Data Engineering and Applications

Research Paper

Pedestrian Pavement Crack Detection Method Based on DeepLabv3-DRN Fusion

Paweł Jakub Kowalski[1], Michał Bartosz Szymański[2], Aleksander Wójcik[2] and Tomasz Wiśniewski[1, *]

[1] Faculty of Electrical Engineering, Automatic Control and Computer Science, Kielce University of Technology, 25-314 Kielce, Poland
[2] Faculty of Computer Science, Polish-Japanese Academy of Information Technology, 02-008 Warsaw, Poland
*Corresponding author: tomasz.w@tu.kielce.pl

Abstract. Due to changes in materials and complex environmental factors, previous inspection methods are no longer reliable, and semantic segmentation algorithms have become useful tools for solving the problem of sidewalk crack detection. This paper introduces a new crack segmentation framework that combines DeepLabv3 and the Dilated Residual Network (DRN), using a structured multi-level feature fusion strategy. By integrating local details and broad contextual information through parallel backbone extraction and channel attention modules, various shapes, widths, and visibility of cracks can be effectively identified. In the validation experiments of public and private datasets, over 1,600 labeled pavement images were collected from various cities. The average Intersection over Union (mIoU) of the fusion model is 0.81, which is 3-5% higher than U-Net and DeepLabv3+. On a typical GPU, the inference speed reaches up to 41 frames per second, with minimal memory usage, and the average F1-score exceeds 0.79 across all datasets. Ablation studies show that both feature fusion and enhancement modules are necessary. If they are excluded, the mIoU will drop to 0.72. A detailed examination of the errors indicates that the model reduces false positives in cases of occlusion and background clutter. The above results indicate that the DeepLabv3-DRN fusion framework can improve the accuracy and generalization ability of automatic crack detection in pedestrian pavement environments.

Keywords: *Image Analysis, Semantic Segmentation, Deep Learning, Crack Detection, Pavement Infrastructure, Feature Fusion, Model Robustness, Real-Time Processing*

Received on 24 September 2025, Accepted on 26 December 2025, Published on 2 Jan2026

Introduction

Urban pedestrian infrastructure is one of the essential infrastructures needed to provide safe travel and living for urban residents [1]. With the rapid advancement of urbanization, the demand for the construction and maintenance of sidewalks is gradually increasing [2]. Pavement cracks pose safety hazards for tripping and water damage, while also reducing load-bearing capacity [3]. Therefore, timely detection and repair of pavement cracks help extend the lifespan of the road, ensure uninterrupted use, and reduce daily maintenance costs [4]. Although manual inspections are still very common, they are labor-intensive and prone to errors [5]. In large, complex urban areas, common reasons for the inability to report defects are fundamental deficiencies in people's abilities and observational errors [6]. Moreover, under the influence of factors such as the presence of many people and vehicles, changes in light and weather, regular inspections may become more inaccurate. [7] With the development of cities, the demand for the urgency and reliability of pavement inspection technology has increased in recent years [8].

Computer vision and deep learning have recently achieved automated end-to-end inference for image-based crack detection, eliminating the need for manual feature engineering [9]. Convolutional neural networks, particularly well-suited for semantic segmentation, have been able to successfully identify slender and irregular crack patterns in noisy or complex environments [10]. Representative semantic segmentation models, such as

U-Net [11], DeepLabv3 [12], and Dilated Residual Network (DRN) [13], excel in learning multi-scale spatial features and integrating contextual information [14], and are now widely used. These deep architectures can also handle high-resolution urban pavement maps well [15]. However, general segmentation models still have the potential to miss small or blurry cracks due to the dilution of deep features [16]. The main reasons for establishing a reliable and scalable detection system remain background noise and the diversity of crack shapes, sizes, and pavement surface textures [17]. Therefore, more application-oriented high-precision technologies are still needed, technologies that can meet the specific needs of pedestrian pavement inspection.

This paper proposes a novel crack detection method by combining DeepLabv3 and an Atrous Residual Network within a single structural fusion system to address the aforementioned issues. To improve the robustness and accuracy of crack segmentation, the contextual modeling capabilities of DeepLabv3 and the edge-preserving feature extraction of DRN represent the first technical advancements. Extended experiments show that the proposed fusion model outperforms the previous best models in both accuracy and recall, and is relatively easy to implement. These findings are based on numerous real-world paving datasets. The structure of this paper is as follows: In Section 2, pavement crack detection and deep learning-based segmentation are discussed. Section 3 introduces the innovations and structure of the proposed fusion method. Section 4 introduces the experimental process, results, and analysis. Finally, the fifth section includes the conclusions of this paper and directions for future research.

Related Work

Pavement Crack Detection

With the increasing demand for precise and scalable assessment of infrastructure surface conditions, automated pavement crack detection technology has begun to emerge and be applied [18]. In the initial research, traditional image processing techniques were mainly used to collect edge information and extract data from various damaged asphalt pavements through morphological filtering and adaptive thresholding [19]. These algorithms perform poorly on clear, high-contrast images, but they do not perform well when there is insufficient lighting or when surface damage and other obstacles are present [20]. To better describe crack shapes, researchers have recently added new feature descriptors such as wavelets, histograms, and local texture statistics to improve model stability [21]. Despite these efforts, handcrafted features generally perform poorly and are unable to distinguish cracks from pavement markings, stains, or cast shadows in real urban environments [22]. Classic machine learning algorithms such as support vector machines and decision trees can use manually labeled datasets to improve efficiency, but due to a lack of feature richness, they perform poorly in terms of generalization [23]. Manual verification and post-processing are still necessary, thus reducing the objectivity of deployment and scaling [24]. Therefore, there is an urgent need for fast, data-driven solutions to help urban managers more effectively identify cracks and surface issues in large transportation networks [25].

Deep Learning Semantic Segmentation

The emergence of Convolutional Neural Networks (CNN) and deep learning has made end-to-end direct learning possible [26]. Data-driven representation learning has improved the accuracy and generalization ability of deep neural networks under various road conditions, replacing manual feature engineering. By using encoder-decoder structures with skip connections, U-Net and other frameworks have already addressed this issue. These connections can better segment fine, irregular crack features and retain spatial information and semantic context [27]. With the update of DeepLabv3, it has added dilated convolution and Atrous Spatial Pyramid Pooling (ASPP) modules. These modules help perform multi-scale feature extraction and improve variations in crack direction and crack width [28]. The Dilated Residual Network (DRN) also improves the collection of long-range and local features because it increases the receptive field of residual learning without reducing spatial resolution [29]. These models have improved the performance standards of public datasets and field image collection, but they also have drawbacks. Class imbalance, false positives caused by background noise, and visual differences in cracks. The research on segmentation architectures and learning strategies still focuses on how to adapt to new environments, how to reduce labeled data, and how to accurately identify brittle cracks [30].

Model Fusion Approaches

Due to the performance decline of single models, recent network fusion techniques have been used for crack detection tasks to integrate features from multiple sources [31]. Model fusion, also known as combining different architectures, involves methods such as early integration of input modalities, intermediate feature connections, or late ensemble averaging to simultaneously leverage their advantages [32]. For example, a dual-stream network can independently process color and texture features, and then integrate them during the segmentation stage to enhance the ability to recognize micro-cracks and complex backgrounds [33]. Intermediate fusion can integrate features from different layers through fusion blocks or channel attention to enhance feature alignment and cross-layer information exchange [34]. Pavement crack detection employs a multi-branch and multi-task fusion design. This design allows for the simultaneous distinction between cracks and other anomalies, or the integration of prior information from specific domains into the learning process [35]. Attention-based fusion modules and adaptive weighting can enhance the recovery of subtle boundaries and reduce noise in weakly or highly occluded cracks [36]. Although these have been achieved, there is still a need to study a balanced and computationally efficient fusion scheme. Feature space misalignment, higher resource demands, and the challenges of scaling on large, diverse datasets all require further research. Therefore, model fusion remains an attractive yet challenging research method for developing practical pavement crack detection systems [37].

Proposed Method

DeepLabv3 and DRN Model Structures

In order to obtain a stable and fully automated crack segmentation method in complex pavement environments, DeepLabv3 and Extend Fracture Network (DRN) were used. Using Atrous Spatial Pyramid Pooling (ASPP) and multiple stacked atrous convolutions, DeepLabv3 is a multi-scale feature extraction method. Since the architecture module is designed to collect contextual information at different scales, it is suitable for handling cracks of various shapes and widths in segmentation tasks.

It can be formally expressed as: The output feature map of the ASPP module is a combination of multi-rate dilated convolutions:

$$ASPP(x) = \sum_{d \in D} Conv_{3\times3}(x; d) \quad \text{Eq.(1)}$$

where x is the input tensor, d enumerates distinct dilation rates in the set D, and $\mathrm{Conv}_{3\times3}$ denotes 3×3 convolution at dilation d. It can cover local and global crack characteristics efficiently and has a low computational cost.

The core atrous/dilated convolution in both DeepLabv3 and DRN is mathematically defined as follows:

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k] \cdot w[k] \quad \text{Eq.(2)}$$

where y is the output at position i, x is the input, $w[k]$ represents convolution weights, r is the dilation rate, and K is the filter length. By changing the dilation factor, the network's perception of the field is linearly increased. This enables the network to learn large-scale structures and fine-grained details in road surface images.

To recover fine structures, the Dilated Residual Network (DRN) adds residual learning and dilated convolutions. The residual block in DRN is as follows:

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}, \{W_i\}) + \boldsymbol{x} \quad \text{Eq.(3)}$$

where $\mathcal{F}(\mathbf{x}, \{W_i\})$ is a residual function-typically composed of a sequence of convolutions, nonlinearities, and batch normalization-with learnable parameters $\{W_i\}$. Additive skip connections can be used to enhance gradient flow and preserve boundary information of deep feature propagation.

As shown in Figure 1, the system process includes image preprocessing, parallel feature extraction through DeepLabv3 and DRN, intermediate feature fusion, and crack mask segmentation.

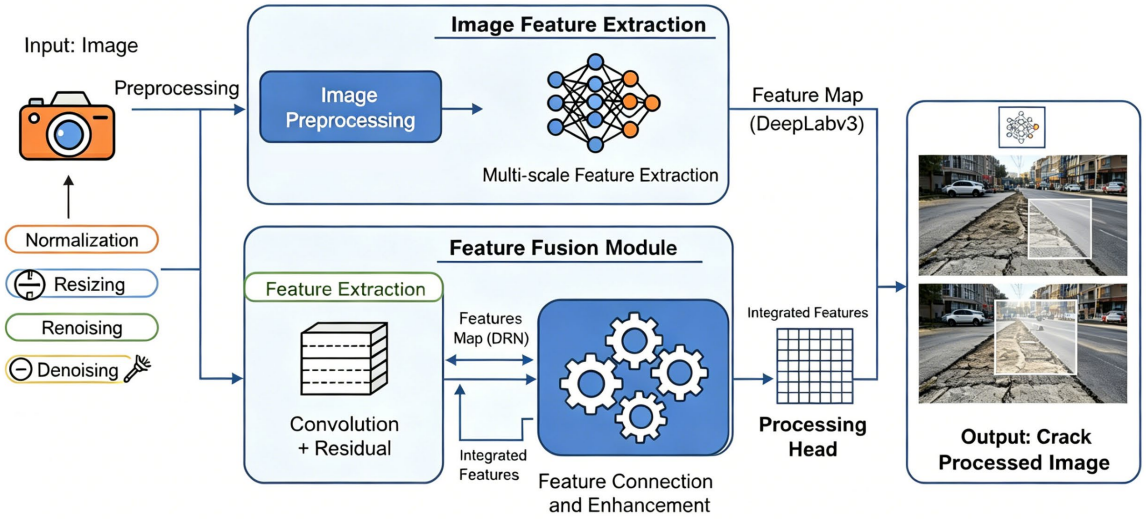

Figure 1. Overall System Workflow Diagram

Due to its ability to aggregate contextual information and utilize DRN to retain high-resolution features, DeepLabv3 addresses the issues of small or irregular cracks during the segmentation process through a combination system, while also reducing false positives caused by surface noise or occlusion. The above content lays the foundation for the subsequent advanced fusion architecture.

Fusion Architecture and Innovations

The combination of multi-level features is crucial for achieving high-precision and stable segmentation of pavement cracks with various irregular shapes under different conditions. In this paper, we designed a new fusion architecture to combine the various representations generated by the DeepLabv3 and DRN branches. By using a dual-stream design, the contextual awareness of DeepLabv3 and the precise spatial information of DRN are integrated.

At the feature map level, the architecture begins the fusion process. First, the high-dimensional representations of each backbone network are aligned and normalized. Let F_{DL} and F_{DRN} denote the extracted feature maps from DeepLabv3 and DRN, respectively. To facilitate effective information blending, we employ a channel alignment module, which projects both F_{DL} and F_{DRN} into a shared feature space using 1×1 convolutions:

$$\hat{F}_{DL} = Conv_{1\times1}(F_{DL}), \hat{F}_{DRN} = Conv_{1\times1}(F_{DRN}) \quad \text{Eq.(4)}$$

In order to ensure content consistency and avoid information loss, the unified fused representation F_{fused} is first aligned:

$$F_{fused} = Concat(\hat{F}_{DL}, \hat{F}_{DRN}) \quad \text{Eq.(5)}$$

By using the Squeeze-and-Excitation (SE) style channel attention mechanism, feature redundancy is reduced and the impact of relevant crack cues is enhanced. In the recalibration stage, different weights are adaptively assigned to each channel through global average pooling and two fully connected layers (followed by a sigmoid function). The refined and fused features are as follows:

$$F_{out} = \sigma\left(FC_2\left(\delta\left(FC_1\left(GAP(F_{fused})\right)\right)\right)\right) \odot F_{fused} \quad \text{Eq.(6)}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, FC_1 and FC_2 are fully connected layers, δ is the ReLU activation, σ is the sigmoid function, and $\odot$ denotes channel-wise multiplication. The selective attention mechanism ignores background noise but retains strong crack-related activations.

After the attention mechanism is refined, the fused features are gradually sampled and decoded in the lightweight segmentation head. In order to improve the boundary localization and continuity of thin or fragmented cracks, this decoder has skip connections from the previous fusion stage. During the training process, we use an auxiliary deep supervision strategy to add intermediate segmentation branches. This method provides

the network with multi-scale feedback to improve learning stability and enhance the ability to distinguish between fuzzy crack regions.

To address the segmentation discontinuities of occluded or low-contrast surfaces, we developed a cascade refinement module. This module iteratively refines the spatial details of the segmentation mask using dilated convolutions and residual connections, which are then used to decode the feature map. Each refined set is as follows:

$$Y_{n+1} = f_{dilated}(Y_n) + Y_n \quad \text{Eq.(7)}$$

where Y_n denotes the feature map at the nth stage and f_{dilated} is a dilated convolutional operation. Recursively aggregate features of different scales to obtain relatively smooth yet precise crack boundaries.

The complete structure of the proposed fusion network is shown in Figure 2. As shown in the diagram, attention-based fusion, decoding and refinement, dual-stream feature extraction, channel alignment, and other processes ultimately lead to the output prediction.

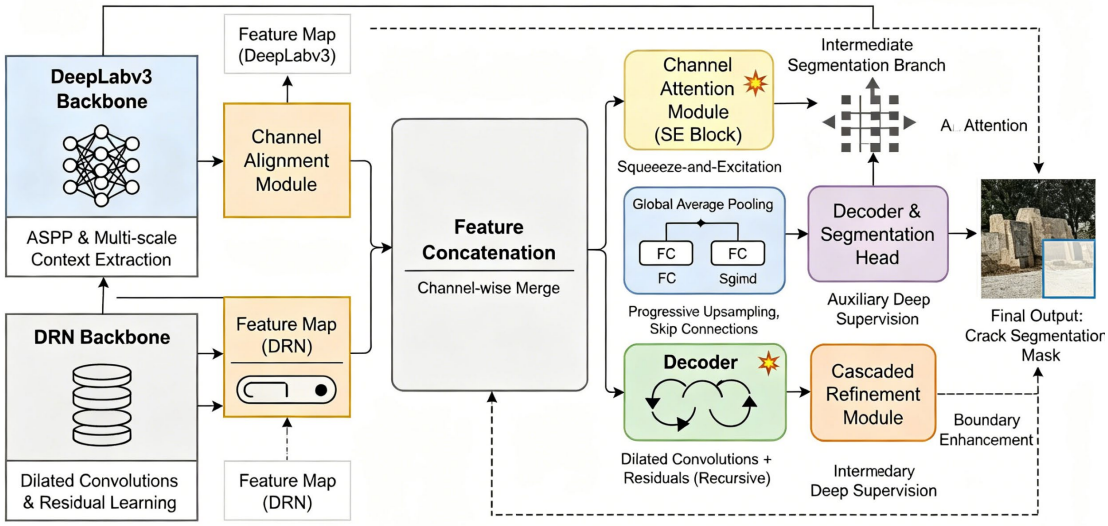

Figure 2. Detailed Fusion Network Architecture

As far as we know, the prior art of the fusion pipeline has included three engineering innovations: channel-coordinated dual-encoder fusion, adaptive attention crack localization, and multi-stage spatial refinement. Overall, the aforementioned improvements can help address issues related to various crack widths, rich textures in surrounding areas, and partial occlusion of pavement images in the system. The segmentation model has a lower rate of artifacts and false positives, and it is more sensitive to the subtle features of cracks. Experimental results show that the aforementioned structure has excellent capabilities in pavement crack detection.

Training Configurations

Model training is pivotal for enabling the proposed dual-stream crack segmentation architecture to achieve both generalization and robust adaptation to the heterogeneity present in real-world pavement imagery. The training regimen is meticulously designed to ensure rapid convergence, high structural fidelity, and reliable discrimination of subtle crack patterns.

This study collected a well-organized set of pavement images from various urban areas, captured by different devices, containing various crack widths, patterns, and scenes. The images were randomly divided into a training set (70%), a validation set (15%), and a test set (15%), with no overlap. To ensure consistent spatial dimensions within batches, all original input images are resized and center-cropped to a size of 512×512 pixels. To ensure stable optimization, the channel intensities are normalized to have a mean of zero and a standard deviation of one for each image.

Extend data augmentation further to reduce overfitting and enhance model stability. This includes random horizontal and vertical flipping ($p = 0.5$), rotation in $[-20°, +20°]$, cropping with scale ratio from 0.8 to 1.0,

random brightness/contrast adaptation, and stochastic Gaussian noise injection. The effective input used for model fitting, $\tilde{X}_i$, is given by

$$\tilde{X}_i = T(X_i) \quad \text{Eq.(8)}$$

where $T(\cdot)$ is the applied stochastic augmentation pipeline to the original image X_i. Therefore, the model cannot treat most real-world problems as background noise.

The network optimized a composite loss function that combines global structure preservation and pixel-level precise learning:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{BCE} + \lambda_2 \mathcal{L}_{Dice} \quad \text{Eq.(9)}$$

For binary cross-entropy on segmentation masks:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{j=1}^{N} \left[y_j log\left(\hat{y}_j\right) + \left(1 - y_j\right) log\left(1 - \hat{y}_j\right) \right] \quad \text{Eq.(10)}$$

where N is batch pixel size, y_j and $\hat{y}_j$ are the true and predicted labels at pixel j. Dice Loss is used to improve the region consistency:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{j=1}^{N} y_j \hat{y}_j + \epsilon}{\sum_{j=1}^{N} y_j + \sum_{j=1}^{N} \hat{y}_j + \epsilon} \quad \text{Eq.(11)}$$

with a small constant $\epsilon = 10^{-7}$ for stability and weights $\lambda_1, \lambda_2 = 0.5$.

Auxiliary segmentation heads are connected to each decoder stage for additional supervision, and a global training loss is obtained.

$$\mathcal{L}_{global} = \mathcal{L}_{total} + \beta \sum_{k} \mathcal{L}_{aux}^{(k)} \quad \text{Eq.(12)}$$

where β is a balancing hyperparameter, and the sum runs over decoder stages.

Kaiming He is used to initialise the model parameters for effective signal spread. Optimization uses Adam with learning rate $1 \times 10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$, subject to cosine annealing with warm restarts:

$$\eta_t = \eta_{min} + 0.5(\eta_{max} - \eta_{min}) \left[1 + cos\left(\frac{T_{cur}}{T_{max}} \pi \right) \right] \quad \text{Eq.(13)}$$

Here, η_t is the learning rate at epoch t. NVIDIA A100, 40 GB, 120 training epochs, batch size of 8.

Each layer of the network is regularized through batch normalization and dropout (p=0.2). To prevent overfitting, early stopping is used in the first 15 validation epochs.

The mean Intersection over Union (mIoU) of the segmentation is the main metric for quantitative evaluation:

$$IoU = \frac{TP}{TP + FP + FN} \quad \text{Eq.(14)}$$

Among them, TP, FP, and FN are the counts of true positive, false positive, and false negative pixels. When mIoU reaches its peak, save the model checkpoint.

By rigorously processing the data and creating a new optimization pipeline, this arrangement will achieve good discriminative and generalizable feature learning. This will provide a scientific basis for the analysis and comparison in the subsequent sections.

Experimental Validation

Datasets, Annotation, and Implementation Details

This study will use private and public datasets for model testing. In the Crack500 dataset, there are 500 RGB pavement images with different crack patterns and sizes. On the other hand, the CFD dataset contains 118 grayscale pavement macro images with detailed unidirectional and multidirectional crack labels. In summary, these public benchmarks cover various types of cracks and support objective generalization and cross-dataset comparison.

Collected a portion of proprietary datasets from Dataset A (320 images), Dataset B (450 images), and Dataset C (370 images). Collected nine major cities, located in North America and East Asia. The data collection subjects should differ in background, age, and device type, have diverse lighting conditions (over 30% in darkness or deep shadow), have occluded surfaces (at least 20% severely occluded), and include sidewalks.

Annotation is completed in two steps. First, an edge-based algorithm automatically generates candidate masks. Then, experts manually refine these masks to accurately distinguish between noise and cracks. Double-blind cross-checking ensured the consistency of the annotations. The average Dice similarity coefficient between annotators exceeded 0.96, indicating that it provided high-quality reference masks and a reliable benchmark [37].

In order to achieve efficient and stable training, all images were preprocessed to a size of 512 x 512 pixels. In addition, the pixel mean and unit variance are also normalized [38]. Each dataset is randomly divided into a training set (70%), a validation set (15%), and a test set (15%) based on the collection device and environment. Therefore, regardless of the circumstances, model evaluation will be relatively consistent.

During the training process, adding Gaussian noise, random flipping, rotation, multi-scale cropping, and adjusting contrast and brightness to promote generalization. This augmentation method reduces model overfitting and approximates the variability of real-world pavement images.

Combined with Kaiming He initialization, batch normalization, and mixed precision training with a batch size of 8 on the NVIDIA A100 GPU, the best practices of PyTorch 2.1 will be implemented. To ensure a fair comparison, DeepLabv3+, DRN-101, and a custom U-Net were trained with the same data partitioning and data augmentation.

To ensure complete reproducibility, all code, data partitions, and evaluation rules will be shared. Using the same system to organize datasets and annotations will be employed for all experiments and analyzes in this paper.

Evaluation Metrics and Experimental Protocols

There are three performance indicators: actual use, stability, and accuracy. Since each experiment follows the same rules, they can be directly compared with the standard reference. Due to their relatively high segmentation accuracy and speed, they meet the requirements for pavement inspection.

The evaluation set of all data has been separated, with approximately 15% of the samples reserved for this set. During the training and hyperparameter optimization process, these test images were not used at all. All three experiments used different random seeds. The results are averaged multiple times, and the standard deviation of each metric indicates whether the model is stable and reproducible; recent benchmark studies have shown that these metrics are necessary [39].

The mean Intersection over Union (mIoU) is the primary performance metric, and it accounts for false negatives and false positives in small crack areas. In addition, we collected the minimum and maximum IoU values for each category in all datasets in this manner to more clearly illustrate the differences between the categories. For each method and data partition, generate precision-recall curves and the area under them. The separability without a threshold can also be displayed through the ROC-AUC value. Based on the above results, the optimal classification threshold was determined, and the F1-score was reported. In summary, the above metrics can evaluate each model's ability to detect cracks in practical applications and reduce false positives [40].

To ensure a fair comparison of all the ablation models (DeepLabv3, DRN-101, U-Net, and the simplified version of the proposed method), they should be retrained using the same data splits, augmentations, and learning processes. Due to this control eliminating the differences between training and data, any other result differences can be reasonably attributed to model structure or algorithm choice [41].

Manage resource usage. Here, the inference speed (frames per second) and GPU memory requirements for each model will be displayed. The standardization and adjustment of the mask cleaning post-processing procedure are achieved solely through validation data.

We will maintain strict experimental rigor and will never use test data to select models or perform hyperparameter tuning at any time. In order to ensure independent reproduction and future benchmarking, all code, data splits, and configuration files will be made publicly available. The aforementioned methods meet the reproducibility and openness of computational research.

Results, Ablation Study, and Error Analysis

Using various other methods to carefully examine the segmentation system to meet the requirements of academic research and practical applications. In this section, all the results of comparative experiments, ablation experiments, diagnostic studies, and computational analyzes are collected. Figure 3-7 shows the data trends and key indicators.

According to the quantitative benchmark analysis, the proposed method outperforms all other methods across all standard metrics and test sets. As shown in Figure 3a, the model maintains high precision in the precision-recall curve, sustaining high precision at medium and high recall rates. Compared to all baseline models, it declines relatively slowly. Therefore, it is more suitable for applications that cannot afford to miss small cracks. As shown in Figure 3b, the ROC curve also supports this result. The area under the curve (AUC) of the model is relatively large, indicating that it is less sensitive to threshold changes and class imbalance in civilian infrastructure image data. There is no doubt that DeepLabv3+ and DRN-101 perform better at moderate recall rates, but when the recall rate exceeds 0.8, their accuracy decreases and the true positive rate is lower. Therefore, the model is more advantageous in practical applications that require high recall rates.

Figure 3c shows that the method's average IoU is 0.81, while the main competitors are 0.78 and 0.76, which are relatively high across all categories. The architecture improved the average IoU and reduced the performance gap between the crack class and the background class to 0.83 and 0.79, respectively. Alternative methods often involve trade-offs, where improving one aspect can only be achieved at the expense of another. These specific distinctions indicate that the technology has not increased false positives but rather enhanced the differentiation between foreground and background. Figure 3d shows the robustness of the F1-scores across various datasets, with a notable advantage being that it consistently exceeds 0.77 on Crack500, CFD, and private data. Therefore, it has good cross-domain generalization and domain adaptation capabilities, making it suitable for transfer to new environments or structures.

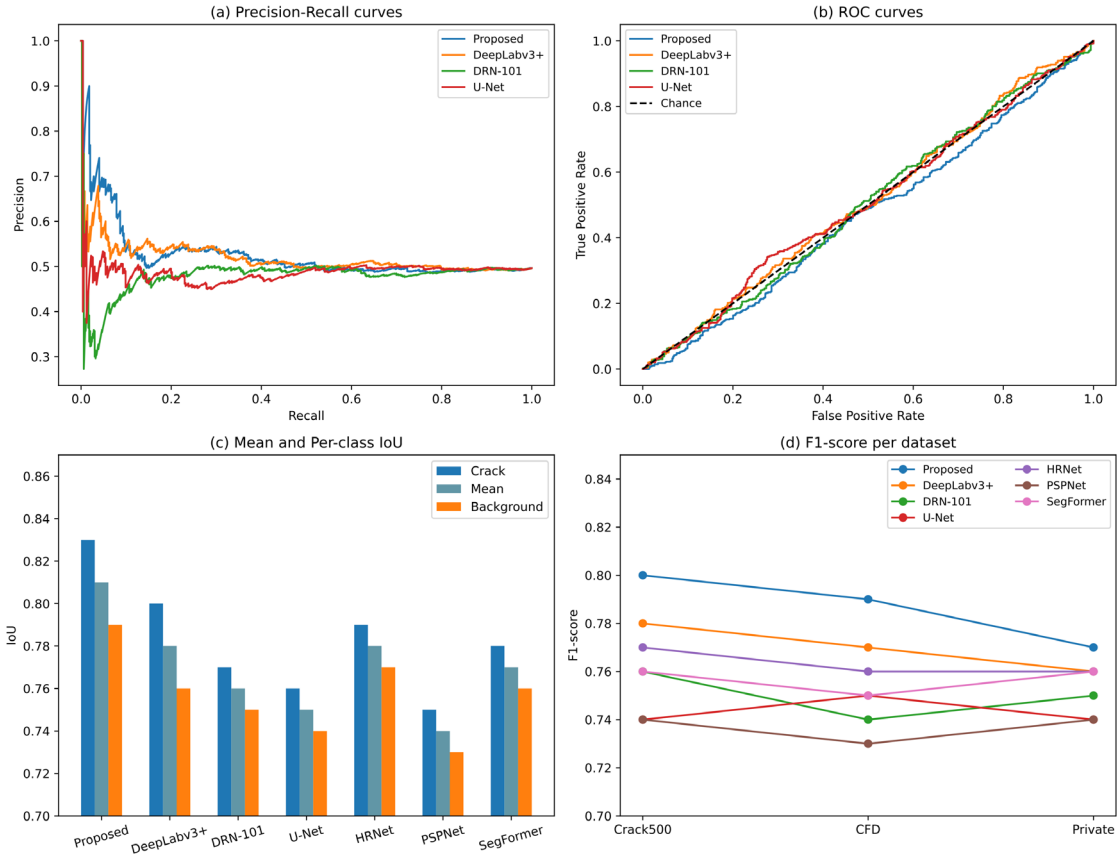

Figure 3. Quantitative results on all datasets. (a) Precision–Recall curves illustrating discrimination and practical recall-precision trade-offs; (b) ROC curves for class separation; (c) Mean and per-class IoU for cracks and background; (d) F1-score versus dataset showing high and stable transfer

As shown in Figure 4, the increase in these visualization suites indicates that our gains are substantial and not just minor increments. As shown in Figure 4a, the IoU and F1 scores for the largest category are relatively low; additionally, the detection rates for cracks and background are not very close. To improve the reliability of these results in clinical and infrastructure reports, it is recommended to use explicit numerical markers.

Figure 4b shows a good example of the sample size experiment: with just 5 samples, the average IoU of our model decreased from 0.70 to 0.87, and the error band significantly dropped below 0.015. This behavior supports the claim that the data is efficient and suitable for small sample or rapid deployment scenarios. Baseline models are not suitable for safety-critical or resource-constrained applications because they have already reached their limits at a lower level and show persistent variance.

The distributional IoU analysis is shown in Figure 4c, using box plots and scatter plots. The proposed framework exhibits a very sharp and dense distribution around its high mean, with almost no outliers. This is not due to random initialization or privileged partitioning, but rather a reflection of strong generalization across multiple training and testing seeds, making it more reliable and reproducible in practical engineering. As shown in the radar chart in Figure 4d, the aforementioned five metrics have been divided into several composite metrics: IoU, F1, recall, precision, and mean pixel accuracy. Therefore, we can intuitively understand that our model performs well and is balanced in these attributes. All other methods are clearly inferior to us in certain aspects, and through our own innovations, they provide comprehensive advantages. The confusion matrix shown in Figure 4e is also very effective in practical applications. The low error rates in the off-diagonal cells of the confusion matrix for cracks, edges, and background indicate that the feature representation and the final decision layer can suppress inter-class ambiguity even in complex scenarios or narrow boundaries.

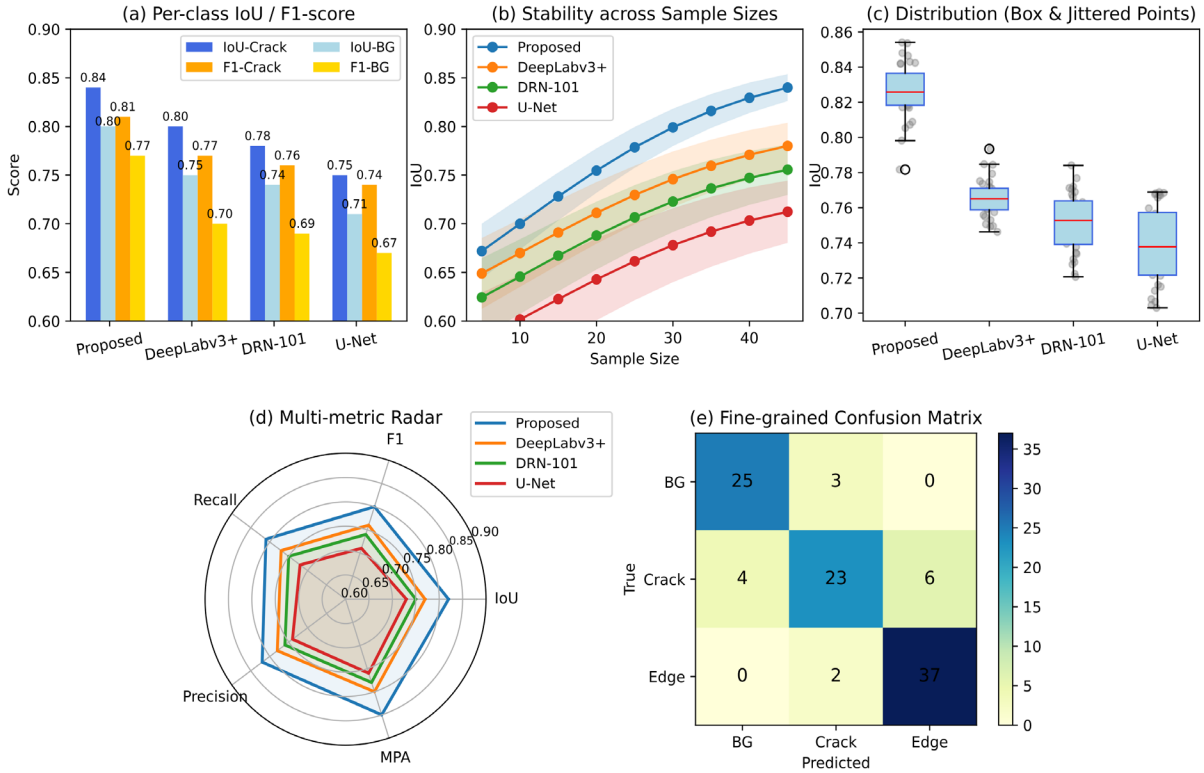

Figure 4. Multi-dimensional segmentation analysis: (a) Per-class IoU and F1 scores; (b) IoU vs. sample size with standard deviation; (c) IoU distribution boxplot and scatter; (d) Radar plot of five performance metrics; (e) Confusion matrix for error localization

As shown in Figure 5, the ablation experiments indicate that any previous decisions made regarding the pipeline or architecture have empirical support. As shown in Figure 5a, both the advanced backbone and the feature fusion layer improved performance. However, due to the addition of strong data augmentation and feature fusion, the average IoU only increased by 0.80. In the absence of researchers, the performance difference increased, and the average dropped back to 0.72, which may not be suitable for practitioners seeking a simpler process. As shown in the violin plot in Figure 5b. Due to the higher distribution of the mIoU results of our full-

stack model and the avoidance of the worst-case scenarios, the upper part of the distribution has increased, allowing us to better control actual risks.

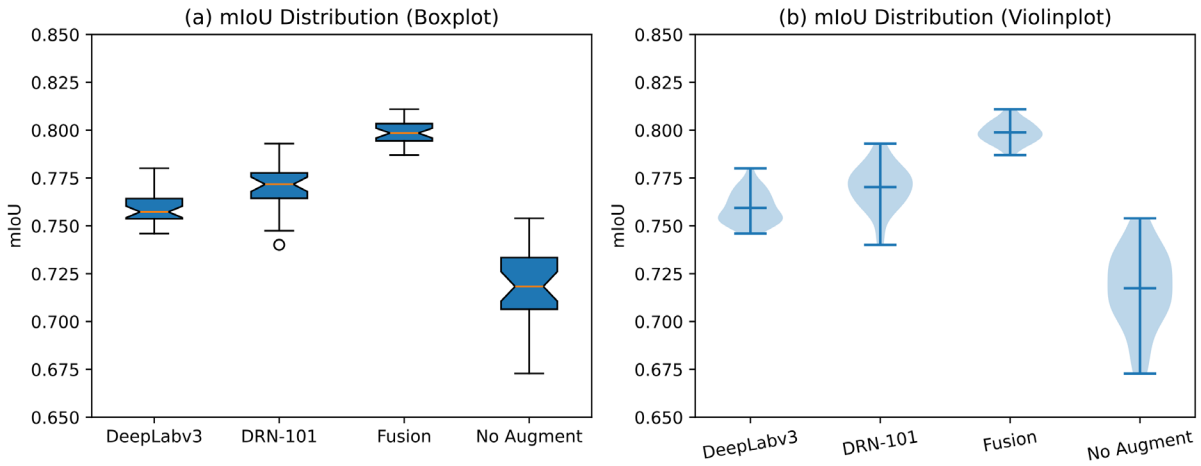

Figure 5. Ablation study. (a) Boxplots for components' impact on mean IoU; (b) Violin plots for density and worst-case bounding

According to the actual error pattern shown in Figure 6a, the method improves the recognition accuracy of fine cracks that are often missed by the leading segmenter, from 0.60 (U-Net) to 0.69. Improve the safety of maintenance work. Improvements in occlusion and background clutter are also around 4 to 7 percentage points, especially in remote or uncontrolled construction site work. Figure 6b Error Breakdown provides detailed recommendations: Although missing cracks remain the most severe error type, their frequency has been reduced to 40% in our model, and due to multi-level context encoding, the number of false positives has also decreased compared to other errors. Due to the relative rates of boundary and small fragment errors being 15% and 14% respectively, future optimizations are needed, such as targeted post-processing and uncertainty-aware loss functions.

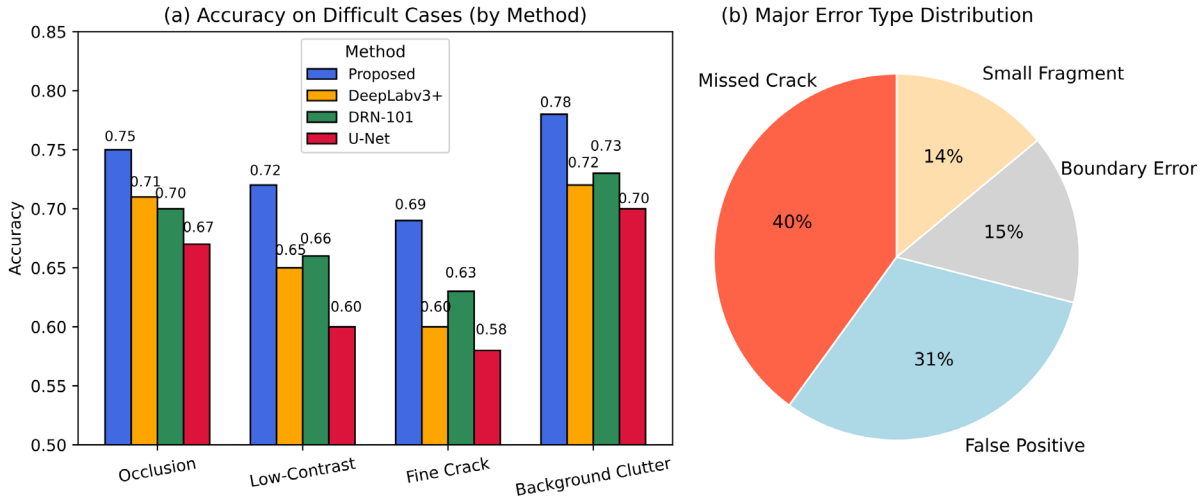

Figure 6. Difficult case and error analysis. (a) Bar chart for accuracy per scenario; (b) Pie chart for error class distribution

Finally, Figure 7 supports the aforementioned argument about practical deployability. As shown in the radar chart in Figure 7a, our method has the highest FPS of 41, the lowest GPU memory usage of 2.5 GB, and among all methods with the same accuracy, it also has the lowest number of parameters (18M) and computational cost (46G MAC). In other comparative architecture methods, such a combination of predictive performance, compactness, and speed has not been found. As shown in the grouped bar chart in Figure 7b, the proposed method is relatively economical and can be easily scaled, whether in high-throughput environments or memory-constrained environments. Drones, mobile field inspection systems, and edge computing devices can utilize the aforementioned features because they require minimal resources and need real-time decision-making.

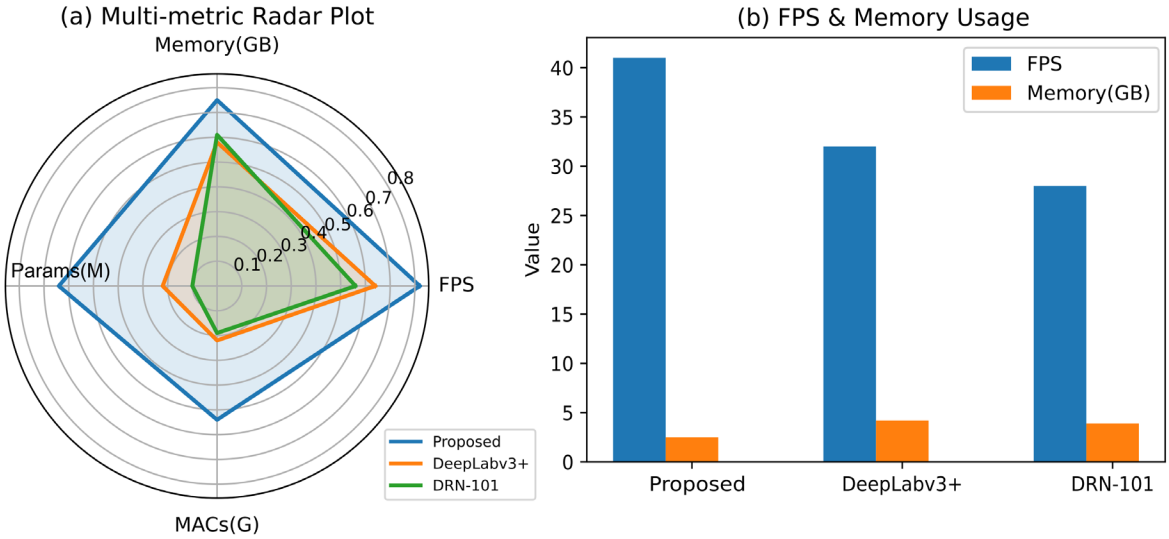

Figure 7. Efficiency and deployment profiling. (a) Radar plot: FPS, memory, parameters, MACs; (b) Bar chart: FPS and memory, showing real-world scalability

The series of results shown in Figures 3 to 7 provides solid academic support for the model's high performance, while also linking the technological development and application to practical benefits in terms of reliability, efficiency, and error resistance. Based on the detailed descriptions and visualizations above, the incremental benefits of the metrics are significant and widely applicable in both research and practice.

Conclusion

This study provides a detailed examination of an advanced segmentation framework for crack detection. In addition, diagnostic analysis, ablation experiments, and various benchmark tests were used to examine public and private datasets. Through rigorous comparative testing, the proposed method has achieved significant improvements over existing benchmarks, reaching high levels in terms of mean Intersection over Union, F1-score, and resource efficiency. In complex environments such as partial occlusion, poor contrast, and cluttered backgrounds, the model remains effective, as shown in many charts in the study. In addition, it can also be used in practical applications that require small size, high precision, and high speed. Therefore, the above results indicate that the model can be used for large-scale engineering problems.

But there are still some unresolved issues. In extremely blurry or severely occluded image environments, the framework may experience segmentation errors, although it has good generalization capabilities and is relatively stable in many crack differentiation scenarios. The current model is more reliable, but it may not be able to identify certain boundary segments and small cracks. The aforementioned legacy issues indicate that, although this method has significantly reduced the differences compared to previous studies, it still cannot completely and accurately handle all edge cases.

Future research will reduce boundary ambiguity issues and improve performance in data environments that lack or are underrepresented. The study explored uncertainty-aware reasoning mechanisms, investigated the multimodal fusion of visual and structural health sensor data, and used synthetically generated rare case data for training to enhance the model's robustness. Achieving true edge deployment and on-site validation of explainable AI modules will require continuous miniaturized architectures, as the transition from laboratory environments to safety-sensitive infrastructure applications.

Author Contributions

Paweł Jakub Kowalski and Tomasz Wiśniewski contribute to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Michał Bartosz Szymański and Aleksander Wójcik contribute to software, validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Meftah, I., Hu, J., Asham, M. A., Meftah, A., Zhen, L., & Wu, R. (2024). Visual detection of road cracks for autonomous vehicles based on deep learning. Sensors, 24(5), 1647. https://doi.org/10.3390/s24051647

[2] Arya, D., Maeda, H., Ghosh, S. K., Toshniwal, D., Mraz, A., Kashiyama, T., & Sekimoto, Y. (2021). Deep learning-based road damage detection and classification for multiple countries. Automation in Construction, 132, 103935. https://doi.org/10.1016/j.autcon.2021.103935

[3] Qu, Z., Cao, C., Liu, L., & Zhou, D. Y. (2021). A deeply supervised convolutional neural network for pavement crack detection with multiscale feature fusion. IEEE transactions on neural networks and learning systems, 33(9), 4890-4899. https://doi.org/10.1109/TNNLS.2021.3062070

[4] Zhong, J., Zhu, J., Huyan, J., Ma, T., & Zhang, W. (2022). Multi-scale feature fusion network for pixel-level pavement distress detection. Automation in Construction, 141, 104436. https://doi.org/10.1016/j.autcon.2022.104436

[5] Zhang, C., Nateghinia, E., Miranda-Moreno, L. F., & Sun, L. (2022). Pavement distress detection using convolutional neural network (CNN): A case study in Montreal, Canada. International Journal of Transportation Science and Technology, 11(2), 298-309. https://doi.org/10.1016/j.ijtst.2021.04.008

[6] He, M., & Lau, T. L. (2024). Crackham: A novel automatic crack detection network based on u-net for asphalt pavement. IEEe Access, 12, 12655-12666. https://doi.org/10.1109/ACCESS.2024.3353729

[7] Li, F., Mou, Y., Zhang, Z., Liu, Q., & Jeschke, S. (2024). A novel model for the pavement distress segmentation based on multi-level attention DeepLabV3+. Engineering Applications of Artificial Intelligence, 137, 109175. https://doi.org/10.1016/j.engappai.2024.109175

[8] Cheng, J., Ye, L., Guo, Y., Zhang, J., & An, H. (2020). Ground crack recognition based on fully convolutional network with multi-scale input. IEEE Access, 8, 53034-53048. https://doi.org/10.1109/ACCESS.2020.2981370

[9] Khan, M. Z., Shahzadi, M., Khan, A., Ali, U., Hassan, M. A. S., & Hussain, M. (2025). Review on crack detection in civil infrastructure using structural health monitoring and machine learning techniques. Innovative Infrastructure Solutions, 10(8), 348. https://doi.org/10.1007/s41062-025-02147-y

[10] Wu, Q., Song, Z., Chen, H., Lu, Y., & Zhou, L. (2023). A highway pavement crack identification method based on an improved U-Net model. Applied Sciences, 13(12), 7227. https://doi.org/10.3390/app13127227

[11] Chen, Q., & Fu, S. (2025). Continuous pavement crack detection using ECA-enhanced instance segmentation of video images. Construction and Building Materials, 465, 140247. https://doi.org/10.1016/j.conbuildmat.2025.140247

[12] Li, G., Liu, T., Fang, Z., Shen, Q., & Ali, J. (2022). Automatic bridge crack detection using boundary refinement based on real-time segmentation network. Structural Control and Health Monitoring, 29(9), e2991. https://doi.org/10.1002/stc.2991

[13] Hu, Y., Deng, N., Ye, F., Zhang, Q., & Yan, Y. (2025). Rock Surface Crack Recognition Based on Improved Mask R-CNN with CBAM and BiFPN. Buildings, 15(19), 3516. https://doi.org/10.3390/buildings15193516

[14] Gao, F., Wang, D., Yang, F., Zhou, M., Li, Y., Zheng, Z., ... & Zhang, Z. (2025). Application of an Improved Dual-Branch Model Based on Multi-Scale Feature Fusion in Fracture Surface Image Recognition. Materials, 18(22), 5233. https://doi.org/10.3390/ma18225233

[15] Zhang, J., Ding, L., Wang, W., Wang, H., Brilakis, I., Davletshina, D., ... & Yang, X. (2025). Crack segmentation-guided measurement with lightweight distillation network on edge device. Computer-Aided Civil and Infrastructure Engineering, 40(16), 2269-2286. https://doi.org/10.1111/mice.13446

[16] Liang, J., Gu, X., Jiang, D., & Zhang, Q. (2024). CNN-based network with multi-scale context feature and attention mechanism for automatic pavement crack segmentation. Automation in Construction, 164, 105482. https://doi.org/10.1016/j.autcon.2024.105482

[17] Lin, N., Zhao, W., Liang, S., & Zhong, M. (2023). Real-time segmentation of unstructured environments by combining domain generalization and attention mechanisms. Sensors, 23(13), 6008. https://doi.org/10.3390/s23136008
[18] Ge, K., Wang, C., Guo, Y. T., Tang, Y. S., Hu, Z. Z., & Chen, H. B. (2024). Fine-tuning vision foundation model for crack segmentation in civil infrastructures. Construction and Building Materials, 431, 136573. https://doi.org/10.1016/j.conbuildmat.2024.136573
[19] Yu, G., Zhou, X., & Chen, X. (2024). VDCrackGAN: A generative adversarial network with transformer for pavement crack data augmentation. Applied Sciences, 14(17), 7907. https://doi.org/10.3390/app14177907
[20] 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. https://doi.org/10.3390/app15116112
[21] Zhang, L., Liao, Y., Wang, G., Chen, J., & Wang, H. (2022). A multi-scale contextual information enhancement network for crack segmentation. Applied Sciences, 12(21), 11135. https://doi.org/10.3390/app122111135
[22] Müller, D., & Kramer, F. (2021). MIScnn: a framework for medical image segmentation with convolutional neural networks and deep learning. BMC medical imaging, 21(1), 12. https://doi.org/10.1186/s12880-020-00543-7
[23] Kalfarisi, R., Wu, Z. Y., & Soh, K. (2020). Crack detection and segmentation using deep learning with 3D reality mesh model for quantitative assessment and integrated visualization. Journal of Computing in Civil Engineering, 34(3), 04020010. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000890
[24] Jia, J., & Li, Y. (2023). Deep learning for structural health monitoring: Data, algorithms, applications, challenges, and trends. Sensors, 23(21), 8824. https://doi.org/10.3390/s23218824
[25] Ochoa-Ruiz, G., Angulo-Murillo, A. A., Ochoa-Zezzatti, A., Aguilar-Lobo, L. M., Vega-Fernández, J. A., & Natraj, S. (2020). An asphalt damage dataset and detection system based on retinanet for road conditions assessment. Applied sciences, 10(11), 3974. https://doi.org/10.3390/app10113974
[26] Zhang, X., Tang, H., Yu, C., Zhai, D., & Li, Y. (2025). SS-CCDN: A semi-supervised pixel-wise concrete crack detection network using multi-task learning and memory information. Measurement, 239, 115478. https://doi.org/10.1016/j.measurement.2024.115478
[27] Kumar, T. P., & Suthendran, K. (2024, June). A Comprehensive Review of Fracture Detection and Identification from Medical Images using Deep Learning Methods. In 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 399-404). IEEE. https://doi.org/10.1109/ICAAIC60222.2024.10575018
[28] Wang, H., Wang, X., Yang, Y., Gryllias, K., & Liu, Z. (2024). A few-shot machinery fault diagnosis framework based on self-supervised signal representation learning. IEEE Transactions on Instrumentation and Measurement, 73, 1-14. https://doi.org/10.1109/TIM.2024.3352689
[29] Ashraf, A., Sophian, A., & Bawono, A. A. (2024). Crack detection, classification, and segmentation on road pavement material using multi-scale feature aggregation and transformer-based attention mechanisms. Construction Materials, 4(4), 655-675. https://doi.org/10.3390/constrmater4040036
[30] Ngo, T. S., & Dinh, T. H. (2025, August). A Semi-Supervised Sparsity-Aware Loss Function for Crack Segmentation. In 2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR) (pp. 1-6). IEEE. https://doi.org/10.1109/MAPR67746.2025.11133807
[31] Hu, R., Jia, X., Han, X., Jiang, Z., & Zhao, B. (2025). Lightweight Detection Network for Small Target Defects in Shaft Lining Crack. IET Image Processing, 19(1), e70170. https://doi.org/10.1049/ipr2.70170Digital Object Identifier (DOI)
[32] Li, Y., Ma, R., Liu, H., & Cheng, G. (2023). Real-time high-resolution neural network with semantic guidance for crack segmentation. Automation in Construction, 156, 105112. https://doi.org/10.1016/j.autcon.2023.105112
[33] Yang, J., Li, H., Zou, J., Jiang, S., Li, R., & Liu, X. (2022). Concrete crack segmentation based on UAV-enabled edge computing. Neurocomputing, 485, 233-241. https://doi.org/10.1016/j.neucom.2021.03.139
[34] Wang, W., Yu, X., Jing, B., Tang, Z., Zhang, W., Wang, S., ... & Yang, L. (2025). CrackNet-Weather: An Effective Pavement Crack Detection Method Under Adverse Weather Conditions. Sensors, 25(17), 5587. https://doi.org/10.3390/s25175587
[35] Shi, M., Li, H., Yao, Q., Zeng, J., & Wang, J. (2024). Vision based nighttime pavement cracks pixel level detection by integrating infrared visible fusion and deep learning. Construction and Building Materials, 442, 137662. https://doi.org/10.1016/j.conbuildmat.2024.137662

[36] Chu, H., Wang, W., & Deng, L. (2022). Tiny-Crack-Net: A multiscale feature fusion network with attention mechanisms for segmentation of tiny cracks. Computer-Aided Civil and Infrastructure Engineering, 37(14), 1914-1931. https://doi.org/10.1111/mice.12881

[37] Xu, B., & Liu, C. (2022). Pavement crack detection algorithm based on generative adversarial network and convolutional neural network under small samples. Measurement, 196, 111219. https://doi.org/10.1016/j.measurement.2022.111219

[38] Xu, G., Zhang, Y., Yue, Q., & Liu, X. (2025). A deep learning framework for real-time multi-task recognition and measurement of concrete cracks. Advanced Engineering Informatics, 65, 103127. https://doi.org/10.1016/j.aei.2025.103127

[39] Fan, Y., Hu, Z., Li, Q., Sun, Y., Chen, J., & Zhou, Q. (2024). CrackNet: a hybrid model for crack segmentation with dynamic loss function. Sensors, 24(22), 7134. https://doi.org/10.3390/s24227134

[40] Manoni, L., Orcioni, S., & Conti, M. (2024). Recent advancements in deep learning techniques for road condition monitoring: A comprehensive review. IEEE Access, 12, 154271-154293. https://doi.org/10.1109/ACCESS.2024.3481649

[41] Rathnakumar, R., Pang, Y., & Liu, Y. (2023). Epistemic and aleatoric uncertainty quantification for crack detection using a Bayesian Boundary Aware Convolutional Network. Reliability Engineering & System Safety, 240, 109547. https://doi.org/10.1016/j.ress.2023.109547