Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.057213

ARTICLE

A Hybrid Approach for Pavement Crack Detection Using Mask R-CNN and Vision Transformer Model

Shorouq Alshawabkeh, Li Wu*, Daojun Dong, Yao Cheng and Liping Li

Faculty of Engineering, China University of Geosciences, Wuhan, 430074, China

*Corresponding Author: Li Wu. Email: lwu@cug.edu.cn

Received: 11 August 2024 Accepted: 14 October 2024 Published: 03 January 2025

ABSTRACT

Detecting pavement cracks is critical for road safety and infrastructure management. Traditional methods, relying on manual inspection and basic image processing, are time-consuming and prone to errors. Recent deep-learning (DL) methods automate crack detection, but many still struggle with variable crack patterns and environmental conditions. This study aims to address these limitations by introducing the MaskerTransformer, a novel hybrid deep learning model that integrates the precise localization capabilities of Mask Region-based Convolutional Neural Network (Mask R-CNN) with the global contextual awareness of Vision Transformer (ViT). The research focuses on leveraging the strengths of both architectures to enhance segmentation accuracy and adaptability across different pavement conditions. We evaluated the performance of the MaskerTransformer against other state-of-the-art models such as U-Net, Transformer U-Net (TransUNet), U-Net Transformer (UNETr), Swin U-Net Transformer (Swin-UNETr), You Only Look Once version 8 (YoloV8), and Mask R-CNN using two benchmark datasets: Crack500 and DeepCrack. The findings reveal that the MaskerTransformer significantly outperforms the existing models, achieving the highest Dice Similarity Coefficient (DSC), precision, recall, and F1-Score across both datasets. Specifically, the model attained a DSC of 80.04% on Crack500 and 91.37% on DeepCrack, demonstrating superior segmentation accuracy and reliability. The high precision and recall rates further substantiate its effectiveness in real-world applications, suggesting that the MaskerTransformer can serve as a robust tool for automated pavement crack detection, potentially replacing more traditional methods.

KEYWORDS

Pavement crack segmentation; transportation; deep learning; vision transformer; Mask R-CNN; image segmentation

1 Introduction

Pavement crack detection is a crucial aspect of infrastructure maintenance, as cracks can severely compromise the structural integrity of roads and highways [1–3]. Timely detection is essential for preventing further damage, reducing maintenance costs, and ensuring public safety [4]. Traditionally, pavement crack detection relied on manual inspections, which are labor-intensive, time-consuming, and subject to human error. In recent years, advances in machine learning (ML) and deep learning (DL) have revolutionized the field, enabling more accurate and automated detection systems [5,6].

The advent of automated pavement inspection systems has revolutionized this field by providing a more efficient and accurate means of detecting cracks [4–6]. These systems leverage advanced image processing techniques and ML models to identify and classify pavement defects. Among the various approaches, the integration of DL models has shown remarkable promise due to their ability to learn complex patterns and features directly from raw data [7].

Several recent studies have explored various DL architectures for crack detection. Liu et al. [8] proposed a transfer learning-based encoder-decoder model with visual explanations for infrastructure crack segmentation. Their study emphasized the importance of transfer learning for improving segmentation accuracy and reducing the computational complexity of models in diverse environmental conditions. Another notable contribution by [9] presented a UNet-based model that integrates visual explanations to address the "black box" nature of Convolutional Neural Networks (CNNs). This model showed improvements in transparency and performance through the inclusion of visual explanations and deeper model architectures. Infrared thermography (IRT) has also been used in conjunction with convolutional neural networks to enhance crack detection capabilities. Despite the advancements, several challenges persist in the accurate detection of pavement cracks. Variability in crack appearances, such as differences in width, length, and orientation, along with environmental factors like lighting and shadow, pose significant obstacles to developing a robust detection model [5,10]. Moreover, the presence of noise and artefacts in images captured in real-world scenarios further complicates the detection process [11].

Recent developments in DL, particularly in computer vision, have introduced innovative architectures that can address these challenges. The Mask R-CNN [12], a popular model for object detection and segmentation, offers precise localization and classification capabilities. Meanwhile, the Vision Transformer (ViT) model, known for its effectiveness in capturing long-range dependencies in images, provides a complementary approach to feature extraction that can enhance the performance of traditional convolutional networks [13]. This study proposes a novel hybrid model that combines the strengths of Mask R-CNN and ViT to improve the accuracy and efficiency of pavement crack detection. By integrating the region proposal and segmentation capabilities of Mask R-CNN with the feature extraction power of ViT, the hybrid model aims to achieve superior performance in crack detection, addressing the gaps in existing methods. This research's main aim is to develop a robust and scalable pavement crack detection system that can operate effectively in diverse and challenging environments. The proposed hybrid model will be evaluated against state-of-the-art methods to demonstrate its efficacy in terms of accuracy, speed, and adaptability to varying conditions.

This paper is organized as follows: Section 2 reviews related work in the domain of pavement crack detection and deep learning models. Section 3 details the methodology and architecture of the proposed hybrid model. Section 4 presents the experimental setup and results, followed by a discussion in Section 5.

2 Related Work

Pavement crack detection is a critical task in infrastructure maintenance, essential for ensuring road safety and minimizing maintenance costs [5]. Traditional manual inspection methods have been largely replaced by automated systems leveraging advanced machine learning techniques. Among these, CNNs have become a cornerstone in detecting and segmenting cracks due to their ability to learn hierarchical features from image data [14–16].

For instance, a study by Xu et al. [17] compared the performance of Mask R-CNN and Faster R-CNN for pavement crack detection. The results indicated that Mask R-CNN outperformed You

Only Look Once (YOLOv3) and was particularly effective for crack detection tasks, achieving higher accuracy in identifying complex crack patterns. However, the study also noted that the bounding box precision of Mask R-CNN could degrade under certain joint training strategies. Wang et al. [18] developed an automatic pavement crack recognition system using the Mask R-CNN model. This study focused on the model's ability to segment cracks at the pixel level and highlighted the importance of dataset size, image resolution, and labelling techniques in enhancing model performance. The research demonstrated that Mask R-CNN could effectively segment cracks, improving maintenance management by facilitating timely inspections.

Ukhwah et al. [19] implemented the YOLO model with three configurations (YOLOv3, YOLOv3 Tiny, and YOLOv3 SPP) for pothole detection. The configurations achieved mean Average Precision (mAP) values of 83.43%, 79.33%, and 88.93%, respectively, and demonstrated rapid detection speeds. This approach shows significant potential for enhancing road assessment processes.

Ghosh et al. [20] applied faster region-based convolutional neural networks (R-CNN) and You Only Look Once (YOLO) v3 to detect and classify distresses in high-resolution 3D images of pavements. The dataset included 625 images with annotated distresses and 798 without. Data augmentation was used to balance class representation and prevent overfitting. Both YOLO and Faster R-CNN showed strong performance, with accuracies of 89.8% and 89.6%, and average precision (AP) values of 90.2% and 89.2%, respectively. ROC curves indicated robust performance with the area under the curve (AUC) values of 0.96 for YOLO and 0.95 for Faster R-CNN. Evaluations against manual quality assurance and quality control (QA/QC) showed high agreement, suggesting the methodology's potential to replace manual QA/QC in automated pavement analysis.

Liu et al. [21] proposed a two-step deep-learning model for automated pavement crack detection and segmentation. The model utilized a modified YOLOv3 in the first step to detect cracks and a modified U-Net in the second step to segment the detected cracks. The study introduced a pre-trained ResNet-34 encoder and spatial and channel squeeze and excitation (SCSE) modules to improve segmentation accuracy. The proposed method achieved an F1-Score of 90.58% for crack detection and 95.75% for crack segmentation, outperforming other state-of-the-art models. This two-step approach demonstrated advantages in accuracy over one-step crack detection and segmentation methods. Liu et al. [22] applied deep learning and infrared thermography (IRT) to classify asphalt pavement crack severity. The study used a dataset with four levels of crack severity (no crack, low, medium, and high severity) and three types of images (visible, infrared, and fusion). The work compared 13 CNN models trained from scratch and 8 pre-trained CNNs using transfer learning. The findings showed that EfficientNet-B3 achieved the highest accuracy across all image types, with fusion images providing the best results for deep learning from scratch, while visible images were more effective for transfer learning.

While traditional models like YOLO and Faster R-CNN have demonstrated high accuracy and precision in pavement distress detection, they face several limitations. These models can struggle with detecting fine-grained crack details due to their reliance on convolutional operations, which may not capture long-range dependencies effectively. Additionally, the performance of these models can degrade in varying lighting conditions and complex road textures [17]. Our proposed hybrid model, combining Mask R-CNN and vision transformer, addresses these weaknesses by leveraging the precise segmentation capabilities of Mask R-CNN and the global context understanding of ViT. This combination enhances the model's ability to detect subtle and complex crack patterns with improved accuracy and robustness across diverse environmental conditions, offering a more comprehensive solution for pavement distress detection.

3 Proposed Methodology

In this section, we present the MaskerTransformer, a novel hybrid model designed to enhance the detection and segmentation of pavement cracks. By combining the localized precision of Mask R-CNN with the global contextual information using ViT, the model is optimized to handle various crack patterns and environmental conditions. The following subsections outline the architecture, data flow, and training configuration of the proposed model, illustrating how each component contributes to improving crack detection performance.

3.1 Dataset Description

The Crack500 dataset [23] is a collection of 500 high-resolution images curated for pavement crack detection. Each image is annotated with pixel-level precision, which makes this dataset ideal for training and evaluating deep learning models focused on crack detection and segmentation tasks. Crack500 includes a variety of crack patterns and types, allowing for the development of models that can generalize across different conditions and environments. This comprehensive dataset provides a robust foundation for testing the accuracy and reliability of pavement crack detection algorithms. The DeepCrack dataset [24] serves as a benchmark for evaluating deep learning models for crack detection. It comprises a diverse set of images capturing various crack types and road conditions (Fig. 1), offering a challenging test bed for model performance. Detailed annotations accompany each image, highlighting crack locations and dimensions, which supports segmentation tasks. These datasets are widely used in research to validate the effectiveness of new crack detection algorithms, providing a comprehensive resource for developing advanced detection solutions.

3.2 Preprocessing

In this study, several preprocessing techniques were applied to enhance the quality and diversity of the training data. These techniques include data normalization, flip, random crop, mask crop, and shuffle, each playing a vital role in preparing the dataset for modelling. Data normalization is a technique used to scale the pixel values of images to a standard range, often between 0 and 1, to ensure that all input features contribute equally to the model's learning process [25]. This can help improve the convergence speed and stability of the training process. The formula for normalization is:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where x is the original pixel value, and x' is the normalized pixel value.

Flipping is a data augmentation technique that involves mirroring the images horizontally or vertically. This helps increase the dataset's diversity and allows the model to learn invariant features under different orientations [26]. The horizontal flip is commonly used and can be represented as:

$$I'(i,j) = I(i, W - j - 1) \tag{2}$$

where I is the original image, I' is the flipped image, i and j are the pixel indices, and W is the width of the image. Random cropping involves selecting a random portion of the image and cropping it to a specified size. This augmentation helps the model focus on different parts of the image and learn to detect features at various scales and locations [27]. The cropping can be defined by selecting random offsets o_x and o_y:

$$I'(i,j) = I\left(i + o_x, j + o_y\right) \tag{3}$$

where o_x and o_y are randomly chosen offsets within the allowable range.

Figure 1: Samples of pavement and cracking

Mask cropping is a specialized form of cropping where the region of interest, often determined by a mask or bounding box, is extracted from the image. This is particularly useful for tasks requiring focused attention on specific areas, such as crack detection in images [28]. Shuffling involves randomly rearranging the order of the data in the dataset. This process ensures that the model does not learn any unintended patterns from the order of the input data and helps improve generalization by mixing up the data batches during training.

3.3 Mask R-CNN

Mask R-CNN is a state-of-the-art deep learning architecture developed for the tasks of object detection and instance segmentation. As an extension of the Faster R-CNN model, Mask R-CNN incorporates an additional branch that predicts segmentation masks, enabling the model to not only classify objects within an image but also provide detailed pixel-level segmentation for each identified

instance [12]. This dual capability renders Mask R-CNN particularly suitable for applications that require precise object localization and segmentation, such as the detection of pavement cracks. The architecture of Mask R-CNN comprises several critical components [28]. First, a backbone network, typically a convolutional neural network like ResNet or ResNeXt, is utilized to extract feature maps from input images. These feature maps form the foundation for further processing within the network. Following this, a region proposal network (RPN) is employed to generate candidate object proposals, identifying regions within the image that are likely to contain objects. The RPN outputs a set of bounding boxes along with associated objectness scores, which indicate the likelihood of each region containing an object [29]. A key innovation in Mask R-CNN is the use of region of interest (RoI) Align, which replaces the RoI Pooling layer found in Faster R-CNN (Fig. 2). RoI Align addresses the potential misalignment issues that can occur during pooling operations by precisely mapping the input feature map to a fixed-size feature map, thereby preserving crucial spatial information necessary for accurate segmentation. For each proposed RoI, the network predicts both the object class and refines the bounding box coordinates to more accurately fit the object. In addition to these capabilities, Mask R-CNN includes a parallel branch dedicated to predicting a binary mask for each RoI. This branch outputs a spatial binary mask that indicates which pixels belong to the detected object, facilitating precise segmentation. The application of Mask R-CNN to pavement crack detection offers several advantages. Its ability to predict segmentation masks allows for the accurate delineation of crack boundaries, which is essential for quantifying crack dimensions and severity [30]. Moreover, Mask R-CNN's robustness to variability in crack appearances, such as differences in width, length, and orientation, enables it to learn to segment cracks effectively from diverse training samples. Additionally, the model can be integrated with ViT to enhance feature extraction, combining the strengths of CNNs and transformers to capture both local and global features within images.

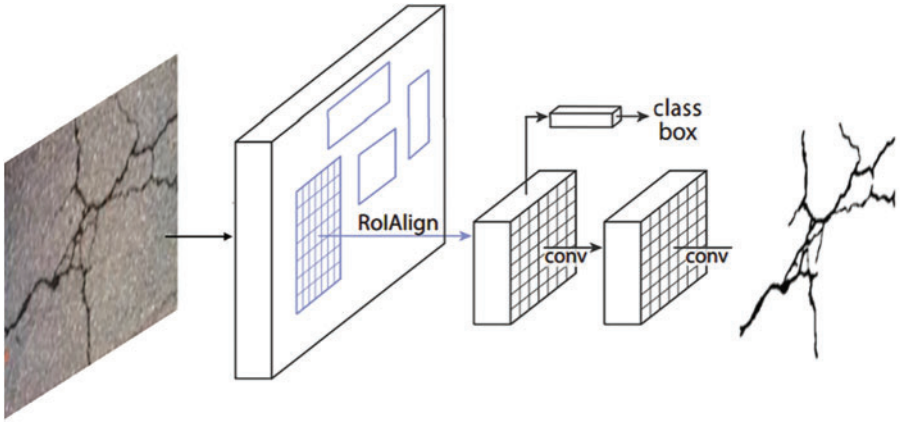

Figure 2: The Mask R-CNN framework for pavement crack segmentation

3.4 Vision Transformer (ViT)

The vision transformer is a groundbreaking model that applies the transformer architecture, originally developed for natural language processing, to image analysis [31]. Unlike traditional CNNs, ViT uses self-attention to capture global context by processing images as sequences of fixed-size patches [13]. This allows ViT to effectively learn long-range dependencies, which is beneficial for detecting intricate features like pavement cracks. ViT divides an image into non-overlapping patches, embeds them linearly, and processes them through a transformer encoder, capturing spatial

relationships across the entire image [13,31]. This approach enables ViT to recognize subtle and complex patterns that might be missed by CNNs, making it particularly suited for tasks requiring a global understanding of image data. The model's strength lies in its ability to achieve competitive performance with state-of-the-art CNNs, particularly when trained on large datasets. However, ViT's reliance on extensive training data requires careful data augmentation and preprocessing to enhance its generalization capabilities. Incorporating ViT into pavement crack detection systems offers significant advantages. By integrating ViT with models like Mask R-CNN, both local and global features can be effectively extracted, improving the accuracy and robustness of detection. Key considerations include the selection of patch size and the implementation of training strategies to prevent overfitting (Fig. 3).

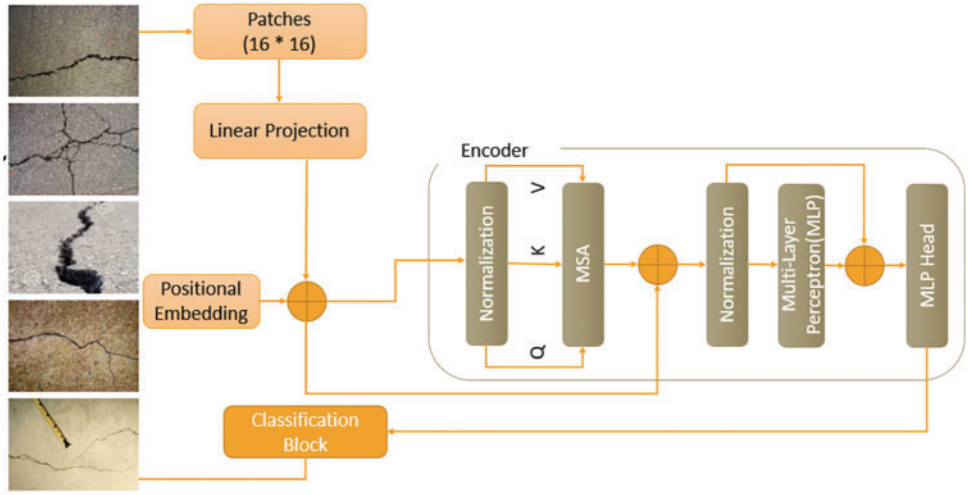

Figure 3: ViT abstract level architecture diagram

3.5 Proposed Hybrid Model

In this study, we introduce the MaskR-Transformer (Fig. 4), a novel hybrid segmentation model that merges the localized precision of Mask R-CNN with the expansive context awareness of the vision transformer. This model is meticulously designed to enhance the detection and segmentation of pavement cracks across varied imagery, as demonstrated on the Crack500 and DeepCrack datasets. The MaskR-Transformer (Table 1) employs a dual objective function approach where the Crack500 dataset utilizes a combination of Dice Loss and weighted binary cross entropy (WBCE), optimizing for a balance between class imbalances and the need for precise pixel-wise segmentation. For the DeepCrack dataset, the model relies solely on Dice Loss, which is effective for handling the sparse and uneven distribution of crack features within the images. This tailored approach to loss functions underscores our model's adaptability to different segmentation challenges. The optimization of the model is carried out using the AdamW optimizer, known for its ability to correct the weight decay integration, thus stabilizing the training process.

Furthermore, data augmentation techniques such as rotations, translations, and scaling are applied to both datasets, ensuring the robustness and generalization of the model by exposing it to a wider variety of training scenarios. Architecturally, the MaskR-Transformer is configured with different settings tailored to each dataset's specific characteristics. The model processes inputs of 256 $\times$ 256 pixels for Crack500 and 224 $\times$ 224 pixels for DeepCrack, optimizing computational efficiency while retaining sufficient detail for accurate crack detection. It incorporates 16 feature maps in the

base layer for Crack500 and 8 for DeepCrack, reflecting the varying levels of abstraction required to capture relevant features from each dataset. The utilization of 12 attention heads facilitates the model's ability to focus on pertinent features across different parts of the input image, enhancing the detection capabilities. The training configuration further refines the model's performance. Employed with a batch size tailored to each dataset—32 for Crack500 and 96 for DeepCrack—the model balances between computational load and learning stability. The training spans 300 epochs to ensure ample learning and convergence to optimal weights. A OneCycleLR scheduler dynamically adjusts the learning rate, starting from an initial rate of 0.0001 and peaking at 0.001 after a warmup period of 10 epochs, to refine training dynamics and enhance convergence rates.

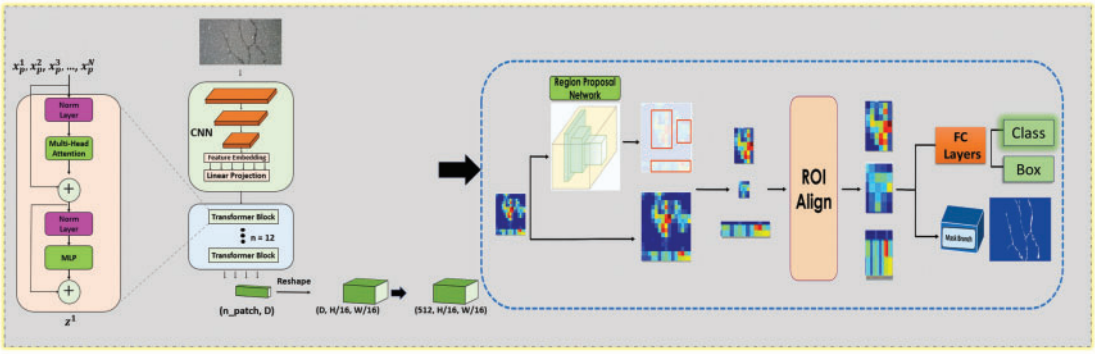

Figure 4: Overview of proposed hybrid pavement detection approach

Table 1: Configuration parameters of the MaskR-Transformer model for Crack500 and DeepCrack datasets

Parameters	Crack500	DeepCrack
Segmentation model	MaskR-Transformer	MaskR-Transformer
Objective function	DiceLoss + WBCE	DiceLoss
2D Sliding window inference batch size	4 with 0.25 overlap	4 with 0.5 overlap
Optimizer	AdamW	AdamW
Augmentation	True	True
Input size	(256, 256)	(224, 224)
# of feature map in base layer	16	8
# of attention head	12	12
MLP dimension	768	768
# of epochs	300	300
# of batch size	32	96
Initial learning rate	0.00001	0.0001
Warmup epoch	10	10
Learning rate scheduler	OneCycleLR	OneCycleLR
Max_lr	0.001	0.001

3.6 Evaluation of Model Performance

In this study, we assess the effectiveness of the proposed model for detecting pavement cracks through various established performance metrics. These metrics are essential for understanding how well the model performs in identifying cracks accurately on pavement surfaces. Among the most referenced metrics in scholarly articles are:

- **Accuracy measure:** This metric indicates the proportion of correct predictions the model makes out of the total predictions. It is widely used for classification models and is defined as:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \tag{4}$$

  Here, TP represents true positives (correctly identified positive cases), TN stands for true negatives (correctly identified negative cases), FP denotes false positives (incorrectly identified positive cases), and FN signifies false negatives (missed positive cases).
- **Precision measure:** Often utilized in pavement crack detection; this metric evaluates the accuracy with which the model identifies positive instances. Precision is the ratio of true positive cases to all cases classified as positive, combining both true positives and false positives. A high precision score suggests fewer misclassifications of non-crack instances as cracks, highlighting the model's accuracy in detecting actual cracks.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{5}$$

- **Recall measure:** This computes the ratio of total relevant pavement cases retrieved relative to the total number of non-pavement cases. This measures the proportion of actual positive cases correctly identified by the model, relative to the total number of positive cases.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{6}$$

- **F-Measure:** This combines precision and recall into a single metric, balancing both the model's accuracy and its ability to retrieve relevant cases.

$$\text{F} - \text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

- **Dice Similarity Coefficient (DSC):** This metric is particularly valuable for segmentation tasks, measuring the overlap between the predicted segmentation and the ground truth. It is defined as:

$$\text{DSC} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{8}$$

  The DSC is a normalized measure that ranges from 0 (no overlap) to 1 (perfect overlap), providing a clear indication of the model's performance in terms of spatial accuracy in the segmentation of cracks.
- **Mean Intersection over Union (mIoU)** is a common evaluation metric for image segmentation tasks. It measures the average overlap between the predicted segmentation mask and the ground truth across all classes. The formula for mIoU is:

$$\text{mIoU} = \frac{1}{C} \sum_{i=1}^{C} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i} \tag{9}$$

where C is the total number of classes (for binary segmentation, $C = 2$), TP_i is the number of true positives for class i (pixels correctly classified as class i), FP_i is the number of false positives for class i (pixels incorrectly classified as class i) and FN_i is the number of false negatives for class i (pixels that belong to class i but were missed). These evaluation metrics provide a comprehensive framework for analyzing the performance of the pavement crack detection model. This study compares the results with those of similar models discussed in the relevant literature. The objective is to ascertain the model's efficacy relative to previous methods and to identify potential enhancements for future models.

4 Results and Analysis

This section presents the comparative analysis of various deep learning models evaluated on two benchmark datasets, Crack500 and DeepCrack, to assess their performance in detecting pavement cracks. The models tested include U-Net, TransUNet, UNETr, Swin-UNETr, YoloV8, Mask R-CNN, nnUNet, and our proposed MaskerTransformer. Each model was evaluated based on several metrics: DSC, precision, recall, and F1-Score, to provide a comprehensive view of their capabilities in both segmentation and classification tasks.

4.1 Comparative Analysis on Models Performance

In this study, we have developed and examined seven state-of-the-art models and then evaluated their performance against our proposed model. The proposed MaskerTransformer model demonstrated superior performance on the Crack500 dataset with a DSC of 80.04%, outperforming other models such as U-Net (66.07%) and TransUNet (68.96%). This significant improvement underscores the effectiveness of MaskerTransformer in achieving precise segmentations in Table 2. In terms of precision, Mask R-CNN led with 70.411%, closely followed by our MaskerTransformer at 72.31%, indicating its strong capability in accurately identifying relevant pixels for crack detection.

Table 2: Comparative analysis evaluation measures

Model	Crack500					DeepCrack				
	DSC	Precision	Recall	F1-Score	mIoU	DSC	Precision	Recall	F1-Score	mIoU
U-Net	66.07	64.298	68.947	66.692	66.692	83.73	85.9	82.178	83.114	85.0
TransUNet	68.96	61.442	69.63	65.565	65.565	84.694	84.968	84.354	84.616	88.5
UNETr	75.39	69.378	78.657	73.88	73.88	89.919	90.017	87.79	88.522	90.0
Swin-UNETr	69.81	69.685	79.485	74.438	73.88	89.222	89.597	88.764	88.601	92.0
YoloV8	76.64	67.848	76.711	72.157	72.157	88.697	90.597	88.36	87.883	79.0
Mask R-CNN	76.44	70.411	**79.954**	75.59	75.59	90.032	**91.712**	89.345	89.654	88.0
nnUNet	74.22	67.917	79.145	72.767	72.767	87.898	91.033	86.09	87.246	90.5
MaskerTransformer	**80.04**	**72.31**	78.892	**76.698**	**75.851**	**91.37**	88.751	**90.36**	**90.726**	**93.5**

On the DeepCrack dataset, MaskerTransformer again showed outstanding performance with the highest DSC of 91.37%, highlighting its robustness across different datasets. It also achieved the highest F1-Score of 90.726%, suggesting an excellent balance between precision and recall. Notably, Mask R-CNN recorded the highest precision at 91.712%, but with a slightly lower recall, indicating a more conservative approach in crack detection (Fig. 5).

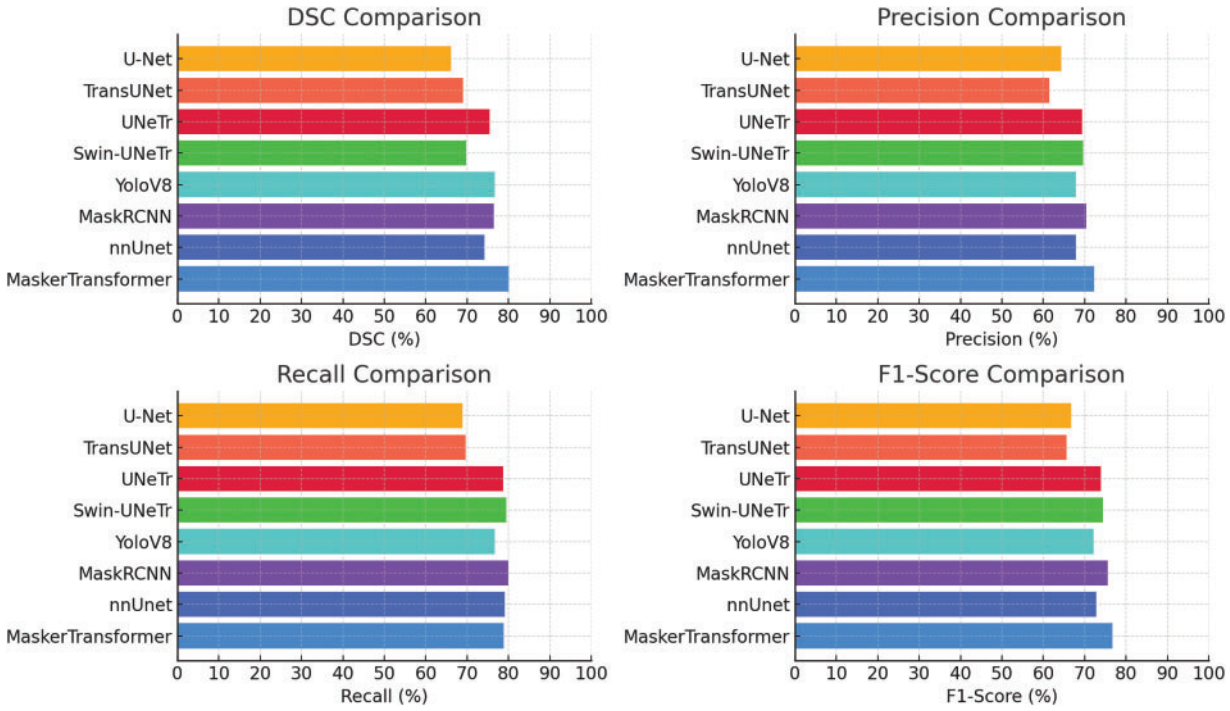

Figure 5: Comparative performance of various deep learning models on crack detection

The models UNETr and YoloV8 also performed commendably across both datasets. UNETr showed a particularly strong recall on the Crack500 dataset (78.657%), which is essential for minimizing false negatives in practical applications. YoloV8, while not leading in any specific metric, maintained consistent performance, underscoring its utility in diverse scenarios. Finally, while all models demonstrated varying strengths, the MaskerTransformer, our proposed hybrid model, consistently outperformed other models in key metrics across both datasets, establishing its potential as a highly effective tool for automated pavement crack detection. This comparative analysis not only validates the superiority of the MaskerTransformer in handling complex crack detection tasks but also highlights the importance of selecting appropriate architectures based on specific dataset characteristics and detection requirements.

4.2 Comparative Analysis on Models Complexity

Table 3 provides a comparative analysis of the models evaluated in this study based on their number of parameters, inference time, and FLOPs. These metrics are important for understanding the computational complexity and efficiency of each model for pavement segmentation. U-Net is the most lightweight model, with fast inference time (10 ms) and low computational demand (~16 GFLOPs). TransUNet and UNETr, which incorporate transformers, have higher computational complexity (~60 and ~150 ms inference time) but offer better segmentation performance for more intricate pavement

details due to their ability to capture global context, albeit at the cost of increased FLOPs (∼12 GFLOPs and ∼75 GFLOPs). For tasks that prioritize speed, YOLOv8N is the fastest with only ∼1–3 ms inference time and low computational requirements (∼10 GFLOPs), though it is more suited for object detection than fine segmentation. Mask R-CNN and nnUNet strike a balance between accuracy and computational complexity, making them suitable for detailed pavement damage analysis. Finally, Mask R-Transformer, while offering the highest level of segmentation detail, has significantly higher FLOPs (∼2635 GFLOPs) and longer inference times (∼65 ms), making it suitable for highly complex segmentation tasks where precision is critical, even at the cost of higher computational resources.

Table 3: Model comparison on number of parameters, inference time, and FLOPs

Model	Number of parameters	Inference time (256 × 256 Image)	FLOPs
U-Net	∼31 M	10 ms	∼16 GFLOPs
TransUNet	∼38 M	∼60 ms	∼12 GFLOPs
UNETr	104 M	∼150 ms	∼75 GFLOPs
Swin-UNETr	∼138 M	∼200 ms	∼90 GFLOPs
YOLOv8N	∼3.5 M	∼1–3 ms	∼10 GFLOPs
Mask R-CNN	∼44 M (ResNet50 Backbone)	∼50 ms	∼937 GFLOPs
nnUNet	∼16 M	∼25 ms	∼650 GFLOPs
Mask R-Transformer	∼91 M	∼65 ms	∼2635 GFLOPs

4.3 Comparative Analysis of Model Performance on Pavement Crack Detection

Fig. 6 illustrates the comparative performance of various deep-learning models on the Crack500 dataset for pavement crack detection. The first column shows raw images of the pavement surfaces, followed by the ground truth annotations in the second column, which highlight the actual cracks. Subsequent columns display the outputs from different models: U-Net, UNETr, nnUNet, YoloV8, and our model (MaskerTransformer). The effectiveness of each model is visualized through their ability to detect and accurately outline the cracks as compared to the ground truth. Notably, our model (rightmost column) demonstrates enhanced precision in delineating the cracks, closely aligning with the ground truth, especially in scenarios with complex crack patterns (highlighted in red boxes).

Fig. 7 presents a similar setup for the DeepCrack dataset, showcasing the ability of the same array of models to identify and segment pavement cracks. Each row corresponds to a different example from the dataset, starting with the raw image, followed by the ground truth, and then the model outputs. This figure underscores the varying levels of success across models in handling different types of cracks, such as thin, discontinuous, or branching cracks. Again, the performance of our model is featured in the final column, showing its superior capability to accurately trace the complex and subtle features of the cracks, which is crucial for effective pavement maintenance.

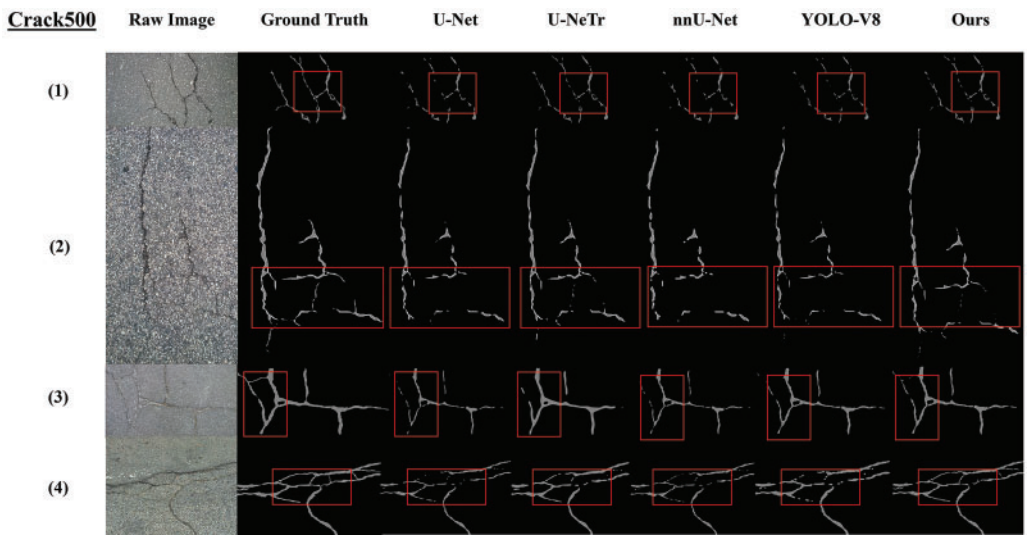

Figure 6: Visualization of crack detection results on the Crack500 dataset across different deep-learning models

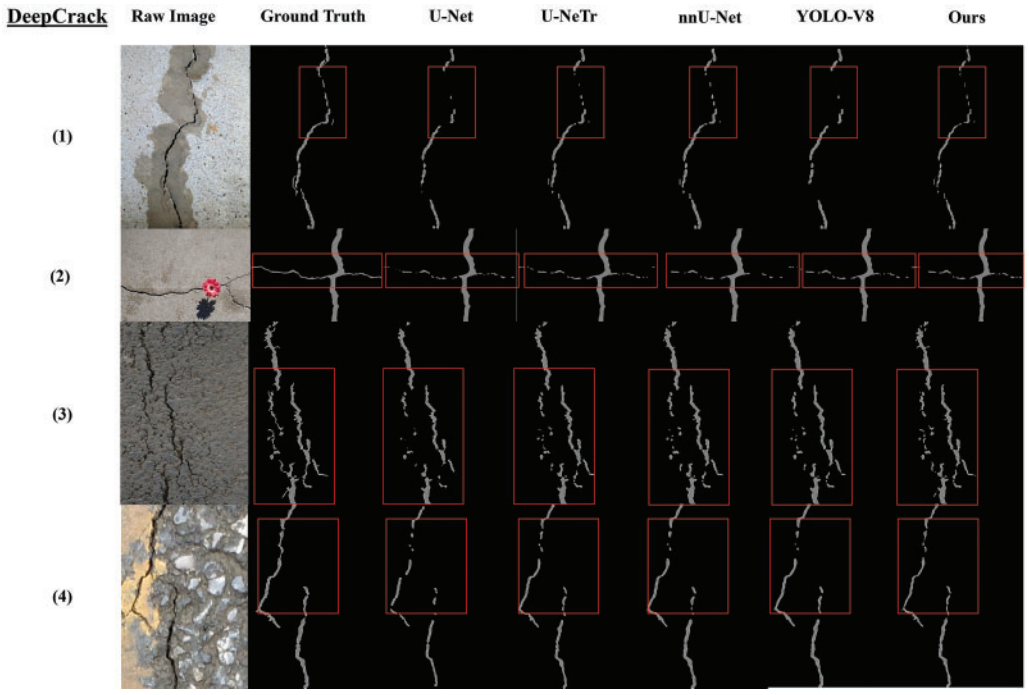

Figure 7: Comparative visualization of crack detection on the DeepCrack dataset

4.4 Benchmarking our Proposed Model Performance against the State-of-the-Art Models

Fig. 8 presents a violin plot displaying the dice similarity coefficient for various deep-learning models evaluated on the Crack500 dataset. The DSC metric quantifies the similarity between the predicted segmentation and the ground truth, offering insights into the accuracy of each model's

segmentation capabilities. From the plot, it is evident that the U-Net, TransUNet, and UNETr models show wider distributions of DSC scores, indicating variability in performance across different test cases. On the other hand, the MaskerTransformer and UNETr models demonstrate higher median DSC values with tighter distributions, suggesting more consistent and accurate crack detection. The Swin-UNETr and YoloV8 models, while achieving commendable median scores, also exhibit wider distributions, highlighting variability in their performance. Notably, our MaskerTransformer model not only shows a high median DSC but also maintains a relatively compact distribution, underscoring its robustness and reliability in detecting pavement cracks.

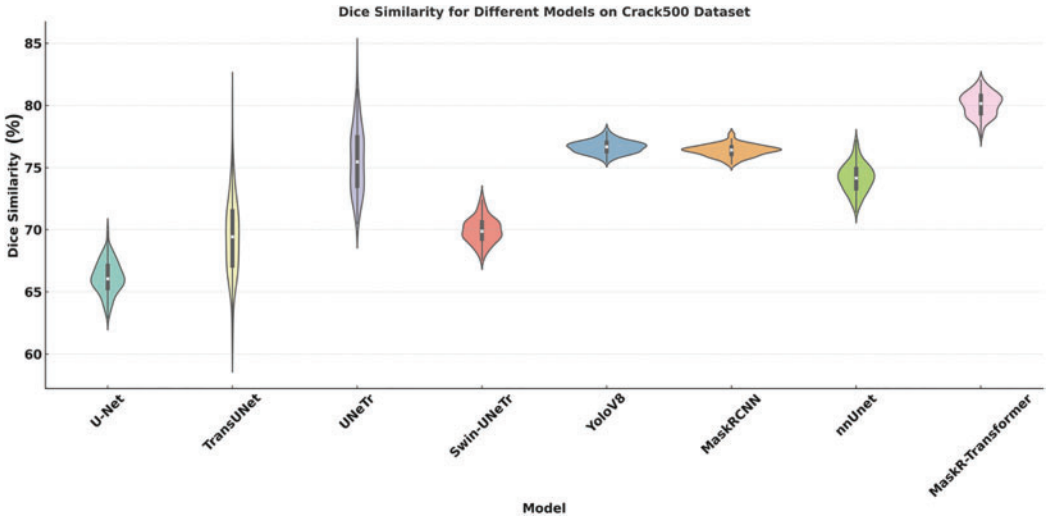

Figure 8: Dice similarity coefficient across models on Crack500 dataset

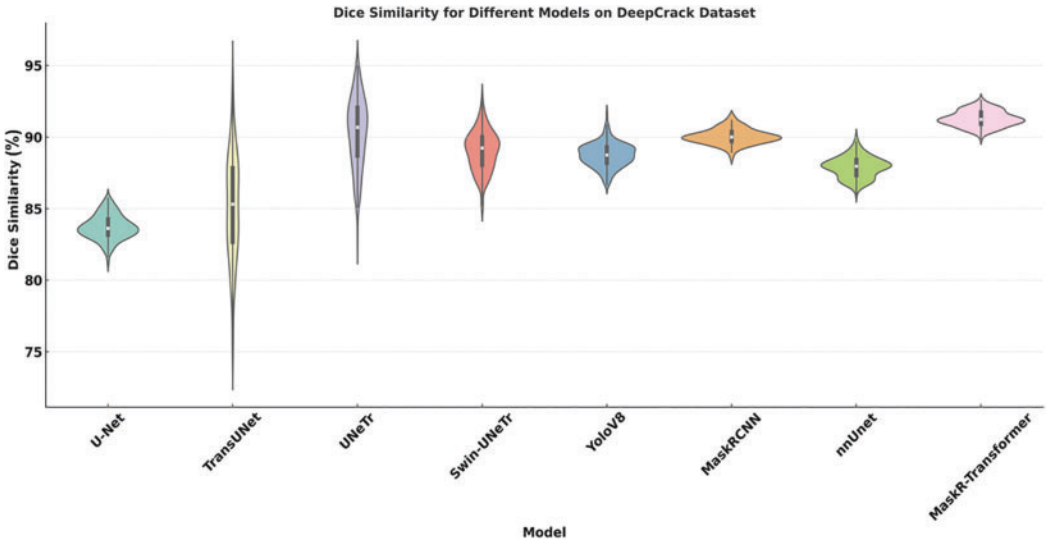

Figure 9: Dice similarity coefficient scores for various models on the DeepCrack dataset

Fig. 9 showcases the DSC scores for various deep-learning models on the DeepCrack dataset. Similar to the results on Crack500, the DSC values provide a measure of each model's effectiveness

in pavement crack segmentation. On the DeepCrack dataset, all models generally exhibit higher DSC scores compared to the Crack500 dataset, indicating better overall performance. The Mask R-CNN and MaskerTransformer models show impressive peaks in their DSC distributions, which are indicative of high-quality segmentation outputs. The consistency in the performance of the Masker-Transformer model across both datasets highlights its robust algorithmic structure that effectively captures complex crack patterns, making it a reliable choice for practical applications in pavement maintenance.

5 Conclusion

This study validates the efficacy of the MaskerTransformer, a novel hybrid model combining Mask R-CNN and ViT, in improving pavement crack detection across diverse conditions. Through comprehensive testing on the Crack500 and DeepCrack datasets, our findings underscore the superior performance of MaskerTransformer over other contemporary models. With a dice similarity coefficient of 80.04% on Crack500 and 91.37% on DeepCrack, alongside the highest precision, recall, and F1 scores, the model has proven its ability to accurately segment and detect diverse crack types under varying conditions, thus validating the hypothesis that integrating local segmentation with global contextual awareness significantly enhances detection accuracy. Despite its strong performance, this study is not without limitations. The high dependency on large, annotated datasets for training may limit the model's applicability in scenarios where such data is scarce or of poor quality. Additionally, the computational demands of processing through both Mask R-CNN and ViT architectures require substantial computational resources, which could be a barrier in low-resource settings. Furthermore, the model's performance in extremely varied weather conditions or on non-standard pavement materials remains less explored, which could affect its generalizability.

Future research will focus on several key areas to further enhance the MaskerTransformer's utility and adaptability. First, efforts will be directed towards optimizing the model's architecture to reduce computational overhead without compromising performance, making it more feasible for real-time applications and deployable on edge devices. Additionally, we plan to expand the training datasets to include a broader range of crack types and pavement conditions, particularly from underrepresented regions and varying weather scenarios, to improve the model's robustness and generalizability. Moreover, integrating additional modalities such as radar and infrared imaging could potentially improve the detection capabilities under adverse conditions, such as poor lighting or wet surfaces. We also aim to explore the use of unsupervised or semi-supervised learning techniques to reduce the dependency on large, annotated datasets, thereby addressing one of the critical limitations identified in this study. Finally, developing lightweight versions of the vision transformer and Mask R-CNN architectures will be explored to reduce computational complexity, making the model more suitable for real-time and low-resource environments.

Acknowledgement: Thanks to the reviewers for their valuable efforts and insightful comments to improve this paper quality and clarity.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Study conception and design: Li Wu, Daojun Dong; data collection: Yao Cheng; analysis and interpretation of results: Shorouq Alshawabkeh, Liping Li; draft manuscript preparation: Shorouq Alshawabkeh. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available. The Crack500 dataset can be accessed from Kaggle (https://www.kaggle.com/datasets/pauldavid22/crack500202205091090436z001) and the DeepCrack dataset is available on GitHub (https://github.com/yhlleo/DeepCrack/tree/master/dataset). Both datasets were accessed on 27 May 2024.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] A. Di Graziano, V. Marchetta, and S. Cafiso, "Structural health monitoring of asphalt pavements using smart sensor networks: A comprehensive review," *J. Traffic Transp. Eng.*, vol. 7, no. 5, pp. 639–651, Oct. 2020.

[2] A. Shtayat, S. Moridpour, B. Best, A. Shroff, and D. Raol, "A review of monitoring systems of pavement condition in paved and unpaved roads," *J. Traffic Transp. Eng.*, vol. 7, no. 5, pp. 629–638, Oct. 2020.

[3] F. Li *et al.*, "Urban ecological infrastructure: An integrated network for ecosystem services and sustainable urban systems," *J. Clean. Prod.*, vol. 163, pp. S12–S18, Oct. 2017. doi: 10.1016/j.jclepro.2016.02.079.

[4] K. Gopalakrishnan, "Deep learning in data-driven pavement image analysis and automated distress detection: A review," *Data*, vol. 3, no. 3, 2018, Art. no. 28. doi: 10.3390/data3030028.

[5] N. Kheradmandi and V. Mehranfar, "A critical review and comparative study on image segmentation-based techniques for pavement crack detection," *Constr. Build. Mater.*, vol. 321, no. 7, Feb. 2022, Art. no. 126162. doi: 10.1016/j.conbuildmat.2021.126162.

[6] D. Ai, G. Jiang, S. K. Lam, P. He, and C. Li, "Computer vision framework for crack detection of civil infrastructure—A review," *Eng. Appl. Artif. Intell.*, vol. 117, no. 4, Jan. 2023, Art. no. 105478. doi: 10.1016/j.engappai.2022.105478.

[7] S. Alshawabkeh, L. Wu, D. Dong, Y. Cheng, L. Li and M. Alanaqreh, "Automated pavement crack detection using deep feature selection and whale optimization algorithm," *Comput. Mater. Contin.*, vol. 77, no. 1, pp. 63–77, 2023. doi: 10.32604/cmc.2023.042183.

[8] F. Liu, W. Ding, Y. Qiao, and L. Wang, "Transfer learning-based encoder-decoder model with visual explanations for infrastructure crack segmentation: New open database and comprehensive evaluation," *Undergr. Space*, vol. 17, no. 10, pp. 60–81, Aug. 2024. doi: 10.1016/j.undsp.2023.09.012.

[9] F. Liu and L. Wang, "UNet-based model for crack detection integrating visual explanations," *Constr. Build. Mater.*, vol. 322, no. 1, 2022, Art. no. 126265. doi: 10.1016/j.conbuildmat.2021.126265.

[10] W. Cao, Q. Liu, and Z. He, "Review of pavement defect detection methods," *IEEE Access*, vol. 8, pp. 14531–14544, Jan. 2020. doi: 10.1109/ACCESS.2020.2966881.

[11] Z. Chen, K. Pawar, M. Ekanayake, C. Pain, S. Zhong and G. F. Egan, "Deep learning for image enhancement and correction in magnetic resonance imaging—State-of-the-art and challenges," *J. Digit. Imaging*, vol. 36, no. 1, pp. 204–230, Feb. 2023. doi: 10.1007/s10278-022-00721-9.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," presented at the IEEE Int. Conf. Comput. Vis., 2017, pp. 2961–2969.

[13] A. Vaswani *et al.*, "Attention is all you need," in *31st Conf. Neural Inf. Process. Syst. (NIPS 2017)*, Long Beach, CA, USA, 2017, vol. 30.

[14] Y. Tang, A. A. Zhang, L. Luo, G. Wang, and E. Yang, "Pixel-level pavement crack segmentation with encoder-decoder network," *Measurement*, vol. 184, no. 1, 2021, Art. no. 109914. doi: 10.1016/j.measurement.2021.109914.

[15] R. Augustauskas and A. Lipnickas, "Improved pixel-level pavement-defect segmentation using a deep autoencoder," *Sensors*, vol. 20, no. 9, 2020, Art. no. 2557. doi: 10.3390/s20092557.

[16] A. Zhang *et al.*, "Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network," *Comput.-Aided Civ. Infrastruct. Eng.*, vol. 32, no. 10, pp. 805–819, 2017.

[17] X. Xu *et al.*, "Crack detection and comparison study based on Faster R-CNN and Mask R-CNN," *Sensors*, vol. 22, no. 3, Feb. 2022, Art. no. 1215. doi: 10.3390/s22031215.
[18] P. Wang *et al.*, "Research on automatic pavement crack recognition based on the Mask R-CNN model," *Coatings*, vol. 13, no. 2, Feb. 2023, Art. no. 430. doi: 10.3390/coatings13020430.
[19] E. N. Ukhwah, E. M. Yuniarno, and Y. K. Suprapto, "Asphalt pavement pothole detection using deep learning method based on YOLO neural network," presented at the 2019 Int. Semin. Intell. Technol. Appl. (ISITIA), Aug. 2019, pp. 35–40.
[20] R. Ghosh and O. Smadi, "Automated detection and classification of pavement distresses using 3D pavement surface images and deep learning," *Transp. Res. Rec.*, vol. 2675, no. 9, pp. 1359–1374, Sep. 2021. doi: 10.1177/03611981211007481.
[21] J. Liu *et al.*, "Automated pavement crack detection and segmentation based on two-step convolutional neural network," *Comput.-Aided Civ. Infrastruct. Eng.*, vol. 35, no. 11, pp. 1291–1305, Nov. 2020.
[22] F. Liu, J. Liu, and L. Wang, "Deep learning and infrared thermography for asphalt pavement crack severity classification," *Autom. Constr.*, vol. 140, no. 7553, Aug. 2022, Art. no. 104383. doi: 10.1016/j.autcon.2022.104383.
[23] PaulDavid22, "Crack500," Kaggle, May 27, 2024. [Online]. Available: https://www.kaggle.com/datasets/pauldavid22/crack50020220509t090436z001
[24] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation," *Neurocomputing*, vol. 338, no. 12, pp. 139–153, Apr. 2019. doi: 10.1016/j.neucom.2019.01.036.
[25] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, no. 5, Dec. 2020, Art. no. 105524. doi: 10.1016/j.asoc.2019.105524.
[26] Y. Li, K. Guo, Y. Lu, and L. Liu, "Cropping and attention based approach for masked face recognition," *Appl. Intell.*, vol. 51, pp. 3012–3025, May 2021. doi: 10.1007/s10489-020-02100-9.
[27] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 13001–13008, Apr. 2020.
[28] J. Zhou, Y. Zheng, J. Tang, J. Li, and Z. Yang, "FlipDA: Effective and robust data augmentation for few-shot learning," Aug. 2021, *arXiv:2108.06332*.
[29] P. Bharati and A. Pramanik, "Deep learning techniques—R-CNN to Mask R-CNN: A survey," in *Computational Intelligence in Pattern Recognition*, Singapore: Springer, 2020, pp. 657–668.
[30] J. H. Shu, F. D. Nian, M. H. Yu, and X. Li, "An improved Mask R-CNN model for multiorgan segmentation," *Math. Probl. Eng.*, vol. 2020, no. 1, 2020, Art. no. 8351725.
[31] M. A. Arshed, A. Alwadain, R. F. Ali, S. Mumtaz, M. Ibrahim and A. Muneer, "Unmasking deception: Empowering deepfake detection with vision transformer network," *Mathematics*, vol. 11, no. 17, Aug. 2023, Art. no. 3710.