Review

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

Mohak Desai and Kaustav Chatterjee *

Posted Date: 24 March 2026

doi: 10.20944/preprints202603.1912.v1

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

 doi:10.20944/preprints202603.1912.v1

Review

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

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

[1] Senior Data Analyst, NIQ, Vadodara, India

[2] PhD, Oklahoma State University, Stillwater, Oklahoma

* Correspondence: kaustav.chatterjee10@okstate.edu

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 You Only Look Once (YOLO), Convolutional Neural Networks (CNNs), and vision transformer architectures 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 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

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.

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 like You Only Look Once (YOLO), Convolutional Neural Network (CNN), Vision Transformer (ViT) have showcased extraordinary performance in identifying pavement distress such as longitudinal and transverse cracks (Chatterjee et al., 2024), rutting, faulting, patching, shoving, raveling and potholes from images and data from sensors (Rana et al., 2025, Parajulee et al., 2025). 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, thanks to their self-attention mechanism (Ansari et al., 2025; Chatterjee et al., 2025, Chatterjee et al., 2026a, Chatterjee et al., 2026b).

 doi:10.20944/preprints202603.1912.v1

This study presents a comprehensive review of application of the state-of-the-art deep learning techniques for pavement distress detection, including their underlying architectures, applications, advantages, and limitations. This review paper focuses on the application of vision transformer algorithm for pavement distress monitoring. Moreover, some of the previous deep learning algorithms such as CNN and YOLO were also covered in this paper. Figure 1 shows the overall flowchart of the research. The research started with selection of different keywords for literature acquisition. The different keywords used were vision transformer and pavement monitoring, YOLO and pavement distress monitoring, and CNN and pavement distress monitoring. Subsequently, using the pair of keywords, literature was acquired from the Science Direct database, American Society of Civil Engineer (ASCE) database, IEEE database, and Multi-Disciplinary Publishing Institute Database. Thereafter, literature was selected and sorted based on article title, keywords and abstract and finally they were summarized in this paper.

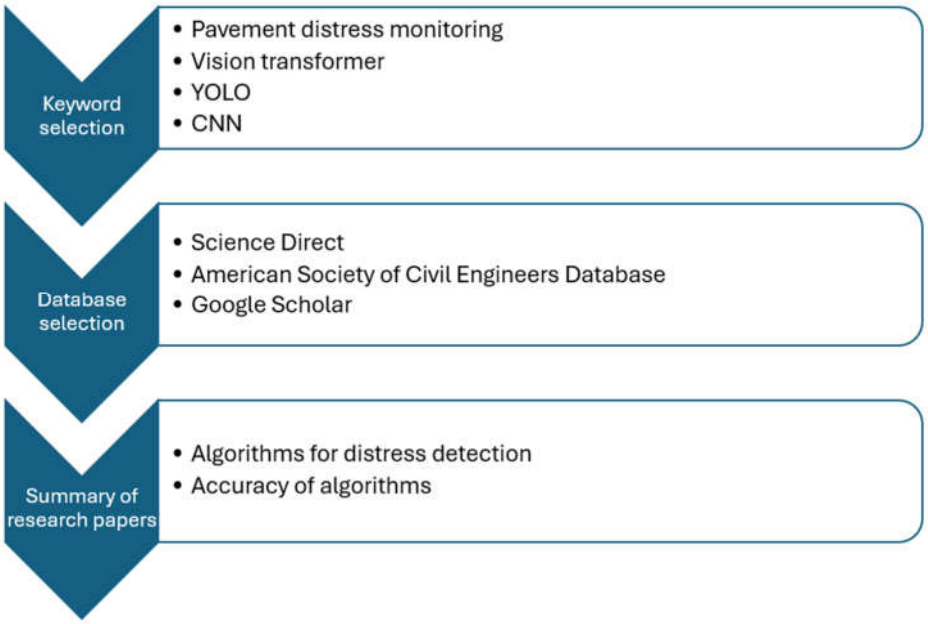

Figure 1. Flowchart of research.

Pavement Distress

Pavement distress refers to any deterioration in a pavement system that showcases reduction of structural capacity, service life and performance. In other words, any kind of damage appeared on the pavement surface or within pavement layers which flags that pavement surface is failing, it is referred as a pavement distress. It involves cracks, surface defects, deformations and moisture-related problems which may occur because of traffic loading, material aging, environmental effects or construction deficiencies. Figure 2 shows the different categories of pavement distress.

 doi:10.20944/preprints202603.1912.v1

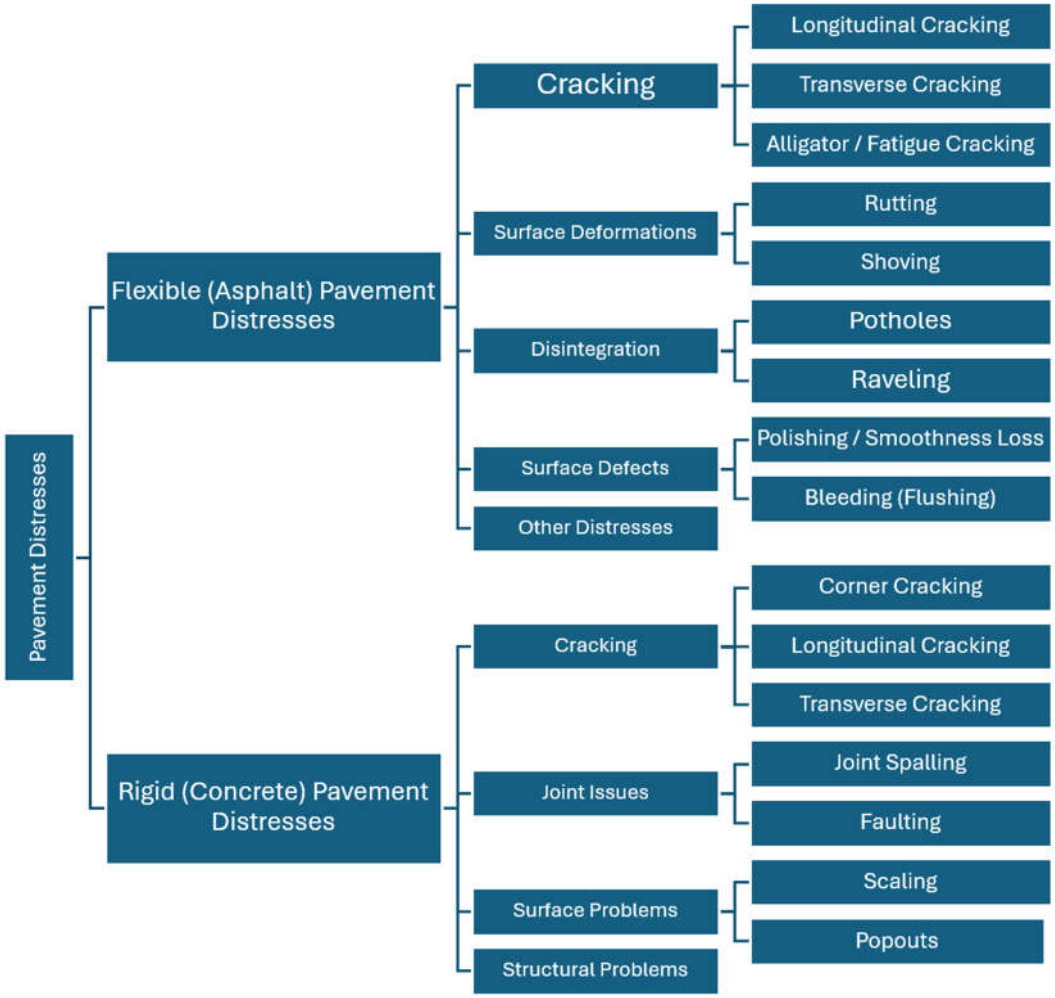

Figure 2. Types of pavement distresses.

Flexible Pavement Distresses

Alligator/Fatigue Cracking: This is the most common type of distress in flexible (asphalt) pavements. These cracks resulted from traffic loads, and they look like interconnected cracks, like crocodile's skin. It generally looks like small, closely spaced and multi-sided cracks in area of the asphalt surface.

Longitudinal Cracking: This type of crack runs parallel to the centerline or traffic direction, and they are caused by shrinkage and poor joints. These cracks basically signal issues related to construction process, pavement structure, or environmental effects.

Transverse Cracking: Transverse cracks, also known as Thermal cracks, runs perpendicular to centerline or traffic direction; and appears mostly because of temperature variations, asphalt shrinkage, or construction issues. They are generally visible as single cracks spaced at regular intervals.

Block Cracking: These are rectangular cracks resulted from asphalt binder aging or shrinkage and not caused by traffic, normally ranging from 0.3 m to 3 m in size. These cracks appears over a large area and are not load-related.

Rutting: Rutting is a longitudinal surface depression in wheel paths caused by subgrade deformation and repeated traffic loading. It is visible as permanent grooves or ruts in area where vehicle tires regularly travel.

Shoving: It 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.

 doi:10.20944/preprints202603.1912.v1

Potholes: These are Bowl-shaped holes formed due to moisture intrusion and traffic load and where pieces of asphalt have broken loose. Potholes usually occur when water infiltrates cracks, weakens the underlying layers, and traffic makes the pavement material break apart.

Raveling: It appears 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. This is a surface distress generally visible in old or poorly constructed asphalt pavements.

Polishing / Smoothness Loss: Polishing, also known as smoothness loss, is a type of distress in which aggregate particles become smooth and rounded because of traffic, which ultimately reduces pavement's skid resistance.

Bleeding: Bleeding, also known as flushing, is a type of pavement surface distress where extra asphalt binder rises to the surface, forming a black, sticky, shiny and smooth film on the pavement. It generally appears in hot weather and results in reduced skid resistance. Bleeding occurs when asphalt binder fills the aggregate voids in hot weather or traffic compaction and then expands onto the pavement surface. Bleeding is irreversible during cold weather or periods of low loading so asphalt binder will appear on the pavement surface over time.

Rigid Pavement Distresses

Longitudinal Cracks: Longitudinal cracks in concrete pavement are the cracks that runs parallel to pavement's centerline or direction of traffic. These types of cracks may develop within a slab or along a joint.

Transverse Cracks: This type of crack runs perpendicular to the pavement's centerline or direction of traffic in concrete pavement; and appears mostly because of temperature variations and shrinkage. These cracks are commonly extended across the full width of the slab.

Corner Cracking: Corners cracks, also known as corner break, are full depth fractures developing near corner of a concrete pavement, usually caused by high stresses at the pavement corners. It occurs due to loss of support or heavy wheel load placed near to the slab corner.

Joint Spalling: Joint spalling is the deterioration of concrete pavement edge near joints, which leads to fragmentation of concrete, cracks, chipping or breaking. It appears when the concrete near joints breaks because of poor joint construction or water infiltration.

Faulting: Faulting is an elevation difference between slabs across a joint or crack, mostly because of pumping. In this case, generally one slab settles lower, and the other slab remains higher, forming a step-like uneven surface. Faulting occurs because of poor load transfer, pumping or loss of support.

Scaling: It is a surface level concrete pavement distress where the uppermost layer of the concrete peels, flakes, or crumbles, generally in small or large patches. It mainly happens due to freeze–thaw cycles or poor concrete finishing.

Popouts: These are small conical holes that appeared on the surface of a concrete pavement made due to expansion of porous aggregate particles and aggregate reactivity. These are cosmetic surface defects and usually do not affect the structural performance unless the number is higher.

Artificial Intelligence

Artificial intelligence is the domain that deals with development of machine/software to perform tasks that are generally done by humans and require human intelligence. In other words, with the help of AI, computers can think, learn and make decisions similar to how humans do. AI systems are built to understand language, recognize images, solve problems and to learn from the data. The pipeline of AI consists of data collection, data processing, model training, evaluation, model deployment and monitoring & retraining. There are five principle technical building blocks based on which AI works, such as Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), Computer Vision and Expert Systems. Figure 3 schematically shows the relationship between artificial intelligence, machine learning and deep learning.

 doi:10.20944/preprints202603.1912.v1

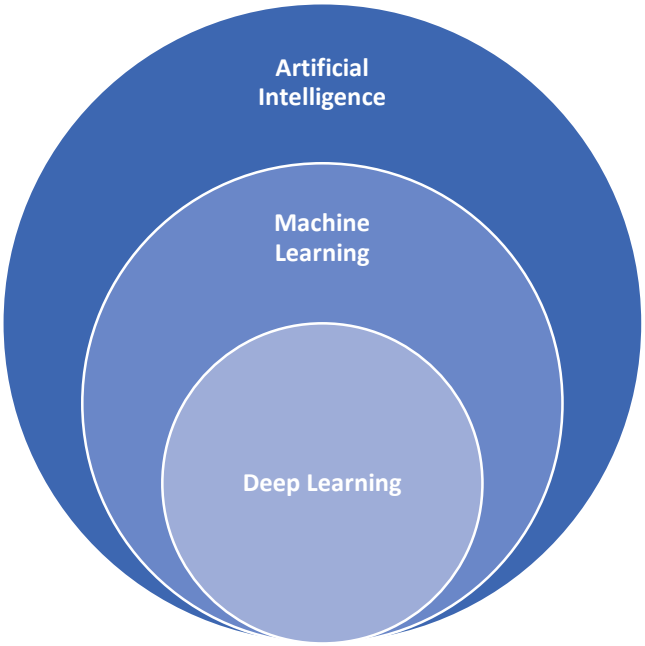

Figure 3. Relationship between AI, ML & DL.

Machine Learning

Machine learning (ML) is a subdivision of AI that facilitates computers to learn patterns from data and make predictions as well as decisions without programming for each task. Instead of writing rules manually, ML models search for rules by exploring examples. There are three types of ML, such as supervised learning, unsupervised learning and reinforcement learning (RL). A few real-life examples of ML in real life are banks predicting frauds and Google Maps predicting traffic. 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, random forest algorithm, gradient boosting algorithm and support vector machine (Chatterjee et al., 2024, Rana et al., 2025, Parajulee et al., 2025, Munna et al., 2025, Desai et al., 2026)

Deep Learning

Deep learning (DL) is a subset of ML, and ML itself is a subset of AI. DL uses multi-layered neural networks to understand complex patterns from large amounts of data. DL outperforms image and speech recognition, Natural language processing as well as self-driving systems. It excels due to high-capacity networks, ability to learn hierarchical features as well as optimization advancements. Deep learning automatically learns features from raw data instead of depending on manual feature engineering. Some of the common deep learning architectures are Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformers, etc.

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 4 shows a schematic representation of the confusion matrix.

 doi:10.20944/preprints202603.1912.v1

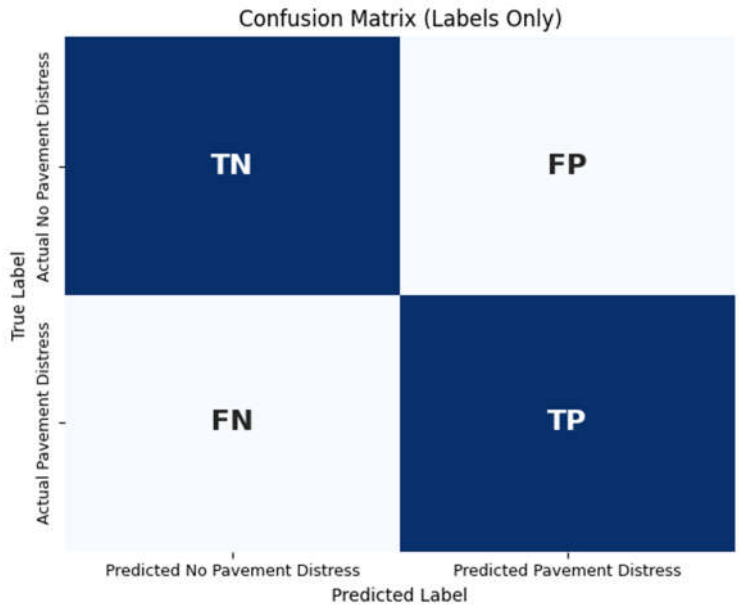

Figure 4. Schematic representation of confusion matrix showing true positive, false positive, true negative, false negative.

Accuracy metrics estimates how accurately a system can identify pavement conditions in applications such as automated pavement distress detection, pavement condition classification and image-based image defect identification.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

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

Precision defines how accurate the model is when it identifies pavement distress. High precision indicates how correctly crack is detected and low precision means false identification of shadow pattern or markings as a crack. For example, if the model identifies 100 cracks, and out of those 100, only 80 are actual cracks, then precision value would be 0.80.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall showcases how well the model identifies all actual pavement distresses. When the model rarely misses a crack, it gives high recall and when a model fails to detect thin cracks or small potholes, there is a low recall value for that model. For instance, if there are 120 cracks but the model only identifies 90, recall value would be 0.75.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1 score represents a balanced measure of both “how many distress events are detected” and “how many detected events are correct”. A framework with high precision but low recall will have a moderate F1 number.

 doi:10.20944/preprints202603.1912.v1

$$F1 = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision+Recall}} \quad (4)$$

Specificity represents how properly a model recognize pavement sections that are not distressed. High specificity represents rare misidentification of smooth pavement as damage, and low specificity means the model labels smooth pavement as cracked due to texture or lighting.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (5)$$

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 help to understand how accurate object detection is. 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.

$$\text{IoU} = \frac{\text{Area of Overlap/Intersaction}}{\text{Area of Union}} \quad (6)$$

Mean Absolute Error (MAE): MAE represents the average magnitude of prediction errors, without considering direction. Here, y_i represents actual pavement condition value and $\hat{y}_i$ represents predicted value. Low MAE represents high accuracy and vice versa.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}i | \quad (7)$$

Mean Squared Error (MSE): MSE evaluates the average of squared prediction errors. Because the errors are squared, large prediction errors are given much larger weight.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}i)^2 \quad (8)$$

Mean Average Precision (mAP): mAP is a main metric for object detection, used to identify how well a model detects pavement distresses such as potholes, alligator cracks, bleeding, raveling, etc. mAP measures precision and recall across multiple thresholds of IoU.

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_i \quad (9)$$

Root Mean Squared Error (RMSE): It is a square root of MSE, showcasing prediction errors in the same units as the pavement measurement. Low RMSE means prediction match field measurement, whereas high RMSE means the model deviates highly from the real pavement data.

$$\text{RMSE} = \sqrt{\text{MSE}} \quad (10)$$

APPLICATIONS OF DEEP LEARNING IN PAVEMENT MONITORING

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

 doi:10.20944/preprints202603.1912.v1

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. Figure 5 shows a schematic representation of pavement feature detection using CNN.

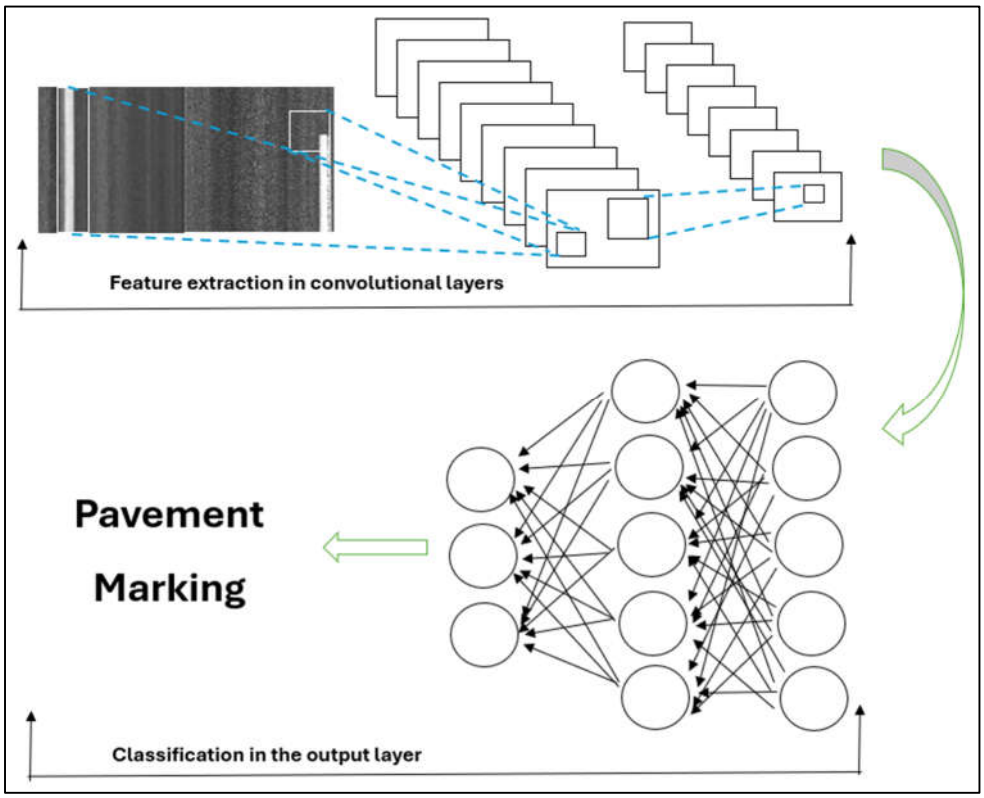

Figure 5. Pavement Marking Detection using CNN.

Over the past couple of years, deep learning has 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 exprolre hybrid architectures that combine CNNs with transformer-based models to improve accuracy as well as real-time performance. For instance, Huayang et al., (2024) developed 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. Feng et al. (2024) 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., (2024a) 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, Fateme et al., (2025) compared CNNs 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

 doi:10.20944/preprints202603.1912.v1

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.

In parallel, Wenlin et al., (2025) presented Spectrum Focus Transformer (SFT) integrated with ResNet34, a frequency-domain attention model to elevate pavement distress detection. The model's ability to handle image data with complex backgrounds and diverse distress levels surpasses SE attention and other CNN models.

Overall, this research showcases 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.

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 (Bochkovskiy et al. 2020) 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, cameras, and robotics. The architecture of YOLO has evolved through multiple versions (YOLOv1 to YOLOv11) as shown in figure 6, 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 7 shows the schematic representation of detection of pavement markings using YOLO.

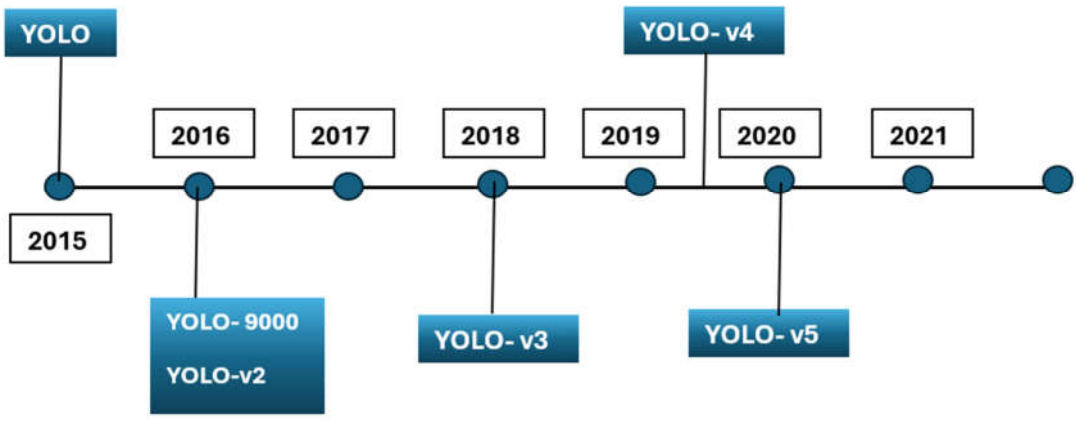

Figure 6. YOLO Timeline.

 doi:10.20944/preprints202603.1912.v1

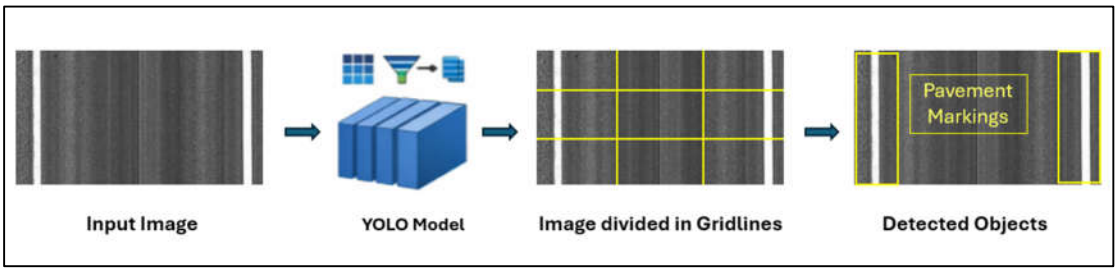

Figure 7. YOLO Workflow.

Recent research in vision-based pavement distress detection has adopted hybrid architecture that combines global and local feature modeling. Hui et al. (2023) 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, Sicheng et al. (2024) extended the scope of transformer architectures with transformer architectures by combining swin transformers with Squeeze-and-Excitation (SE) attention into 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, Peng et al. (2024) 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 of video data, outperforming the YOLOv5 baseline while minimizing model parameters by 27%.

In the same year, Yang et al. (2024) 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. 104 hours of field video data was collected, and model was tested on it. 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 a work done by Zhenglong et al. (2025), who worked on a comprehensive review of pavement distress detection methodologies that include manual inspection and state-of-the-art AI-driven systems, highlighting algorithms such as ViT, 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 researcher compared YOLO and Faster R-CNN 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.

Vision Transformer

 doi:10.20944/preprints202603.1912.v1

Vision transformer (ViT) is a deep learning architecture based on the principles of transformer architecture (Vaswani et al. 2017), developed for image analysis. 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 the recent times, transformer architecture was used for determining the profiles of highway-railway grade crossings (Chatterjee et al. 2026). Figure 8 shows the schematic representation of pavement feature detection using ViT framework.

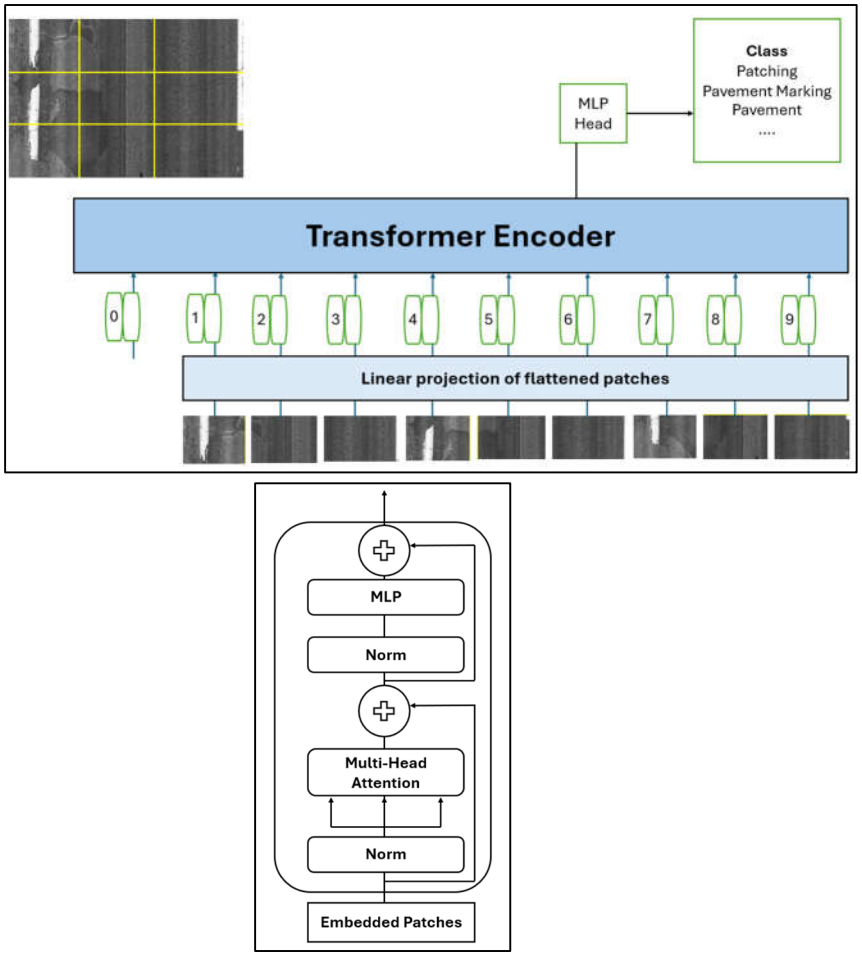

Figure 8. Vision Transformer.

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. Elyas et al., (2022) presented an early benchmark by introducing 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 and limited edge optimization. Similarly, Wenhao et al., (2022) introduced PicT, a lightweight ViT integrated 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, improved patch-level localization of distressed regions, and addressed challenges posed by sparse distress areas and weak supervision. For better local-global feature integration, Zhengsen et al., (2022) presented LETNet, which merges

 doi:10.20944/preprints202603.1912.v1

ViT with local enhancement modules for global-local feature fusion. The model achieved high IoU 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.

Shaozhang et al., (2023) developed a CrackFormer with a hybrid window and weighted multi-head attention for fine-grained crack segmentation. The model excels CNNs and capsule networks in terms of F1 score (0.9364). Shortcomings of the model included the requirement for a large dataset and high computational resources. Feng et al., (2023) developed a crack transformer model that uses a Swin Transformer encoder and a SegFormer decoder for crack segmentation. The model achieved high F1 (94.60) and recall (96.41) and demonstrated robust performance with noisy data and complex crack patterns across multiple datasets. Feng et al., (2023) proposed the Swin Transformer with Attention model, merging the Swin Transformer encoder with the UperNet decoder and Convolutional Block Attention Module attention for increased feature refinement. This model outperforms CNN-based computer vision models in terms of accuracy and recall, and can handle noisy backgrounds and complex crack geometries. Huajun et al., (2023) conducted a study to upgrade CrackFormer to CrackFormer-II, adding local self-attention and feed-forward layers to improve global-local feature fusion. The model improved segmentation of fine and coarse cracks and achieved state-of-the-art F1 scores (0.912) across multiple databases, while reducing FLOPs. Zheng et al., (2023) combined the Dempster–Shafer Theory (a mathematical framework for combining evidence and handling uncertainty in decision-making) with ViT to improve uncertainty calibration and robustness in segmentation. The model achieved high pixel-level accuracy and mIoU compared to probabilistic models, enabling multimodel fusion without retraining. The study also effectively highlights annotation noise and class imbalance. Hang et al., (2023) developed ShuttleNetV2, merging transformer blocks into a repeatable encoder-decoder structure for multi-type distress identification, including sealed cracks, patches, and potholes. The model achieved 94.21% F1 and an IoU score of 0.8914, outperforming the CNN and SegFormer baselines.

By the year 2024, the researchers worked on modifying and refining hybrid and lightweight transformer models. Lili et al., (2024) developed ViT, CNNs, 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. Cheng et al., (2024) introduced SwinCrack, a hybrid U-shaped model that integrates the Swin transformer and convolutional modules for global-local feature extraction. The model outperformed CNN and transformer-only models for six databases and improved crack boundary precision and robustness under noise. Song et al., (2024) developed iSwin-Unet, which uses residual Swin transformer blocks and skip attention modules to enhance global-local feature fusion. The model achieved the best F1 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. Hubing et al., (2024) applied SegFormer for crack identification, leveraging a hierarchical transformer encoder and a lightweight decoder to model global context. This model outperformed the CNN model in Dice score (92.41%) and F1 score (91.86%), especially for thin cracks, while maintaining efficiency across model sizes. Zaiyan et al., (2024) introduced an Improved Segmentation Transformer-based Distress Network, built on SegFormer, with blended attention and transposed convolutional upsampling 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, Wen-Qing et al., (2024) presented a Lightweight Transformer Patch Labeling Network by utilizing the Swin Transformer 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 Fady et al., (2024) combined ViT with feature selection techniques to optimize crack identification. 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.

 doi:10.20944/preprints202603.1912.v1

Collectively, all these research showcases a clear direction: from hybrid CNN–ViT frameworks to highly specialized lightweight transformers, the domain 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.

SUMMARY

This study emphasizes the impact of deep learning models in automated pavement distress detection and pavement monitoring. Conventional manual methods are expensive, time-intensive, and require traffic control. Alternatively, advanced AI techniques offer efficient, precise, and scalable solutions.

Three deep learning architectures, including YOLO, CNN, and ViT, were reviewed. Among the reviewed architectures, YOLO is projected for real-time object detection, making it appropriate for on-site pavement monitoring. CNNs remain a cornerstone for image-based distress classification and segmentation because of their ability to learn spatial hierarchies of features. In recent times, ViTs have emerged as a powerful alternative that leverages a self-attention mechanism to capture global context and outperform CNNs in complex conditions, especially with large datasets.

Overall, deep learning-driven methods have proven revolutionary for assessing pavement conditions, enabling proactive maintenance and contributing to safe, sustainable transportation infrastructure maintenance practices. Table 1 summarizes the application of different deep learning architectures for automated pavement distress detection.

Table 1. Summary of application of deep learning in pavement distress detection.

Author	Application, performance, and algorithm
Owor et al., 2025	Application: Pavement distress (longitudinal, transverse, alligator, block, cracks, Patches, Manholes) segmentation Performance: Dice improvement ≈ 35% over Segment Anything Model Algorithms: ViT, CNN
Ansari et al., 2025	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., 2025	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., 2025	Application: Crack segmentation () with Context-Aware Global Module Performance: mIoU - 82.91%, Pixel Accuracy- 96.12% Algorithm: CNN
Hu et al., 2025	Application: Distress (Cracks, Rutting, Potholes) detection on mobile device with MobileNet V2 Performance: Accuracy ≈ 96.38%; model size 0.6 MB Algorithm: CNN, ResNet
Lv et al., 2025	Application: A review highlighting ViT/PVT and hybrid models Algorithm: ViT, PVT, CNN, YOLO series, Faster R-CNN, CrackGAN

 doi:10.20944/preprints202603.1912.v1

Li et al., 2025	Application: 3D distress (Polishing & Raveling) analysis with ASViT-Net. Performance: Achieved high accuracy (MPA ≈ 85.75%; MIoU ≈ 83.78%) Algorithm: ViT
Alshawabkeh et al., 2025	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., 2024	Application: Real-time crack (Longitudinal, Transverse, Block and Alligator cracks) detection Performance: 83 FPS with iBiSeNetv2; F1 score: 10.14% Algorithm: iBiSeNetv2, Swin Transformer, Crack Transformer
Wang et al., 2024	Application: Crack (Longitudinal, Transverse, Block and Alligator cracks) detection Performance: ViT higher Optimal Image Scale score - 0.849 Algorithm: Swin Transformer, CNN, Convolutional Swin-Transformer Block, Depth-convolution Forward Network
Cano-Ortiz et al., 2024	Application: Multi-defect (Alligator cracks, Longitudinal cracks, Pothole, Raveling, Patch, Sealed crack, Drain, Manhole) detection Performance: mAP@0.5 ≈ 0.59; integrated GIS Algorithm: YOLOv5, DeepCrack, Faster R-CNN
Zhang et al., 2024b	Application: Crack detection PerformanceF1 ≈ 90.13%; mIoU ≈ 82.68% Algorithm: CNN-based architecture “DARNet”
Zheng et al., 2024	Application: Deep learning based intelligent distress detection Performance: YOLOv5l (88.1%), YOLOv8-DGS (91.6% precision), Optimized CNN: 98.2%
Zhang et al., 2024c	Application: Multi-scene distress (Cracks, Potholes) Performance: mAP50 ≈ 79.4% Algorithm: SMG-YOLOv8 based on YOLOv8s
Tao et al., 2024	Application: Distress (Cracks, raveling, spalling) Performance: mAP ≈ 49.2% Algorithm: YOLOv5
Liu et al., 2024	Application: Multi-distress (Cracks, Pothole) detection Performance: Fusion images mAP ≈91.3%, Accuracy ≈ 93.5% Algorithm: R-CNN and YOLO series
Fan et al., 2024	Application: Review of pavement defect (Cracks, Potholes, Rutting, Raveling, Patching, Bleeding, Depressions) detection algorithms Algorithm: CNN based AlexNet, VGGNet, ResNet, U-Net, FCN, YOLO series, SDD
Wang et al., 2024	Application: Integration of Swin transformer with YOLOv8 for distress (traverse, longitudinal, cross and alligator cracks, potholes) detection

 doi:10.20944/preprints202603.1912.v1

	Performance: Swin Transformer integration (MAP 94.8%) Algorithm: Enhanced YOLOv8, Swin Transformer module
Chen et al., 2024	Application: Crack segmentation Performance: F1 score ≈ 95% Algorithm: iSwin-Unet, Swin-Unet, Swin Transformer, U-Net
Guo et al., 2024	Application: Crack detection transformer-based model CTv2 Performance: High precision and recall (>90%) Algorithm: Crack Transformer v2, CNN
Li et al., 2024	Application: Crack segmentation with SegFormer Performance: SegFormer (Dice ≈ 0.84; accuracy ~96.3%) Algorithm: SegFormer, CNN based models
Zhang et al., 2024d	Application: Crack segmentation with Mix-Graph CrackNet Performance: F-measure ≈ 90.37%; IoU ≈ 82.43% Algorithm: Integrated CNN and Transformer architecture.
Zhang et al., 2024e	Application: Multi-distress (Cracks, Potholes, Patches) segmentation Performance: F1 ≈ 95.51%; mIoU ≈ 91.67% Algorithm: ISTD-DisNet (SegFormer backbone)
Huang et al., 2024	Application: Distress (Cracks, Potholes, Repairs, Rutting) detection Performance: AUC ≈ 94.2%; real-time 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., 2024	Application: Crack classification Performance: F1 ≈ 98.72% Algorithm: CNN
Wu et al., 2024	Application: Crack (Longitudinal, Transverse, Alligator) detection Performance: Crack500 (F1 score: 70.84%), DeepCrack (F1 score: 84.50%) Algorithm: Multi-scale feature fusion (CNN-based or Transformer-based), SSD, YOLOv8
Liu et al., 2024	Application: Introduced dataset for distress (Cracks, Patches, Potholes, Background images without defects) detection. Algorithm: CNN, Mask R-CNN, DeepLab v3+, Hybrid CNN +ViT
Zaman et al., 2024	Application: Crack Detection Performance: Accuracy (99%) Algorithm: ViT, CNN
Xiao et al., 2023	Application: Crack segmentation with hybrid-window transformer architecture Performance: F1 score ≈ 0.9364, Precision: 0.9376, Recall: 0.9352 Algorithm: CrackForme (ViT backbone)
Wang et al., 2023	Application: Detection & classification of asphalt pavement cracks

 doi:10.20944/preprints202603.1912.v1

	Performance: ViT-improved YOLO V5 enhances detection accuracy & speed (11.9 ms/image), enabling real-time crack detection. Algorithm: YOLOv5, ViT
Guo et al., 2023a	Application: Crack segmentation Performance: 109 FPS; high F1 and recall Algorithm: CT (VIT based model using Swin transformer), CrackFormer, CNN based - U-Net, DeepLabv3+, HRNet.
Guo et al., 2023b	Application: Crack detection Performance: High recall and accuracy (>90%); 115 FPS Algorithm: STA (Swin Transformer + UperNet), CNN
Liu et al., 2023	Application: Crack (Thin, Thick, Grid cracks) segmentation Performance: F1 ≈ 0.912; fewer parameters than CNN Algorithm: CNN, VIT
Luo et al., 2023	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., 2023	Application: Distress (Cracks (longitudinal, transverse, block cracks), Potholes, Repair areas, Background pavement) segmentation Performance: Better uncertainty calibration; high pixel accuracy Algorithm: ES-transformer (ViT+ Dempster-Shafer Theory (DST)), Baseline model: P-FCN
Zhang et al., 2023	Application: Multi-distress segmentation Performance: F-measure ≈ 94.21%; IoU ≈ 0.8914 Algorithm: CNN+ Transformers
Li et al., 2023	Application: Pavement distress (Alligator, Longitudinal, Block and Transverse crack, Pothole, Patch) classification and detection Performance: Accuracy 96.24%, Algorithm: ResNet-34, ResNet-50, VGG-16, VGG-19
Ali et at., 2023	Application: Crack detection, localization, and segmentation Performance: Accuracy (96%, test accuracy) Algorithm: ViT, CNN-based baselines: U-Net, FCN, DeepLabv3+
Lin et al., 2023	Application: Distress (Cracks (including longitudinal and transverse), Potholes, Spalling, Faulting) Algorithm: UNet-based models, FCN, YOLACT
Zheng et al., 2023	Application: Real-time concrete pavement crack detection on edge devices Performance: Precision 0.896, lightweight design, 89 FPS real-time performance. Algorithm: Bottleneck Transformer and YOLOv5
Xu et al., 2022	Application: Crack detection Performance: Precision: ~93.04%, Recall: ~92.85% Algorithm: ViT based architecture

 doi:10.20944/preprints202603.1912.v1

Shamsabadi et al., 2022	Application: Crack detection in asphalt and concrete surfaces with TransUNet Architecture Performance: TransUNet (CNN-ViT) achieved ~61% better mean IoU than DeepLabv3+ Algorithm: CNN + ViT backbone, Baselines: DeepLabv3+, U-Net
Guan et al., 2021	Application: Distress (Transverse, Longitudinal, Block, Alligator, Edge, Reflection, Potholes, Raveling) detection & classification Performance: Captures global context for complex cracks, F1~ 91.9% with hybrid model Algorithm: CNN, ViT
Gopalakrishnan., 2018	Application: Automated pavement distress detection using 2D/3D images Performance: High detection accuracy ~99.8%, better generalization than traditional method Algorithm: CNN

DISCUSSION

A CNN is a generic deep learning backbone used for several vision tasks. It is generally applied for tasks such as cracks classification, damage severity estimation, semantic segmentation (pixel-wise crack mapping). CNN showcases high accuracy for pixel-level crack segmentation and distress severity mapping. However, this algorithm is not suitable for real-time analysis unless heavily optimized and relies heavily on spatial hierarchies that sometimes limited their capability to capture global relationships. In addition to that, CNN requires post-processing for localization. As the datasets grew eventually and required more flexible models, researchers started looking for architecture which can look at an image more comprehensively.

ViTs were developed to solve some of these shortcomings. ViT applies self-attention instead of relying on convolution and it obtain the attention mechanism from NLP transformers to capture global dependencies across the whole image. This allow ViT to understand spatial relationships more effectively and scale better with large datasets. ViT outperforms CNN when trained on large datasets because it can learn and understand more general representations. However, ViT also has some limitations such as it requires high computational power and large database to meet full potential. This requirement makes ViT less suitable for small and real-time applications.

Furthermore, object detection faces another challenge, that is not just identifying the object but determining the exact location of the object. To solve this problem, YOLO was developed to perform object detection very quickly for real time scenarios such as surveillance, autonomous driving and video analytics. YOLO is designed specifically to localize and classify multiple objects in a single pass which makes it more suitable than ViT and CNNs. Both CNN and ViTs require additional frameworks such as R-CNN, Faster R-CNN, or DETR which may not match YOLO's performance. YOLO further enhanced in series of versions such as YOLOv3, YOLOv5, YOLOv7, and YOLOv8, and continuously improved speed and accuracy.

Overall, ViT works well with global reasoning, CNN outperforms features extraction and YOLO performs well in quick and precise object detection. Each model developed to solve different problems and not to replace the previous one.

FUTURE DIRECTION

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, requirement of large training time of models and unsupervised learning. The problem of the

 doi:10.20944/preprints202603.1912.v1

conventional deep learning algorithms can be solved using Large Language Models (LLMs) and Digital Twins (DTs).

Although deep learning has been extensively used by researchers for pavement monitoring, some challenges exist in the pavement industry, such as the generalizability of models, the need for long training times, and the use of unsupervised learning. The limitations of conventional deep learning algorithms can be overcome using Large Language Models (LLMs) and Digital Twins (DTs).

LLMs are based on the transformer architecture 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.

References

1. Ali, L., Jassmi, H. A., Khan, W., & Alnajjar, F. (2023). Crack45K: integration of vision transformer with tubularity flow field (TuFF) and sliding-window approach for crack-segmentation in pavement structures. *Buildings*, *13*(1), 55.
2. Alshawabkeh, S., Wu, L., Dong, D., Cheng, Y., & Li, L. (2025). A Hybrid Approach for Pavement Crack Detection Using Mask R-CNN and Vision Transformer Model. *Computers, Materials & Continua*, *82*(1).
3. Ansari, F., Chatterjee, K., Li, J. Q., Wang, K., & Golalipour, A. Multi-Object Pavement Surface Feature Detection with CNN and Transformer Deep Learning Architecture. In *Airfield and Highway Pavements 2025* (pp. 350-359).
4. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.
5. Cano-Ortiz, S., Iglesias, L. L., del Árbol, P. M. R., Lastra-González, P., & Castro-Fresno, D. (2024). An end-to-end computer vision system based on deep learning for pavement distress detection and quantification. *Construction and Building Materials*, *416*, 135036.
6. Chatterjee, K., Vivanco, D., Yang, X., & Li, J. Q. (2024). Enhancing Pavement Performance through Balanced Mix Design: A Comprehensive Field Study in Oklahoma. In *International Conference on Transportation and Development 2024* (pp. 511-522).
7. Chatterjee, K., Li, J., Parajulee, K., & Schwennesen, J. (2025). Network Level Evaluation of Hangup Susceptibility of HRGCs using Deep Learning and Sensing Techniques: A Goal Towards Safer Future. *arXiv preprint arXiv:2512.12832*.
8. Chatterjee, K., Li, J. Q., Ansari, F., Munna, M. R., Parajulee, K., & Schwennesen, J. (2026a). Hybrid LSTM-Transformer Models for Profiling Highway–Railway Grade Crossings. *Journal of Transportation Engineering, Part A: Systems*, *152*(2), 04025138.
9. Chatterjee, K., Desai, M., & Li, J. (2026b). Application of Large Language Models in Geotechnical Engineering: A Movement Towards Safe and Sustainable Future.
10. Chen, S., Feng, Z., Xiao, G., Chen, X., Gao, C., Zhao, M., & Yu, H. (2024). Pavement crack detection based on the improved Swin-Unet model. *Buildings*, *14*(5), 1442.
11. Desai, M., & Chatterjee, K. (2026). Application of Machine Learning Techniques for Prediction of Soil Water Characteristics Curve: A State of the Art Review.
12. Elghaish, F., Matarneh, S., Abdellatef, E., Rahimian, F., Hosseini, M. R., & Farouk Kineber, A. (2025). Multi-layers deep learning model with feature selection for automated detection and classification of highway pavement cracks. *Smart and Sustainable Built Environment*, *14*(2), 511-535.
13. Fan, L., Wang, D., Wang, J., Li, Y., Cao, Y., Liu, Y., ... & Wang, Y. (2023). Pavement defect detection with deep learning: A comprehensive survey. IEEE Transactions on Intelligent Vehicles, 9(3), 4292-4311.

 doi:10.20944/preprints202603.1912.v1

14. Guan, S., Liu, H., Pourreza, H. R., & Mahyar, H. (2023). Deep learning approaches in pavement distress identification: A review. arXiv preprint arXiv:2308.00828.
15. Guo, F., Qian, Y., Liu, J., & Yu, H. (2023a). Pavement crack detection based on transformer network. Automation in Construction, 145, 104646.
16. Guo, F., Liu, J., Lv, C., & Yu, H. (2023b). A novel transformer-based network with attention mechanism for automatic pavement crack detection. Construction and Building Materials, 391, 131852.
17. Guo, F., Liu, J., Xie, Q., & Yu, H. (2024). A two-stage framework for pixel-level pavement surface crack detection. Engineering Applications of Artificial Intelligence, 133, 108312.
18. Gopalakrishnan, K. (2018). Deep learning in data-driven pavement image analysis and automated distress detection: A review. Data, 3(3), 28.
19. Hu, Y., Chen, N., Hou, Y., Lin, X., Jing, B., & Liu, P. (2025). Lightweight deep learning for real-time road distress detection on mobile devices. *Nature Communications*, *16*(1), 4212.
20. Huang, W. Q., Feng, L., & He, Y. L. (2024). LTPLN: Automatic pavement distress detection. *PLoS One*, *19*(10), e0309172.
21. Kyem, B. A., Asamoah, J. K., & Aboah, A. (2025). Context-cracknet: A context-aware framework for precise segmentation of tiny cracks in pavement images. *Construction and Building Materials*, *484*, 141583.
22. Li, D., Duan, Z., Hu, X., Zhang, D., & Zhang, Y. (2023). Automated classification and detection of multiple pavement distress images based on deep learning. *Journal of Traffic and Transportation Engineering (English Edition)*, *10*(2), 276-290.
23. Li, H., Zhang, H., Zhu, H., Gao, K., Liang, H., & Yang, J. (2024). Automatic crack detection on concrete and asphalt surfaces using semantic segmentation network with hierarchical Transformer. *Engineering Structures*, *307*, 117903.
24. Li, Y., Liu, C., Weng, Z., Wu, D., & Du, Y. (2025). Aggregate-level 3D analysis of asphalt pavement deterioration using laser scanning and vision transformer. *Automation in Construction*, *178*, 106380.
25. Lin, W., Li, X., Han, H., Yu, Q., & Cho, Y. H. (2023). A novel approach for pavement distress detection and quantification using RGB-D camera and deep learning algorithm. *Construction and Building Materials*, *407*, 133593.
26. Liu, F., Liu, J., Wang, L., & Al-Qadi, I. L. (2024). Multiple-type distress detection in asphalt concrete pavement using infrared thermography and deep learning. *Automation in Construction*, *161*, 105355.
27. Liu, H., Yang, J., Miao, X., Mertz, C., & Kong, H. (2023). Crackformer network for pavement crack segmentation. *IEEE Transactions on Intelligent Transportation Systems*, *24*(9), 9240-9252.
28. Liu, Z., Wu, W., Gu, X., & Cui, B. (2024). PaveDistress: A comprehensive dataset of pavement distresses detection. *Data in Brief*, *57*, 111111.
29. Luo, H., Li, J., Cai, L., & Wu, M. (2023). STrans-YOLOX: Fusing swin transformer and YOLOX for automatic pavement crack detection. *Applied Sciences*, *13*(3), 1999.
30. Lv, Z., Hao, Z., Zhu, Y., & Lu, C. (2025). A Review on Automated Detection and Identification Algorithms for Highway Pavement Distress. *Applied Sciences*, *15*(11), 6112.
31. Munna, M. R., Chatterjee, K., & Li, J. Q. (2025). Evaluating Pavement Surface Texture with LTPP Database. *International Journal of Pavement Research and Technology*, 1-14.
32. Owor, N. J., Adu-Gyamfi, Y., Aboah, A., & Amo-Boateng, M. (2025). PaveSAM–segment anything for pavement distress. *Road Materials and Pavement Design*, *26*(3), 593-617.
33. Parajulee, K., Chatterjee, K., & Li, J. (2025). Leveraging original equipment manufacturer vehicle sensor data for enhanced roadway safety. *International Journal of Pavement Research and Technology*, 1-18.
34. Rana Munna, M., Chatterjee, K., Parajulee, K., & Li, J. Q. Effect of Pavement Surface Characteristics on Adverse Road Conditions. In Airfield and Highway Pavements 2025 (pp. 360-369).
35. Shamsabadi, E. A., Xu, C., Rao, A. S., Nguyen, T., Ngo, T., & Dias-da-Costa, D. (2022). Vision transformer-based autonomous crack detection on asphalt and concrete surfaces. *Automation in Construction*, *140*, 104316.
36. Tao, Z., Gong, H., Liu, L., Cong, L., & Liang, H. (2025). A weakly-supervised deep learning model for end-to-end detection of airfield pavement distress. *International Journal of Transportation Science and Technology*, *17*, 67-78.

 doi:10.20944/preprints202603.1912.v1

37. Tong, Z., Ma, T., Zhang, W., & Huyan, J. (2023). Evidential transformer for pavement distress segmentation. *Computer-Aided Civil and Infrastructure Engineering*, *38*(16), 2317-2338.
38. Wang, C., Liu, H., An, X., Gong, Z., & Deng, F. (2024). SwinCrack: Pavement crack detection using convolutional swin-transformer network. *Digital Signal Processing*, *145*, 104297.
39. Wang, S., Cai, B., Wang, W., Li, Z., Hu, W., Yan, B., & Liu, X. (2024). Automated detection of pavement distress based on enhanced YOLOv8 and synthetic data with textured background modeling. *Transportation Geotechnics*, *48*, 101304.
40. Wang, S., Chen, X., & Dong, Q. (2023). Detection of asphalt pavement cracks based on vision transformer improved YOLO V5. *Journal of Transportation Engineering, Part B: Pavements*, *149*(2), 04023004.
41. Wu, P., Wu, J., & Xie, L. (2024). Pavement distress detection based on improved feature fusion network. *Measurement*, *236*, 115119.
42. Wu, W., Zhu, F., Li, Z., Li, X., Li, X., & Wang, J. (2025). Optimized deep learning model with integrated spectrum focus transformer for pavement distress recognition and classification. *Scientific Reports*, *15*(1), 3803.
43. Xiao, S., Shang, K., Lin, K., Wu, Q., Gu, H., & Zhang, Z. (2023). Pavement crack detection with hybrid-window attentive vision transformers. *International Journal of Applied Earth Observation and Geoinformation*, *116*, 103172.
44. Xu, Z., Guan, H., Kang, J., Lei, X., Ma, L., Yu, Y., ... & Li, J. (2022). Pavement crack detection from CCD images with a locally enhanced transformer network. *International Journal of Applied Earth Observation and Geoinformation*, *110*, 102825.
45. Yang, X., Li, J. Q., Zhan, Y., & Yu, W. (2024). Real-time automated deep learning based railroad trespassing violation detection and tracking at highway-rail grade crossing. Intelligent Transportation Infrastructure, 3, liae003.
46. Yu, H., Deng, Y., & Guo, F. (2024). Real-time pavement surface crack detection based on lightweight semantic segmentation model. *Transportation Geotechnics*, *48*, 101335.
47. Zaman, F., Xu, Z., Hussain, A., Aslam, A., & Abdullahi, M. R. (2024, September). Concrete and Pavement Cracks Detection Leveraging Vision Transformer (ViT) Model. In *2024 International Conference on Electrical and Information Technology (IEIT)* (pp. 90-95). IEEE.
48. Zhang, H., Zhang, A. A., Dong, Z., He, A., Liu, Y., Zhan, Y., & Wang, K. C. (2024a). Robust semantic segmentation for automatic crack detection within pavement images using multi-mixing of global context and local image features. *IEEE Transactions on Intelligent Transportation Systems*, *25*(9), 11282-11303.
49. Zhang, J., Sun, S., Song, W., Li, Y., & Teng, Q. (2024b). Automated Pavement Distress Detection Based on Convolutional Neural Network. *IEEE Access*.
50. Zhang, S., Bei, Z., Ling, T., Chen, Q., & Zhang, L. (2024c). Research on high-precision recognition model for multi-scene asphalt pavement distresses based on deep learning. *Scientific Reports*, *14*(1), 25416.
51. Zhang, H., Zhang, A. A., He, A., Dong, Z., & Liu, Y. (2024d). Pixel-level detection of multiple pavement distresses and surface design features with ShuttleNetV2. *Structural Health Monitoring*, *23*(2), 1263-1279.
52. Zhang, Z., Song, W., Zhuang, Y., Zhang, B., & Wu, J. (2024e). Automated Multi-Type Pavement Distress Segmentation and Quantification Using Transformer Networks for Pavement Condition Index Prediction. *Applied Sciences*, *14*(11), 4709.
53. Zheng, L., Xiao, J., Wang, Y., Wu, W., Chen, Z., Yuan, D., & Jiang, W. (2024). Deep learning-based intelligent detection of pavement distress. *Automation in Construction*, *168*, 105772.
54. Zheng, X., Qian, S., Wei, S., Zhou, S., & Hou, Y. (2023). The combination of transformer and you only look once for automatic concrete pavement crack detection. Applied Sciences, 13(16), 9211.