavement assessment and monitoring.

Pavement evaluation is generally divided into two different components including structural parameters and functional parameters. Structural parameters define the capacity of a pavement to carry load from vehicular traffic

and are generally represented by different parameters of the deflection basin such as D0, D300, and other parameter including effective structural number and tensile strain at the bottom of the asphalt layer. Generally, structural parameters are determined by using different devices such falling weight deflectometer and traffic speed deflectometer. Functional parameters are the metrics which govern the ride quality, safety, and driver comfort. The functional parameters are measured by international roughness index, rut depth, pavement macro-texture and cracking on the surface of the pavement.

Recent progress in Artificial Intelligence (AI), especially in deep learning, has revolutionized distress detection methods by introducing accurate, efficient, and scalable solutions. Deep learning models including You Only Look Once (YOLO), Convolutional Neural Network (CNN), and Vision Transformer (ViT) have showcased extraordinary performance in identifying pavement distress such as longitudinal and transverse cracks [1], rutting, faulting, patching, shoving, raveling and potholes from 2D and 3D images acquired using different sensors [2]. CNNs have been a foundation for computer vision tasks, while YOLO is suitable for on-site monitoring due to its real-time object detection capabilities. Transformers have evolved into a powerful technique for capturing global context and outperforming traditional methods in complex conditions by leveraging self-attention mechanism [3–6].

1.1. Problem Statement and Methodology

This study was performed to demonstrate the application of vision-transformer architecture and some conventional architecture such CNN and YOLO for pavement distress detection. The objectives of this paper are (1) to highlight the application, benefits, and limitations of using transformer architecture for pavement distress detection, (2) outline the shortcomings in existing research work, and (3) put forward future application of these deep learning frameworks for pavement distress detection.

This review paper contributes to the knowledge base of civil engineering by addressing three different questions: (1) how ViT architecture can enhance assessment of pavement distress? (2) What are the different challenges faced by researchers in developing ViT-based framework? (3) How ViT can be integrated with modern technologies for pavement assessment?

This study briefly introduces the concept of ViT architecture and deep learning. It presents an overview of the transformer architecture and its application of pavement distress detection. Moreover, some of the previous deep learning algorithms such as CNN and YOLO were also covered in this paper. This study will facilitate civil engineering engineers and researchers to gain understanding about the process of automated pavement distress detection, the level of accuracy achieved by different methods and different distress detected by different types of algorithms.

This literature review paper commenced with a systematic identification of research paper following a systematic methodology. The researchers used a pair of different keywords during the process of literature acquisition. The pair of keywords include transformer architecture and pavement distress, transformer architecture and pavement monitoring, pavement cracking and transformer architecture, YOLO and pavement distress, and CNN and pavement distress.

The literature was acquired from different databases, including ScienceDirect, American Society of Civil Engineers database, Institute of Electrical and Electronics Engineers (IEEE) database, and other databases. From the literature search, around 25 papers were obtained, which demonstrated the application of ViT architecture for pavement monitoring. The information about the application of YOLO and CNN for pavement distress detection were obtained from different review papers. After literature acquisition, the literature were organized into three different categories: (a) application of ViT architecture for pavement distress detection, (b) application of CNN for pavement distress detection and (c) application of YOLO for pavement distress detection. This framework of literature search proved to be an effective methodology for showcasing the application of ViT architecture for pavement distress detection.

1.2. Pavement Distress

Generally, there are two types of pavements namely: (a) flexible pavement and (b) rigid pavement. Flexible pavements are constructed using asphalt material as the top layer of the pavement. These pavements transfer load to the underlying layers using grain-to-grain contact and consist of four different layers namely: surface layer, base layer, subbase layer and subgrade layer. These pavements require different types of maintenance actions over their lifespan due to cracking and other distresses. Rigid pavements are constructed using concrete surfaces as the top layer of the pavement. These pavements transfer load using slab action to underneath layers. These pavements require less maintenance action over their lifespan. Figure 1 shows two different types of pavements.

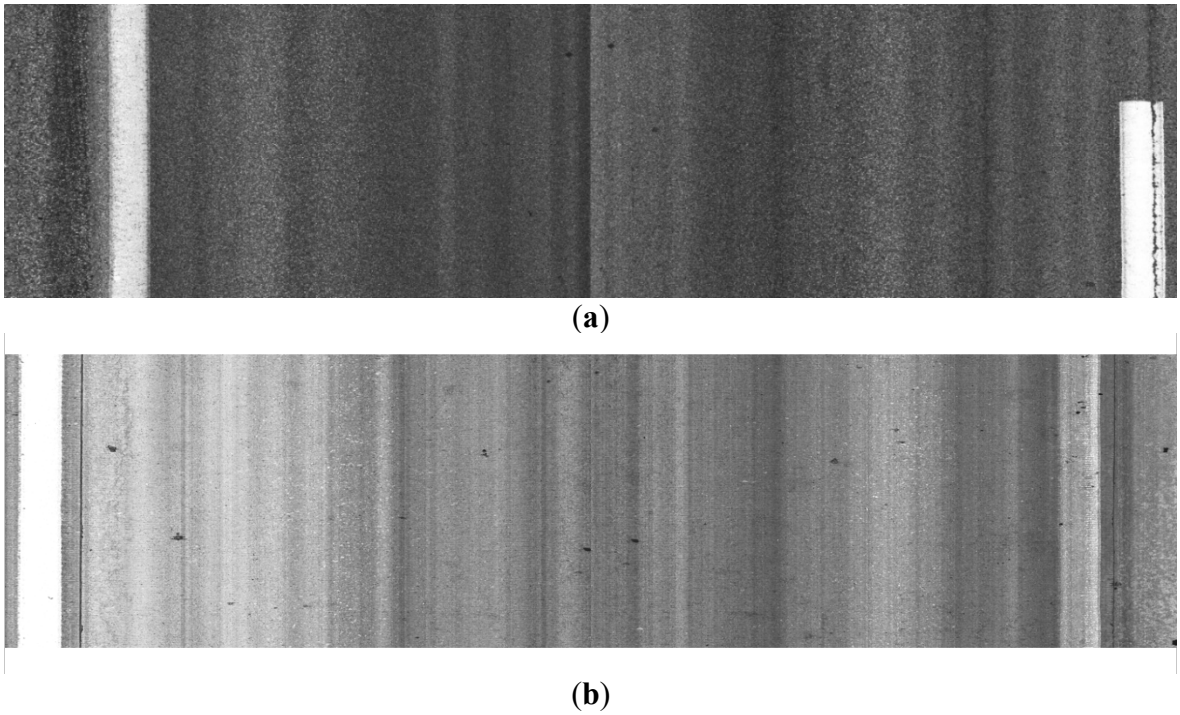

Figure 1. Pavements: (**a**) flexible pavement, (**b**) rigid pavement.

Pavement distress refers to any deterioration in a pavement system that degrades its structural capacity, service life, and performance. In other words, any damage that appears on the pavement surface or within pavement layers that indicates the pavement surface is failing is referred to as a pavement distress. It involves cracks, surface defects, deformations, and moisture-related problems that may occur due to traffic loading, material aging, environmental effects, or construction deficiencies. Table 1 provides brief description about the different types of pavement distresses.

1.3. Artificial Intelligence

Artificial intelligence (Asteris, [7]) is the domain of computer science that deals with development of systems to perform tasks that are generally performed by humans and require human intelligence. In other words, with the help of AI, computers can think, learn and make decisions, similar to humans actions. AI systems are built to understand language, recognize images, solve problems and to learn patterns from the data. The pipeline of AI consists of data collection, data processing, model training, evaluation, model deployment monitoring and retraining. There are five principle technical building blocks based on which AI works, such as Machine Learning (ML), Deep Learning (DL), Natural Language Processing, and Computer Vision Systems.

1.4. Machine Learning

ML is a subdivision of AI that facilitates computers to learn patterns from data and estimate predictions and perform decision-making without programming for each task. Instead of writing rules manually, ML models search for rules by exploring data. There are three types of ML, such as supervised learning, unsupervised learning, and reinforcement learning. There are various types of ML problems, such as classification, regression, clustering, outlier detection, ranking problems, recommendation tasks, time-series forecasting, natural language processing task and generative modeling. Some of the common machine learning algorithms are decision tree, artificial neural network, random forest algorithm, gradient boosting algorithm, extreme gradient boosting algorithm and support vector machine [1,2,8–15].

1.5. Deep Learning

Deep learning (DL) is a subset of ML, and uses multi-layered neural networks to understand complex patterns using large volumes of data. DL outperforms conventional AI techniques in image and speech recognition, natural language processing and self-driving systems. It excels due to deep networks, ability to learn hierarchical features as well as optimization advancements. Deep learning automatically learns features from raw data, without the

requirement of manual feature engineering. Some of the common deep learning architectures are CNNs), Recurrent Neural Networks, and Transformers, etc.

Table 1. Types of pavement distresses.

Pavement	Distress	Description
Flexible pavement	Alligator/Fatigue Cracking	Resulted from traffic loads, most common type of distress in flexible (asphalt) pavements
	Longitudinal Cracking	Cracks forming parallel to the centerline, caused by shrinkage and poor joints. Indicates issues related to construction process, pavement structure, or environmental effects.
	Transverse Cracking	Runs perpendicular to centerline; occurs mostly because of temperature variations, asphalt shrinkage, or construction issues
	Block Cracking	Rectangular cracks resulted from asphalt binder aging or shrinkage, not caused by traffic.
	Rutting	Longitudinal surface depression in wheel paths caused due to subgrade deformation or asphalt deformation.
	Shoving	Appears due to localized bulging due to braking or pushing forces, normally occurring perpendicular to traffic direction. Shoving is caused by instability of the asphalt mix, resulting in pavement material being pushed and displaced because of traffic load.
	Potholes	Bowl-shaped holes formed due to moisture intrusion and traffic load and where pieces of asphalt have broken loose.
	Raveling	Occurs due to loss of aggregate (both fine and coarse) from the pavement surface due to binder aging, poor compaction or disintegration of the asphalt binder–aggregate bond.
	Polishing/Smoothness Loss	Aggregate particles become smooth and rounded because of traffic, which ultimately reduces pavement's skid resistance
	Bleeding	Excess asphalt binder rises to the surface, forming a black, sticky, shiny and smooth film on the pavement.
Rigid Pavement	Longitudinal cracks	Cracks that run parallel to pavement's centerline, may develop within a slab or along a joint.
	Transverse Cracks	Runs perpendicular to the pavement's centerline; appears mostly because of temperature variations and shrinkage.
	Corner cracking	Full depth fractures are developed near corners of a concrete pavement, usually caused by high stresses at the pavement corners.
	Joint spalling	Deterioration of concrete pavement edge near joints, which leads to fragmentation of concrete, cracks, chipping or breaking
	Faulting	Elevation difference between slabs across a joint or crack, mostly because of pumping
	Scaling	Surface level concrete pavement distress where the uppermost layer of the concrete peels, flakes, or crumbles, generally in small or large patches.
	Popout	Small conical holes that appeared on the surface of a concrete pavement made due to expansion of porous aggregate particles and aggregate reactivity.

2. Evaluation Metrics for Vision-Based Deep Learning Model

Different evaluation metrics are used by researchers for determining the effectiveness of deep learning models for pavement distress monitoring. Some of the common evaluation metrics are accuracy, precision, F-1 score, and recall. Figure 2 shows a schematic representation of the confusion matrix. The different metrics are represented in Equations (1)–(5).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

where,

Accuracy represents the accuracy of the deep learning models
TP = True Positive: Represents cases where the model correctly detects a pavement distress
TN = True Negative: Represents cases where the model correctly detects absence of distress
FP = False Positive: Represents cases where the model reports distress even though none exist
FN = False Negative: Represents cases where the model fails to identify/detect an existing distress

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

Intersection over union (IoU) represents how precisely a model segments pavement distresses such as potholes, cracks, patches, by comparing the predicted region with the actual labeled region. In simple words, IoU facilitates understanding accuracy of object detection. High IoU represents predicted distress area closely matches the ground condition, whereas low IoU represents poor capability of the model to outline cracks. The range of IoU is 0 to 1. Where, 0 means there's no overlap between the predicted box and the real object and 1 means the predicted box overlaps with the real object (6).

$$\text{IoU} = \frac{\text{Area of Overlap/Intersaction}}{\text{Area of Union}} \tag{6}$$

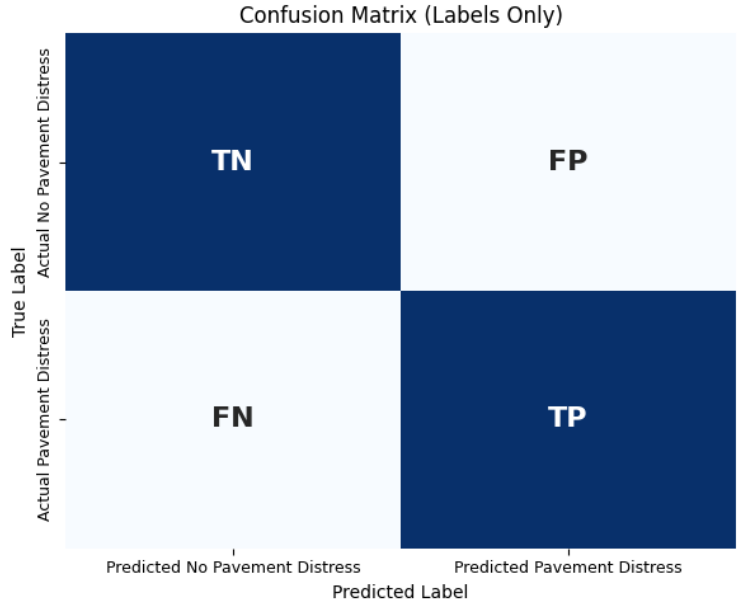

Figure 2. Schematic representation of confusion matrix showing true positive, false positive, true negative, false negative.

3. Applications of Deep Learning in Pavement Monitoring

3.1. Image Preprocessing

Image preprocessing is one of the important steps for developing different deep learning frameworks for automated distress detection. Generally, it is performed in four different steps:

(1) Image normalization: Image normalization is performed to convert the information into depth images or point cloud data for pavement distress 3D reconstruction and quantification with the help of normalization.

(2) Noise removal and correction: Feature pyramid and hierarchical boosting network (FPHBN) can be employed for bifurcating samples and to show robustness in cracks detection under shadow [16].

(3) Image processing: Techniques such as edge detection for partial crack datasets can be utilized for semantic segmentation. Image processing enables high quality of input data for feature extraction; which ultimately gives important information to DL models.

(4) Data Augmentation: Data augmentation includes creation of new data samples from already available data to artificially increase the size of dataset. This technique can be employed to tackle uneven sample distribution and monotonous scenarios, as well as to enhance data quality and diversity.

3.2. Dataset for Model Development

Majority of the studies utilized publicly available datasets such as CRACK500, DeepCrack, CrackTree200, Crack Forest Dataset, AEL, GAPs384, CrackLS315, FMA, Stone331, CrackPV14. On the other hand, Ansari et al., [3] used custom datasets with 2D pavement images collected using OSU 3D laser scanning system. Each image of the dataset was manually labeled with pavement features such as background, crack, pothole, sealed crack, marking, expansion joint, patch, and manhole. After manual image labelling, data augmentation techniques were applied to increase numbers of training dataset by applying random rotation, translation, and scaling. Cano-Ortiz et al., [17] prepared custom dataset named Mosquitonet with 7099 road images and 13 types of defects.

3.3. Convolutional Neural Network (CNN)

CNN is a deep learning algorithm for analyzing visual data, such as images and videos. The CNN architecture consists of three principal components: convolutional layers, max pooling layers, and fully connected layers. Convolutional layers are designed to automatically learn spatial hierarchies of features such as shape, texture, and edges. The max-pooling layer reduces the dimensionality of the input data. CNNs employs local connectivity, shared weights, and hierarchical feature extraction, allowing them to understand spatially invariant representations with considerably fewer parameters than fully connected networks.

Over the past decade, deep learning emerged in pavement distress detection domain, with CNN playing key role in it. Unlike conventional ML algorithms, which require manual feature extraction, CNNs automatically extract features from raw data, making them very efficient for tasks such as image recognition and object detection. Moreover, CNNs are adept at capturing local and global patterns in data, making them a mainstay of computer vision applications.

On this basis, researchers attempted to explore hybrid architectures that combine CNNs with transformer-based models to improve accuracy as well as real-time performance. For instance, Yu et al., [18] used a Swin Transformer model and a lightweight CNN model that achieved high accuracy but low real-time suitability due to high Floating Point Operations (FLOPs). Higher FLOPs indicate more computation and slower inference. Moreover, the study proposes an improved version of Bilateral Segmentation Network v2 (BiSeNetv2) that offers the best trade-off between speed and accuracy, achieving a high recall for fine cracks. Guo et al., [19] integrated a CNN-based classification architecture (ResNet-34) with a transformer-based segmentation architecture (CTv2 using Swin Transformer + SegFormer) to improve pixel-level crack detection. The CTv2 model achieved high precision and recall with real-time inference (110 FPS). Sufficiently handles high-resolution images and data imbalance.

In the same year, Zhang et al., [20] further advanced this trend by developing a mix-graph CrackNet and a hybrid CNN-Transformer architecture with graph-based skip connections and dual attention fusion. The researchers achieved an F1-score of 90.37% and an Intersection over Union (IoU) of 82.43%, excelling state-of-the-art models under noisy, complex pavement conditions.

Continuing this research advancement, Ansari et al., [3] compared CNN and ViT (SegFormer-B4) for multi-class pavement feature detection, achieving high pixel-level accuracy and detecting complex features such as sealed cracks and pavement markings. Data augmentation and high robustness were achieved with limited samples by using Generative Adversarial Networks (GANs). The authors developed the ViT model using transfer learning techniques. Transfer learning is a technique in which a model trained for one task is used for another.

Wu et al., [21] presented Spectrum Focus Transformer (SFT) integrated with ResNet34, a frequency-domain attention model to enhance pavement distress detection. From the experiments, the authors noted higher classification accuracy (97.73%) than ResNet34, ResNet34 + SE, DenseNet, ResNet50, and MobileNet on the benchmark dataset.

Overall, these research works showcase clear progress in the domain from only CNN based models to CNN-transformer hybrids and graph-integrated architectures and spotlight a common trend: integration of strong feature extraction capabilities of CNNs with the global representation power of transformers produces robust, accurate, and real-time pavement distress detection systems reliable for real-world scenarios.

3.4. You Only Look Once (YOLO)

YOLO is a single stage real-time object detection algorithm that uses a single convolutional neural network (CNN) with superior processing speed and accuracy [22] and performs its function by looking at the entire image at once rather than searching through different parts of the image. This makes YOLO suitable for real-time detection applications. YOLO divides an image into a grid and estimates bounding boxes, identifying the object's position and type. This feature makes YOLO quick, efficient, and suitable for real-time jobs such as self-driving cars, and robotics. The architecture of YOLO has evolved through multiple versions (YOLOv1 to YOLOv11) as

shown in Figure 3, ultimately improving performance in terms of speed, accuracy, and the identification of tiny objects. Recent studies have combined YOLO with transformer-based models to further enhance its performance. YOLO is mostly used for real-time object detection jobs such as real-time video analytics and real-time video surveillance. Figure 4 shows the schematic representation of detection of pavement markings using YOLO.

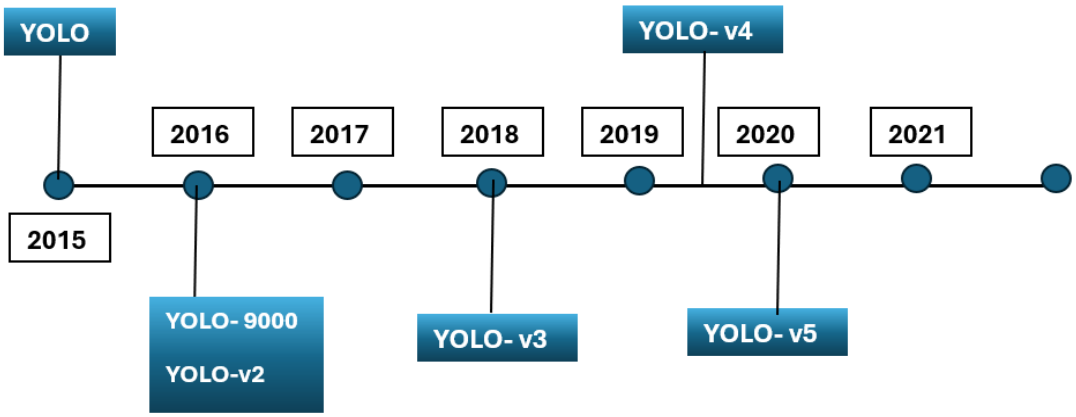

Figure 3. YOLO timeline.

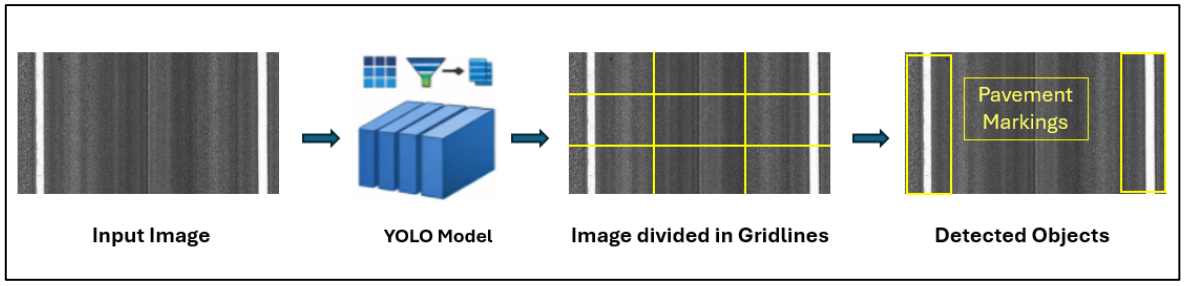

Figure 4. YOLO workflow.

Recent research in vision-based pavement distress detection has adopted hybrid architecture that combines global and local feature modeling. Luo et al., [23] combined YOLO framework with swin transformer, enabling a hybrid global–local feature extraction strategy. Their model showcased notable robustness under occlusion and noise, achieving a Mean Average Precision (mAP) of 63.73%. Moreover, this study also presented a global attention guidance module and α-IoU-NMS to improve multi-scale fusion and minimize false negatives, marking an effort in the direction of transformer-enhanced detection in pavement evaluation.

Progressing from this basis, Wang et al., [24] extended the scope of transformer architectures by combining swin transformers with Squeeze-and-Excitation (SE) attention and an enhanced YOLOv8 framework. This merger provided improved global perception as well as maintained strong local feature representation, which significantly improved models' performance in complicated real-world situations such as severe noise and pothole identification. By employing a synthetic-to-real dataset ratio of 2:1, enhanced YOLOv8 attained mAP of 94.8%, showcasing the effectiveness of transformer–attention hybrids when supported by data-driven augmentation strategies.

In a similar manner, Wu et al., [25] worked on architectural optimization by enhancing YOLOv5 with a crossed feature pyramid network and Wise-IoU v2 loss function. This model achieved MAP of 69.3% with 118 frames per second, outperforming the YOLOv5 baseline while minimizing model parameters by 27.1% and GFLOPs by 24.1%.

In the same year, Yang et al., [26] performed a real-time study on automated trespassing violation detection and tracking framework for highway–rail grade crossings (HRGCs) by employing an enhanced deep learning method. YOLOv4 for object detection and Deep SORT for multi-object tracking were integrated and augmented with a low confidence tracking filter to minimize false positives and identity switches. The model identified 436 trespassing events, with precision of 95.6%, recall of 93.2%, and an F1 score of 94.4%. The authors underlined model's potential integration into real-time warning/alarming systems and need for additional data collection and improved sensing technologies (e.g., thermal imaging) for improved robustness under different environments conditions.

All this research aligns with the work by Lv et al., [27], who conducted a comprehensive review of pavement distress detection methodologies including manual inspection and state-of-the-art AI-driven systems, highlighting algorithms such as pyramid vision transformer, and hybrid models for segmentation and detection. The Pyramid Vision Transformer is an advanced deep learning model for dense predictions, such as object detection and semantic segmentation. Also, the researchers compared YOLO and Faster R-CNN for object detection and highlighted lightweight models such as Efficient CrackNet.

Collectively, all this research showcases a development in pavement distress detection research—from CNN-based designs to transformer-enhanced, hybrid, and lightweight architectures, as the domain leaps forward toward more robust, scalable, and real-time solutions for complicated pavement conditions.

3.5. Vision Transformer

According to Dosovitskiy et al., [28], ViT is a deep learning architecture based on the principles of transformer architecture developed for image analysis. Transformer architecture was developed by Vaswani et al., [29]. Rather than using convolutional layers like a CNN, ViT breaks an image into fixed-size patches, flattens them, and treats each patch as a token, like words in a sentence. Subsequently, the patches are processed through a self-attention mechanism, allowing the model to record long-range dependencies and global context across the whole image. The self-attention technique allows ViT to gain state-of-the-art performance in image classification, especially when trained using large databases. ViT's capability to model global relationships without depending on convolutions creates notable shifts in computer vision research. ViT is applied in image classification, object detection, medical imaging, satellite and remote sensing, semantic/instance segmentation, etc. In recent times, transformer architecture was used for determining the profiles of highway-railway grade crossings [4,5]. Figure 5 shows the schematic representation of pavement feature detection using ViT framework.

The progression of transformer-based models for crack detection has grown in recent years, beginning with hybrid architecture and over time evolving toward lightweight and robust frameworks. Shamsabadi et al., [30] presented an early benchmark by evaluating TransUNet (CNN + ViT Hybrid) for crack segmentation. This hybrid model achieved high accuracy (99.55%) for small and thin crack detection, but with high inference time. Similarly, Tang et al., [31] introduced PicT, a slim weakly supervised ViT with a patch-labeling teacher model for weakly supervised learning, which outperformed Swin Transformers and CNNs. This model achieved high accuracy with faster training and inference, and addressed challenges posed by sparse distress areas and weak supervision. For better local-global feature integration, Xu et al., [32] presented LETNet, inspired by transformer architecture integrating with convolution-based local enhancement modules. The model achieved high IoU (94.07%) for Sthone331 dataset, and good robustness for noisy data. The study highlighted ViT's limitations in local detail modeling and computational overhead, along with improved observation of fine cracks.

Xiao et al., [33] developed a CrackFormer, a hybrid window attentive vision transformer and weighted multi-head self-attention for fine-grained crack segmentation. The model excels different CNN and capsule networks in terms of F1 score. Shortcomings of the model included the requirement for a large dataset. Guo et al., [34] developed a crack transformer model that uses a Swin Transformer encoder and a lightweight MLP-based decoder for crack segmentation. The model achieved high mF1 (94.60) and mRecall (96.41) and demonstrated robust performance with noisy data and complex crack patterns across multiple datasets. Guo et al., [35] proposed the Swin Transformer with attention model, merging the Swin Transformer encoder with the UperNet decoder and Convolutional Block Attention Module for increased feature refinement. This model outperforms ANN, Fully connected network, PSPNet, UNet, and STU (mostly CNN-based computer vision models) in terms of accuracy and recall, and can handle noisy backgrounds and complex crack geometries. Liu et al., [36] conducted a study to upgrade CrackFormer to CrackFormer-II, by adding local self-attention (LSA) and local feed-forward (LFF) layers to improve global-local feature fusion. The model improved segmentation of fine and coarse cracks and achieved state-of-the-art F1 scores (≈0.9) across multiple databases, while reducing FLOPs. Tong et al., [37] combined the Dempster–Shafer Theory (a mathematical framework for combining evidence and handling uncertainty in decision-making) with a transformer network to improve uncertainty calibration and robustness in segmentation. The Evidential Segmentation Transformer (ES-Transformer) model achieved high pixel-level accuracy and mIoU compared to probabilistic models, enabling multi-model fusion without retraining. The study also effectively highlights annotation noise and class imbalance. Zhang et al., [38] developed ShuttleNetV2, merging transformer blocks into a repeatable encoder-decoder structure for pixel-level detection of multiple pavement distresses, including sealed cracks, patches, and potholes. The model achieved 94.21% F1 and an IoU score of 0.8914, outperforming the original ShuttleNet model. Table 2 presents the computational expenses of three different architectures.

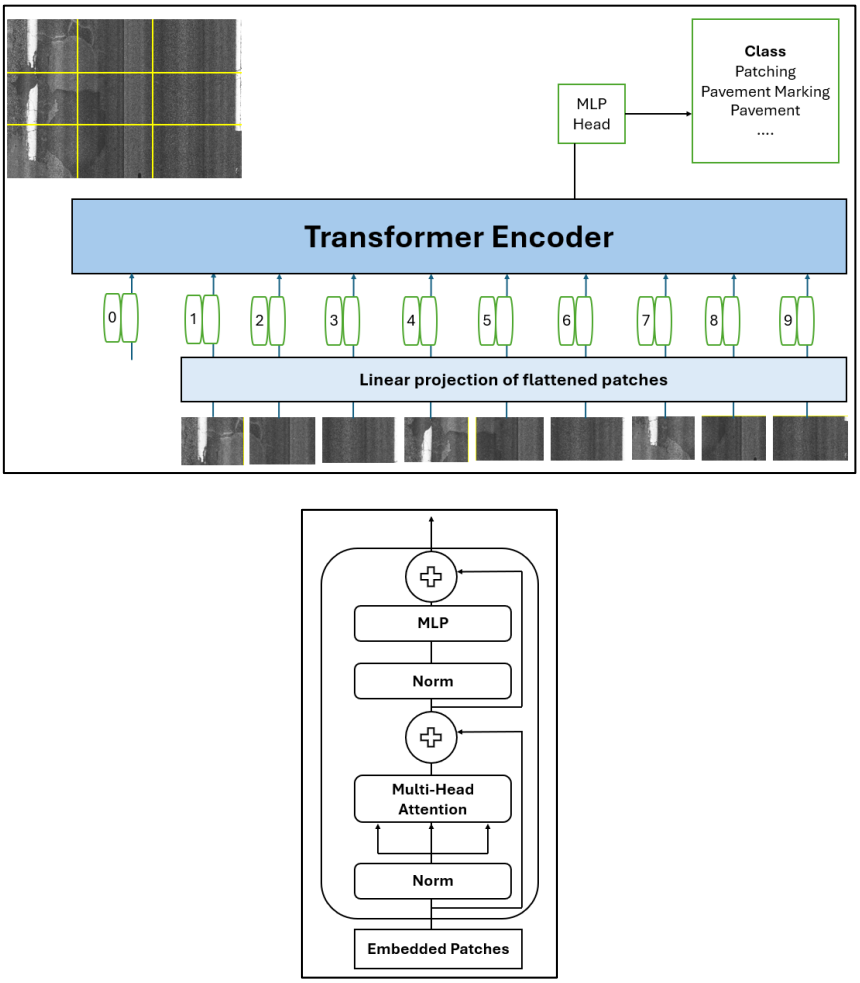

Figure 5. Vision transformer.

Table 2. Computational expenses of different architectures.

Authors	Model/Architecture	Computational Expense (FLOPs)	Inference Speed
Zim et al., [39]	EfficientCrackNet (hybrid CNN + Transformer module)	DeepCrack—15.42 EfficientCrackNet—0.483 ShuffleNetV2—1.105	FPS not mentioned
Yu et al., [18]	iBiSeNetv2	iBiSeNetv2—7.21	iBiSeNetv2—147.20 FPS
	FCN	FCN—115.83	FCN—66.30
	UNet	UNet—118.69	UNet—25.30
	CCNet	CCNet—117.41	CCNet—63.20
	Swin Transformer	Swin Transformer—25.62	Swin Transformer—57.30
Kyem et al., [40]	Context-CrackNet	243.78 GFLOPs	15.63

By the year 2024, the researchers worked on modifying and refining hybrid and lightweight transformer models. Fan et al., [41] used ViT-based, CNNs-based, and hybrid models for pavement defect detection using 2D image data and 3D point cloud-based techniques. Moreover, the study examined the architecture and performance of deep learning for classification and segmentation. Wang et al., [42] introduced SwinCrack, a hybrid U-shaped model that integrates the Swin transformer and convolutional modules for global-local feature extraction. The model outperformed several CNN-based models and transformer-only models for six databases and improved crack boundary precision and robustness under noise. Chen et al., [43] developed iSwin-Unet, which uses residual Swin transformer blocks and skips attention modules to enhance feature fusion and global crack feature modeling. The model achieved the best mF1 score (86.98%) and inference speed (50.65 FPS) among benchmarks, while also improving the identification of thin and branched cracks under noisy conditions.

During the same time, transformer-based segmentation frameworks were showing increasing efficiency. Li et al., [44] applied SegFormer for crack identification, leveraging a hierarchical transformer encoder (MiT) and a lightweight decoder (All-MLP) to model global context. This model outperformed the CNN model including FCN, U-Net, and DeepLabV3, especially for thin cracks, while maintaining efficiency across model sizes. Zhang et al., [45]

introduced an Improved Segmentation Transformer-based Distress Network, built on SegFormer, with blended Mixed Attention Module (MAM), transposed convolutional upsampling module (TCUM) for multi-type distress segmentation. The model achieved an F1 score of 95.51% and an mIoU of 91.67%, enabling automated pavement condition index estimation with strong correlation with distress density. Similarly, Huang et al., [46] presented a Lightweight Transformer Patch Labeling Network (LTPLN) by utilizing the Swin Transformer backbone with token fusion and depth-wise convolution for efficiency. The model achieved an Area Under the Curve of 94.2% and real-time speed (23 FPS) under weak supervision, minimizing annotation cost considerably. In the same year Elghaish et al., [47] developed a five-layer CNN model and three optimization algorithms (ADAM, SGDM, RMSProp) applied to improve the accuracy of CNN model. The study achieved an F1 score of 98.72% using particle swarm optimization, excelling pre-trained CNN models. Supported proactive maintenance through accurate distress detection.

Sometimes, due to surroundings, lighting and condition of cracks, it becomes challenging to recognize cracks automatically with high quality [33]. Furthermore, CrackFormer architecture missed detection of cracks because of occlusion of on-surface objects, immensely slender and shallow cracks demonstrating low contrast. Similarly, complicated topography structure leads to loss of low-level details and considerable accuracy drop in edge detection. CNN-based models such as FCN, UNet, DeepLab, etc., have been employed to detect pavement surface cracks with complex topography [18].

Collectively, all these research showcases a clear direction: from hybrid CNN–ViT frameworks to highly specialized lightweight transformers, the state-of-the-art model has constantly evolved towards high accuracy, efficiency, robustness to noise, and minimized reliance on extensive labeled datasets. Moreover, Transformers are becoming foundation element for next-generation pavement crack and distress detection systems. Table 3 shows the performance of deep learning frameworks in different research papers.

Table 3. Performance of different deep learning model in different research papers.

Metrics	Values from Different Paper
F1 score	Kyem et al., [40] —0.64–0.86,
	Yu et al., [18]—86.05%,
	Zhang et al., [48]—90.13%
	Chen et al., [43] ≈ 86%
	Zhang et al., [45] ≈ 95.51%
	Elghaish et al., [47] ≈ 98.72%
	Wu et al.,[25]—69.3%
	Xiao et al., [33] ≈ 0.9364
	Guo et al., [34] 94.60%
	Liu et al., [36] ≈ 0.9
	Ali et al., [49]—0.96
Recall	Kyem et al., [40] —0.78–0.94
	Xiao et al., [33]—0.9352
	Guo et al., [34] ≈ 96.41
	Ali et al., [49]—0.95
	Xu et al., [32] ~ 92.85%
Precision	Kyem et al., [40] —0.61–0.90
	Elghaish et al., [47]—99.02%
	Xiao et al., [33]—0.9376
	Ali et al., [49]—0.97
	Zheng et al., [50]—0.896
	Xu et al., [32] ~ 93.04%

3.6. Deployment of Deep Learning for Pavement Distress Detection

One the technology for pavement distress detection was developed at Oklahoma State University, Stillwater, OK, USA. The technology can acquire pavement distresses at traffic speed (around 60–70 mph or 96.5–112.654 kmph). The technology consists of different sensors including inertial measurement unit, global positioning system sensor, 3D camera, distance measuring units, and computer hardware system. All of these sensors were attached to a vehicle for collection pavement distress images from the highway.

Once, the images were collected, they were processed in the transportation infrastructure lab at Oklahoma State University using deep learning models to determine the cracks, rut depth and other functional feature from the pavement surface.

4. Summary and Future Works

This study enlisted and elaborated deep learning techniques for automatic pavement distress detection (i.e., longitudinal cracks, transverse cracks, alligator cracks, rutting, patching, raveling, faulting, and potholes) as well as classification. The observation in this study demonstrates that deep learning architectures such as ViT, CNN and YOLO have remarkably transformed pavement monitoring by providing precise and efficient alternatives compared to historical inspection methods which are costly as well as obstructive to traffic.

Among the studied architectures, CNN was found to be the most effective for feature extraction and image classification because of the ability of CNN to learn hierarchical spatial representations. Some of the CNN based architectures showcased accuracy more than 90%, proving it the most appropriate algorithm for pavement applications. Despite this, in complicated pavement with multiple distress types and varying lighting conditions, the localized receptive field of CNN lowers the ability to capture long-range spatial dependencies.

On the other hand, YOLO architecture has proven to be relevant for real-time pavement monitoring because of its single-stage detection framework and high inference speed. YOLO with attention mechanisms and transformer modules has shown improved detection performance, with mAP values more than 94% and processing speed more than 100 FPS.

This analysis further observed that ViT has the most significant evolution in pavement distress detection. By harnessing self-attention mechanism ViT algorithm can model global contextual relationships across images, ultimately eliminates limitations linked with traditional convolutional operations. Additionally, hybrid CNN + ViT architecture has proven to be achieving outstanding performance in crack segmentation and detection. This hybrid architecture achieved F1 score more than 90% and mIoU value 80–90%. Architectures such as Swin Transformer, CrackFormer, TransUNet, SegFormer exhibits robustness for complex cracks, noisy background as well as diverse environmental conditions. Furthermore, integration of LLM and Digital Twins showcases a high potential for developing next-generation autonomous pavement monitoring systems.

Overall, the advancement from CNNs to YOLO-based real-time sensors as well as transformer driven architecture demonstrates a technological shift to autonomous pavement monitoring system. Whereas ViT and hybrid CNN–Transformer models are proven to be best for making data-driven maintenance decisions across various pavement conditions. Table 4 summarizes the application of different deep learning architectures for automated pavement distress detection.

There are three limitations of this work. First, this study did not consider some deep learning frameworks such as U-Net, SegNet, Mask R-CNN and some of the conventional deep learning models. Another limitation of this research is not considering the application of large language models for pavement distress detection. The application of LLM for pavement distress detection is an area of future work. The third limitation of this work is not considering the structural parameters of pavement. Generally, structural parameters of the pavements are determined by traffic speed deflectometer and falling weight deflectometer. The application of deep learning techniques for determining the structural parameter is scope of future work.

Table 4. Summary of application of deep learning in pavement distress detection.

Author	Application, Performance, and Algorithm
Owor et al., [51]	Application: Pavement distress (longitudinal, transverse, alligator, block, cracks, Patches, Manholes) segmentation Performance: Dice improvement ≈ 35% over Segment Anything Model, 215% performance increase across entire dataset Algorithms: ViT, CNN
Ansari et al., [3]	Application: Multi-object feature (e.g., cracks, potholes, sealed cracks, markings, joints, patches, manholes) segmentation Performance: ViT accuracy—89.23%, CNN accuracy—88.24% Algorithms: CNN, ViT, Generative Adversarial Networks
Wu et al., [21]	Application: Distress (crack, pothole, patch) classification Performance: ResNet34 (baseline)—96.75%, ResNet34 + SE—97.24%, ResNet34 + SFT—97.73% Algorithms: CNN, Spectrum Focus Transformer, SE Attention
Kyem et al., [40]	Application: Crack segmentation with Context-Aware Global Module Performance: mIoU: 0.47–0.76 (dataset wise), F1 score: 0.64–0.86, Recall: 0.78–0.94, Precision: 0.61–0.90 Algorithm: CNN
Hu et al., [52]	Application: Distress (Cracks, Rutting, Potholes) detection on mobile device with MobileNet V2 Performance: Accuracy ≈ 96.38 ± 0.15%; model size 0.6 MB Algorithm: CNN, ResNet
Lv et al., [27]	Application: A review highlighting different deep learning and hybrid models Algorithm: CNN, YOLO series, Faster R-CNN, CrackGAN

Table 4. *Cont.*

Li et al., [53]	App Application: 3D distress (Polishing & Raveling) analysis with ASViT-Net. Performance: Achieved high accuracy (MPA ≈ 85.75%; MIoU ≈ 83.78%) Algorithm: ViT
Alshawabkeh et al., [54]	Application: Automated pavement crack (thin, discontinuous, branching) detection and segmentation Performance: High accuracy (DSC 80.04% on Crack500, 91.37% on DeepCrack) Algorithm: MaskerTransformer (Mask R-CNN + ViT)
Yu et al., [18]	Application: Real-time crack (Long, Thin, Block and Alligator cracks) detection Performance: 147.20 FPS with iBiSeNetv2; F1 score: 86.05% (Crack500 dataset) Algorithm: iBiSeNetv2 (proposed model), Swin Transformer, FCN, UNet, and CCNet
Wang et al., [42]	Application: Crack (Longitudinal, Transverse, Block and Alligator cracks) detection Performance: Higher Optimal Image Scale score—0.781 to 0.849 Algorithm: Swin Transformer, CNN, Convolutional Swin-Transformer Block, Depth-convolution Forward Network
Cano-Ortiz et al., [17]	Application: Multi-defect (Alligator cracks, Longitudinal cracks, Pothole, Raveling, Patch, Sealed crack, Drain, Manhole) detection Performance: mAP@0.5; integrated geotagged images Algorithm: YOLOv5
Zhang et al., [48]	Application: Crack detection Performance: F1 ≈ 90.13%; mIoU ≈ 82.68% Algorithm: CNN-based architecture "DARNet"
Zheng et al., [55]	Application: Deep learning based intelligent distress detection Performance: F1 score: 0.7–0.85 (Crack500 dataset) Precision: 0.7–0.85 Algorithm: CNN, GAN, YOLOv3, YOLO v5, YOLO v8, Faster R-CNN
Zhang et al., [56]	Application: Multi-scene distress (Cracks, Potholes) Performance: mAP50 ≈ 79.4% Algorithm: SMG-YOLOv8 based on YOLOv8s
Tao et al., [57]	Application: Distress (Cracks, corner break, joint damage, and patch) detection Performance: mAP ≈ 49.2% Algorithm: YOLOv5
Liu et al., [58]	Application: Multi-distress (Cracks, Pothole) detection Performance: Fusion images (25IRT) mAP@0.5 ≈ 0.65 Algorithm: Faster R-CNN and YOLO series (YOLOv3/YOLOv3-SPP/YOLOv5)
Fan et al., [41]	Application: Review of pavement defect (Cracks, Potholes, Rutting, Raveling, Patching, Bleeding, Depressions) detection Algorithm: CNN based-(AlexNet, VGGNet, ResNet), Semantic segmentation networks-(U-Net, FCN, SegNet), Object detection models-(YOLO series, SDD)
Wang et al., [24]	Application: Integration of Swin transformer with YOLOv8 for distress (traverse, longitudinal, cross and alligator cracks, potholes) detection Performance: Swin Transformer integration (MAP 94.8%) Algorithm: Enhanced YOLOv8, Swin Transformer module
Chen et al., [43]	Application: Crack segmentation Performance: F1 score ≈ 86% Algorithm: iSwin-Unet, Swin-Unet, Swin Transformer, U-Net
Guo et al., [19]	Application: Crack detection transformer-based model CTv2 Performance: High mPrecision and mRecall, mFscore (>90%) Algorithm: Crack Transformer v2, Two-stage framework ResNet-34 using CNN
Li et al., [44]	Application: Crack segmentation with SegFormer Algorithm: SegFormer, CNN based models: FCN, U-Net, DeepLabV3
Zhang et al., [38]	Application: Pixel-level crack segmentation Performance: mean F-measure ≈ 94.21%; mean IoU ≈ 0.8914 Algorithm: ShuttleNetV2 (combination of CNN based encoder-decoder + Transformer self-attention mechanism)
Zhang et al., [45]	Application: Multi-distress (Cracks, Potholes, Patches) segmentation Performance: F1 ≈ 95.51%; mIoU ≈ 91.67% Algorithm: ISTD-DisNet (SegFormer MiT-B2 backbone)
Huang et al., [46]	Application: Distress (Cracks, Potholes, Repairs, Rutting) detection Performance: AUC ≈ 94.2%; real-time processing speed Algorithm: Lightweight Transformer Patch Labeling Network (Swin Transformer backbone), CNN based models—ResNet, DenseNet, EfficientNet, object detection framework—YOLO, SSD, Faster RCNN, RetinaNet
Elghaish et al., [47]	Application: Crack classification Performance: F1 ≈ 98.72%, Precision—99.02% Algorithm: CNN
Wu et al., [25]	Application: Crack (Longitudinal, Transverse, Alligator) detection Performance: F1 score: 69.3% Algorithm: Multi-scale feature fusion, SSD, YOLOv5

Table 4. *Cont.*

Liu et al., [59]	Application: Introduced dataset for distress (Cracks, Patches, Potholes, Background images without defects) detection.
Zaman et al., [60] 2024	Application: Crack Detection Performance: Accuracy (99%) Algorithm: ViT, CNN
Xiao et al., [33]	Application: Crack detection with hybrid-window transformer architecture Performance: F1 score ≈ 0.9364, Precision: 0.9376, Recall: 0.9352 Algorithm: CrackFormer Transformer-based High-Resolution Network (HRNet)
Wang et al., [61]	Application: Detection & classification of asphalt pavement cracks Performance: ViT-improved YOLO V5 enhances detection accuracy & speed (11.9 ms/image), enabling real-time crack detection. Algorithm: YOLOv5, ViT
Guo et al., [34]	Application: Crack detection and segmentation Performance: 109 FPS; high mF1 ≈ 94.60% and recall ≈ 96.41% Algorithm: CT (VIT based model using Swin transformer encoder and SegFormer-inspired MLP decoder), CNN based semantic segmentation model—U-Net
Guo et al., [35]	Application: Crack detection Performance: High recall and accuracymF1 up to 87%, Recall up to 90%; 115 FPS Algorithm: STA (Swin Transformer encoder + UperNet decoder + CBAM attention module)
Liu et al., [36]	Application: Crack segmentation Performance: F1 ≈ 0.9; fewer parameters than several CNN-based models like DeepCrack. Algorithm: CNN, VIT
Luo et al., [23]	Application: Crack (Longitudinal, Transverse, Alligator) detection Performance: mAP ≈ 63.73%; robust in complex scenes Algorithm: YOLOX, Swin transformer, Global Attention Guidance Module, FPN
Tong et al., [37]	Application: Distress (Cracks, Potholes, Repair areas, Background pavement) segmentation Performance: Better uncertainty calibration; high pixel accuracy Algorithm: ES-transformer (ViT + Dempster-Shafer Theory (DST)), Baseline model: PS-Transformer (primary baseline), also compared with P-FCN
Li et al., [62]	Application: Pavement distress (Alligator, Longitudinal, Block and Transverse crack, Pothole, Patch) classification and detection Performance: Accuracy 96.24%—binary classification task (distressed vs. non-distressed pavement images) Algorithm: ResNet-34, ResNet-50, VGG-16, VGG-19
Ali et al., [49]	Application: Crack detection, localization, and segmentation Performance: Test accuracy: 0.96, Precision: 0.97, Recall: 0.95, F1: 0.96 Algorithm: ViT, Tubularity-Flow-Field (TuFF) Algorithm, Sliding Window (SW)
Lin et al., [63]	Application: Distress (Cracks (including longitudinal and transverse), Potholes, Spalling, Faulting) Algorithm: UNet-based models, FCN, YOLACT
Zheng et al., [50]	Application: Real-time concrete pavement crack detection on edge devices Performance: Precision 0.896, lightweight design, 89 FPS real-time performance. Algorithm: BT-YOLO (Bottleneck Transformer and YOLOv5)
Xu et al., [32]	Application: Crack detection Performance: Precision: ~93.04%, Recall: ~92.85% Algorithm: Transformer-based architecture
Shamsabadi et al., [30]	Application: Crack detection in asphalt and concrete surfaces with TransUNet Architecture Performance: TransUNet (CNN-ViT) achieved up to ~61% better mean IoU on CT260 dataset Algorithm: CNN + ViT backbone, Baselines: DeepLabv3+, U-Net
Guan et al., [64]	Application: Distress (Transverse, Longitudinal, Block, Alligator, Edge, Reflection, Potholes, Raveling) detection & classification Performance: Captures global context for complex cracks, F1-88% to 91.9% with hybrid model Algorithm: Hybrid-Window Attentive Vision Transformer (HRNet-based hybrid deep learning model)
Gopalakrishnan, [65]	Application: Automated pavement distress detection using pre-trained VGG-16 CNN Performance: High detection accuracy ~90% Algorithm: CNN (VGG-16) with transfer learning

Although deep learning was extensively used by researchers for pavement monitoring, however, some of the challenges exist in the pavement industry such as generalizability of models, and requirement of large training time of models and large volumes of data. The problem of the conventional deep learning algorithms can be solved using vision or multimodal Large Language Models (LLMs) and Digital Twins (DTs).

LLMs are based on transformer architecture [6,66] and were initially developed for performing various language translation tasks; however, they can be deployed to solve a range of computer vision problems. As LLMs are trained on large volumes of data, they can achieve better generalization and shorter model development time than conventional deep learning models.

DTs are copies of the infrastructure and consist of a digital model, sensors, and a communication system that exchanges information between the sensors and the communication system. DTs can be developed using embedded pavement sensors and a finite-element pavement model. DT would provide insights about the reasons for cracking on the pavement surface and could provide information about crack formation within the asphalt layer. It is expected that, along with deep learning models, LLMs and DTs will be employed by researchers for pavement monitoring. The advanced models, ViT can extract different pavement features including the cracks, potholes, man holes from the surface of the pavements. This cracks and other features can be incorporated into the virtual model, this can result in realistic modelling pavement distresses.

Author Contributions

Conceptualization, K.C., M.D.; Methodology, K.C., M.D. and J.L.; Formal analysis, K.C. and M.D.; Investigation, K.C., M.D. and J.L.; Data curation, K.C., and M.D.; Writing—original draft preparation, K.C., and M.D.; Writing—review and editing, K.C., M.D. and J.L.; Visualization, K.C.; All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not Applicable.

Informed Consent Statement

Not Applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

Grammarly was used exclusively to improve the clarity and readability of the language. It did not contribute to the generation of data, results, or scientific conclusions presented in this manuscript.

References

1. Chatterjee, K.; Vivanco, D.; Yang, X.; et al. Enhancing Pavement Performance through Balanced Mix Design: A Comprehensive Field Study in Oklahoma. In Proceedings of the International Conference on Transportation and Development, Atlanta, GA, USA, 15–18 June 2024; pp. 511–522.
2. Parajulee, K.; Chatterjee, K.; Li, J. Leveraging original equipment manufacturer vehicle sensor data for enhanced roadway safety. *Int. J. Pavement Res. Technol.* **2025**, 1–18. https://doi.org/10.1007/s42947-025-00619-z.
3. Ansari, F.; Chatterjee, K.; Li, J.Q.; et al. Multi-Object Pavement Surface Feature Detection with CNN and Transformer Deep Learning Architecture. In Proceedings of the Airfield and Highway Pavements, Glendale, AZ, USA, 8–11 June 2025; pp. 350–359.
4. Chatterjee, K.; Li, J.Q.; Ansari, F.; et al. Hybrid LSTM-Transformer models for profiling highway–railway grade crossings. *J. Transp. Eng. Part A Syst.* **2026**, *152*, 04025138.
5. Chatterjee, K.; Li, J.; Parajulee, K.; et al. Assessing Hangup Susceptibility of Highway-Railway Grade Crossings Leveraging CNN-Transformer Hybrid Model. In Proceedings of the International Conference on Transportation and Development, Detroit, MI, USA, 28 June–1 July 2026; pp. 579–590.
6. Chatterjee, K.; Desai, M.; Li, J. Application of large language models in geotechnical engineering: A movement towards safe and sustainable future. *Geotechnics* **2026**, *6*, 38.
7. Asteris, P.G. Computational intelligence: From nature and Aristotle to Meta-Heuristic algorithms. *Bull. Comput. Intell.* **2025**, *1*, 1–2. https://doi.org/10.53941/bci.2025.100001.

8. Benzaamia, A.; Ghrici, M.; Rbouh, R.; et al. Prediction of chloride resistance level in concrete using optimized tree-based machine learning models. *Bull. Comput. Intell.* **2025**, *1*, 104–117.
9. Khatti, J.; Kontoni, D.P.N. Assessment of bearing capacity of concrete piles in alluvial soils using bio and swarm-optimized artificial neural network models. *Bull. Comput. Intell.* **2025**, *1*, 53–75.
10. Ansari, S.S.; Ansari, M.A.; Ibrahim, S.M. Explainable predictive modelling of sustainable slag-fly ash based geopolymer concrete with life cycle and carbon-neutrality assessment. *Bull. Comput. Intell.* **2026**, *2*, 54–82.
11. Bardhan, A.; Kardani, N. An efficient meta-ensemble paradigm for modelling Poisson's ratio and maximum horizontal stress in casing collapse hazard. *Bull. Comput. Intell.* **2026**, *2*, 146–163.
12. Desai, M.; Chatterjee, K. Application of machine learning techniques for prediction of soil water characteristics curve: A state of the art review. *preprints* **2026**, *Epub ahead of print*. https://doi.org/10.20944/preprints202602.0792.v1.
13. Muhammed, A.M.; Omer, M.A.I.; Haider, R.S.; et al. Predicting the compressive strength of fly ash composite foam concrete using artificial neural networks and soft computing techniques. *Bull. Comput. Intell.* **2026**, *2*, 83–102.
14. Iqbal, M.; Raheel, M.; Biswas, R. Hybrid artificial neural network models for predicting flexural strength of FRP-reinforced concrete beams. *Bull. Comput. Intell.* **2026**, *2*, 196–212.
15. Subedi, A.; Subedi, S.; Adhikari, S.; et al. Explainable hybridized machine learning for prediction of compressive strength of fly-ash based geopolymer concrete. *Bull. Comput. Intell.* **2026**, *2*, 213–234.
16. Yang, F.; Zhang, L.; Yu, S.; et al. Feature pyramid and hierarchical boosting network for pavement crack detection. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1525–1535.
17. Cano-Ortiz, S.; Iglesias, L.L.; del Árbol, P.M.R.; et al. An end-to-end computer vision system based on deep learning for pavement distress detection and quantification. *Constr. Build. Mater.* **2024**, *416*, 135036.
18. Yu, H.; Deng, Y.; Guo, F. Real-time pavement surface crack detection based on lightweight semantic segmentation model. *Transp. Geotech.* **2024**, *48*, 101335.
19. Guo, F.; Liu, J.; Xie, Q.; et al. A two-stage framework for pixel-level pavement surface crack detection. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108312.
20. Zhang, H.; Zhang, A.A.; Dong, Z.; et al. Robust semantic segmentation for automatic crack detection within pavement images using multi-mixing of global context and local image features. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 11282–11303.
21. Wu, W.; Zhu, F.; Li, Z.; et al. Optimized deep learning model with integrated spectrum focus transformer for pavement distress recognition and classification. *Sci. Rep.* **2025**, *15*, 3803.
22. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
23. Luo, H.; Li, J.; Cai, L.; et al. STrans-YOLOX: Fusing Swin Transformer and YOLOX for automatic pavement crack detection. *Appl. Sci.* **2023**, *13*, 1999.
24. Wang, S.; Cai, B.; Wang, W.; et al. Automated detection of pavement distress based on enhanced YOLOv8 and synthetic data with textured background modeling. *Transp. Geotech.* **2024**, *48*, 101304.
25. Wu, P.; Wu, J.; Xie, L. Pavement distress detection based on improved feature fusion network. *Measurement* **2024**, *236*, 115119.
26. Yang, X.; Li, J.Q.; Zhan, Y.; et al. Real-time automated deep learning based railroad trespassing violation detection and tracking at highway-rail grade crossing. *Intell. Transp. Infrastruct.* **2024**, *3*, liae003.
27. Lv, Z.; Hao, Z.; Zhu, Y.; et al. A review on automated detection and identification algorithms for highway pavement distress. *Appl. Sci.* **2025**, *15*, 6112.
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
30. Shamsabadi, E.A.; Xu, C.; Rao, A.S.; et al. Vision transformer-based autonomous crack detection on asphalt and concrete surfaces. *Autom. Constr.* **2022**, *140*, 104316.
31. Tang, W.; Huang, S.; Zhang, X.; et al. Pict: A Slim Weakly Supervised Vision Transformer for Pavement Distress Classification. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 3076–3084.
32. Xu, Z.; Guan, H.; Kang, J.; et al. Pavement crack detection from CCD images with a locally enhanced transformer network. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *110*, 102825.
33. Xiao, S.; Shang, K.; Lin, K.; et al. Pavement crack detection with hybrid-window attentive vision transformers. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *116*, 103172.
34. Guo, F.; Qian, Y.; Liu, J.; et al. Pavement crack detection based on transformer network. *Autom. Constr.* **2023**, *145*, 104646.
35. Guo, F.; Liu, J.; Lv, C.; et al. A novel transformer-based network with attention mechanism for automatic pavement crack detection. *Constr. Build. Mater.* **2023**, *391*, 131852.

36. Liu, H.; Yang, J.; Miao, X.; et al. CrackFormer network for pavement crack segmentation. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 9240–9252.
37. Tong, Z.; Ma, T.; Zhang, W.; et al. Evidential transformer for pavement distress segmentation. *Comput. Aided Civ. Infrastruct. Eng.* **2023**, *38*, 2317–2338.
38. Zhang, H.; Zhang, A.A.; He, A.; et al. Pixel-level detection of multiple pavement distresses and surface design features with ShuttleNetV2. *Struct. Health Monit.* **2024**, *23*, 1263–1279.
39. Zim, A.H.; Iqbal, A.; Al-Huda, Z.; et al. EfficientCrackNet: A Lightweight Model for Crack Segmentation. In Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 26 February–6 March 2025; pp. 6279–6289.
40. Kyem, B.A.; Asamoah, J.K.; Aboah, A. Context-CrackNet: A context-aware framework for precise segmentation of tiny cracks in pavement images. *Constr. Build. Mater.* **2025**, *484*, 141583.
41. Fan, L.; Wang, D.; Wang, J.; et al. Pavement defect detection with deep learning: A comprehensive survey. *IEEE Trans. Intell. Veh.* **2024**, *9*, 4292–4311.
42. Wang, C.; Liu, H.; An, X.; et al. SwinCrack: Pavement crack detection using convolutional swin-transformer network. *Digit. Signal Process.* **2024**, *145*, 104297.
43. Chen, S.; Feng, Z.; Xiao, G.; et al. Pavement crack detection based on the improved Swin-Unet model. *Buildings* **2024**, *14*, 1442.
44. Li, H.; Zhang, H.; Zhu, H.; et al. Automatic crack detection on concrete and asphalt surfaces using semantic segmentation network with hierarchical transformer. *Eng. Struct.* **2024**, *307*, 117903.
45. Zhang, Z.; Song, W.; Zhuang, Y.; et al. Automated multi-type pavement distress segmentation and quantification using transformer networks for pavement condition index prediction. *Appl. Sci.* **2024**, *14*, 4709.
46. Huang, W.Q.; Feng, L.; He, Y.L. LTPLN: Automatic pavement distress detection. *PLoS ONE* **2024**, *19*, e0309172.
47. Elghaish, F.; Matarneh, S.; Abdellatef, E.; et al. Multi-layers deep learning model with feature selection for automated detection and classification of highway pavement cracks. *Smart Sustain. Built Environ.* **2025**, *14*, 511–535.
48. Zhang, J.; Sun, S.; Song, W.; et al. Automated pavement distress detection based on convolutional neural network. *IEEE Access* **2024**, *12*, 105055–105068.
49. Ali, L.; Jassmi, H.A.; Khan, W.; et al. Crack45K: Integration of vision transformer with tubularity flow field (TuFF) and sliding-window approach for crack-segmentation in pavement structures. *Buildings* **2023**, *13*, 55.
50. Zheng, X.; Qian, S.; Wei, S.; et al. The combination of Transformer and You Only Look Once for automatic concrete pavement crack detection. *Appl. Sci.* **2023**, *13*, 9211.
51. Owor, N.J.; Adu-Gyamfi, Y.; Aboah, A.; et al. PaveSAM—Segment anything for pavement distress. *Road Mater. Pavement Des.* **2025**, *26*, 593–617.
52. Hu, Y.; Chen, N.; Hou, Y.; et al. Lightweight deep learning for real-time road distress detection on mobile devices. *Nat. Commun.* **2025**, *16*, 4212.
53. Li, Y.; Liu, C.; Weng, Z.; et al. Aggregate-level 3D analysis of asphalt pavement deterioration using laser scanning and vision transformer. *Autom. Constr.* **2025**, *178*, 106380.
54. Alshawabkeh, S.; Wu, L.; Dong, D.; et al. A hybrid approach for pavement crack detection using Mask R-CNN and Vision Transformer model. *Comput. Mater. Contin.* **2025**, *82*, 561–577.
55. Zheng, L.; Xiao, J.; Wang, Y.; et al. Deep learning-based intelligent detection of pavement distress. *Autom. Constr.* **2024**, *168*, 105772.
56. Zhang, S.; Bei, Z.; Ling, T.; et al. Research on high-precision recognition model for multi-scene asphalt pavement distresses based on deep learning. *Sci. Rep.* **2024**, *14*, 25416.
57. Tao, Z.; Gong, H.; Liu, L.; et al. A weakly-supervised deep learning model for end-to-end detection of airfield pavement distress. *Int. J. Transp. Sci. Technol.* **2025**, *17*, 67–78.
58. Liu, F.; Liu, J.; Wang, L.; et al. Multiple-type distress detection in asphalt concrete pavement using infrared thermography and deep learning. *Autom. Constr.* **2024**, *161*, 105355.
59. Liu, Z.; Wu, W.; Gu, X.; et al. PaveDistress: A comprehensive dataset of pavement distresses detection. *Data Brief* **2024**, *57*, 111111.
60. Zaman, F.; Xu, Z.; Hussain, A.; et al. Concrete and Pavement Cracks Detection Leveraging Vision Transformer (ViT) Model. In Proceedings of 2024 International Conference on Electrical and Information Technology (IEIT), Malang, Indonesia, 12–13 September 2024; pp. 90–95.
61. Wang, S.; Chen, X.; Dong, Q. Detection of asphalt pavement cracks based on vision transformer improved YOLOv5. *J. Transp. Eng. Part B Pavements* **2023**, *149*, 04023004.
62. Li, D.; Duan, Z.; Hu, X.; et al. Automated classification and detection of multiple pavement distress images based on deep learning. *J. Traffic Transp. Eng.* **2023**, *10*, 276–290.

63. Lin, W.; Li, X.; Han, H.; et al. A novel approach for pavement distress detection and quantification using RGB-D camera and deep learning algorithm. *Constr. Build. Mater.* **2023**, *407*, 133593.
64. Guan, S.; Liu, H.; Pourreza, H.R.; et al. Deep learning approaches in pavement distress identification: A review. *arXiv* **2023**, arXiv:2308.00828.
65. Gopalakrishnan, K. Deep learning in data-driven pavement image analysis and automated distress detection: A review. *Data* **2018**, *3*, 28.
66. Chatterjee, K. Application of Large Language Models in Intelligent Transportation Systems: A Systematic Review. In Proceedings of International Conference on Transportation and Development, Detroit, MI, USA, 28 June–1 July 2026; pp. 565–572.