Human Factors, Architecture, Sustainable Urban Planning and Infrastructure, Vol. 229, 2026, 180–190
https://doi.org/10.54941/ahfe1007966

Bridge Inspection Support Framework Incorporating Human Reliability

Yuki Murata[1], Kosei Koizumi[1], Akito Sakurai[2], and Yusaku Okada[3]

[1] Graduate School of Science and Technology, Keio University, Japan
[2] Ohsaki Research Institute, Inc., Japan
[3] Faculty of Science and Technology, Keio University, Japan

ABSTRACT

In Japan, periodic road-bridge inspections rely primarily on inspectors' close visual examination at five-year intervals. Concurrently, the proportion of aging bridges is increasing while the availability of experienced inspectors is expected to decline, amplifying the demand for support tools that reduce workload without compromising safety assurance. This study proposes a bridge-deck inspection support framework grounded in human–AI collaboration and presents its core application for crack-candidate visualization on concrete decks. The framework enforces an explicit division of roles: the AI is restricted to presenting and visualizing crack candidates, whereas inspectors retain responsibility for close visual inspection, on-site measurement, and final condition assessment. As the AI component, we developed a semantic-segmentation model that estimates crack presence at the pixel level and implemented post-processing to suppress short, spurious detections while preserving line-like crack structures. Evaluation results indicate that missed detections of cracks as continuous line-like structures were rare in our test set, including images with diverse surface appearances. False positives occurred predominantly as short segments, which—while requiring verification—are typically less hazardous than misses in safety-critical inspections and can be treated as conservative prompts for closer examination. Accordingly, this study frames crack detection not solely as an accuracy problem but as a human reliability intervention: a "safety-oriented filter" that redistributes attention from exhaustive search to prioritized verification, thereby mitigating lapse-type inspection errors. The findings suggest that the proposed framework can support safer attention allocation while maintaining inspector accountability.

Keywords: Bridge inspection, Human–AI collaboration, Semantic segmentation, Cognitive load, Attention guidance, Fail-safe mechanism, Human reliability

INTRODUCTION

In Japan, road bridges are, in principle, subject to periodic inspections every five years, with close visual inspection conducted by human inspectors. Demographic change is expected to reduce the pool of experienced inspectors and increase the proportion of less experienced personnel in the field. In addition, projections indicate that by 2035 approximately 65% of bridges

Received February 5, 2026; Revised April 8, 2026; Accepted April 27, 2026; Available online July 20, 2026

will be more than 50 years old, implying a growing share of structures requiring repair and, consequently, an increasing inspection workload under the current approach.

While AI-based image analysis has advanced and commercial services for automated defect detection have emerged, field adoption remains limited. Two issues are particularly salient. First, many deployed systems are designed around end-to-end automation (e.g., detecting cracks and producing inspection records), effectively positioning AI as a substitute for inspector judgment. In practice, however, stakeholders do not necessarily expect or accept full replacement. Overreliance on automated judgments can blur responsibility boundaries and may degrade vigilance, undermining the safety function of inspection. For example, a Ministry of Land, Infrastructure, Transport and Tourism survey reported practitioner feedback indicating that new technologies were not adopted because close visual inspection remained indispensable—suggesting a mismatch between technological assumptions and operational reality. Second, as highlighted by recent studies, the generalizability of AI models to variability in deck surface appearance remains insufficient (see Current Research). These considerations motivate treating inspection support as a human reliability problem, rather than reducing it to detection accuracy alone.

CURRENT RESEARCH

Recent literature indicates that sustaining model performance across concrete deck images with heterogeneous visual characteristics remains challenging. Rashid et al. (2025) reported that models can achieve high accuracy under in-dataset training/testing but degrade substantially under cross-dataset evaluation, implying unresolved robustness to domain shifts. Flotzinger et al. (2024) similarly noted that widely used datasets for reinforced-concrete defect recognition often reflect limited numbers of bridges and acquisition conditions, raising concerns about real-world applicability. Even when diversity is increased through real-world inspection datasets, the reported outcomes suggest that practical-level generalization remains difficult to achieve.

Collectively, these findings position operational generalizability—maintaining usable performance despite variability in deck surface appearance (e.g., texture, staining, weathering, efflorescence)—as a central open challenge for image-based defect recognition. Importantly, from a human factors perspective, the risk is not only performance degradation per se, but also the inspector's inability to anticipate when and how the AI will fail under domain shifts. Such unpredictability impedes calibrated trust and can lead to inappropriate reliance ("overtrust") or disuse ("undertrust"), both of which can degrade joint system performance.

RESEARCH OBJECTIVES

This study focuses on cracks, which are often early manifestations of diverse deterioration processes, and pursues two objectives:

1. To propose a bridge inspection support framework that explicitly allocates cognitive responsibilities between humans and AI to enhance inspection reliability.
2. To examine whether the system remains usable under variability in bridge-deck surface conditions (i.e., robustness to appearance variability).

Rather than pre-empting a fully automated future, the goal is to establish a pragmatic collaborative model that bridges current inspection practice and plausible near-term operational deployment.

PROPOSED BRIDGE INSPECTION SUPPORT FRAMEWORK

Compared with conventional bridge-deck inspections that rely on inspectors' exhaustive close visual search, the proposed framework introduces AI-based crack-candidate presentation to support a shift from broad, continuous scanning (search) to prioritized verification (confirmation). Figure 1 outlines the procedure, which consists of four steps:

1. Capture images of the bridge deck.
2. Visualize crack-candidate lines using the AI application.
3. Inspectors review the AI output and determine where to focus attention.
4. Inspectors conduct close visual inspection, measurement, and documentation.

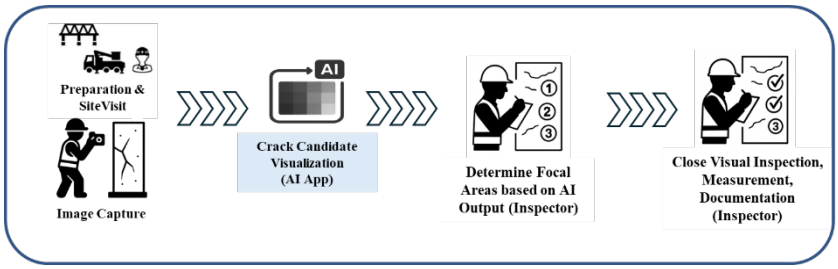

Figure 1: Proposed bridge inspection support framework.

Premised on human–AI collaboration, the AI is not positioned as a substitute for judgment but as a perceptual aid that supports damage search. Accordingly, the AI is limited to estimating and visualizing crack candidates; it does not replace close inspection, measurement, or record preparation.

Notably, the proposed system does not estimate crack width. Crack width is not uniquely determined because recorded values depend on the measurement location and conditions, and the basis for strict definition and rigorous validity assessment remains limited. In current practice, inspectors' on-site measurements are treated as official values. Therefore, the framework intentionally assigns width measurement and final interpretation to the inspector, while using AI to support the upstream decision of where careful checking is warranted.

The primary contribution of the framework is to enable a controlled transition from exhaustive search to focused verification while preserving accountability: ultimate responsibility for inspection outcomes remains with the human inspector. In addition, because AI outputs can be stored as image-based evidence linked to human decisions, the framework may improve explainability and consistency in the judgment process. Finally, less experienced inspectors may benefit from candidate presentation as a scaffold for verification, suggesting potential training value.

To operationalize this framework, an AI application that visualizes crack candidates in a manner aligned with inspection needs is required. The next section describes the implementation.

IMPLEMENTATION: CRACK CANDIDATE VISUALIZATION APPLICATION

To operationalize the proposed workflow, we developed a prototype application that estimates crack candidates from bridge-deck images and visualizes them for inspection support. The application employs deep learning–based semantic segmentation to predict crack presence at the pixel level.

The training data comprised bridge-deck images collected from 8 bridges (8 panels, 28 locations). For training and testing, each image was divided into 256 × 256 pixel patches. Figure 2 shows examples of patch images.

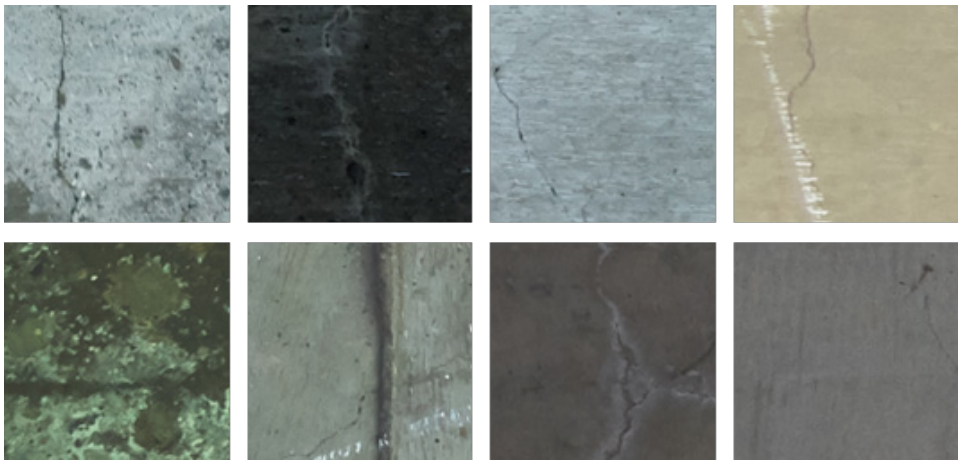

Figure 2: Examples of patch images used in the experiments.

1) Dataset Construction

Training requires paired data consisting of each patch image and its corresponding ground-truth mask. Based on expert annotations of crack locations and widths, together with image resolution, we generated ground-truth masks as illustrated in Figure 3. Importantly, the masks adopt a point-set representation (sparse crack pixels) rather than a densely filled region mask, reflecting the thin and line-like nature of cracks.

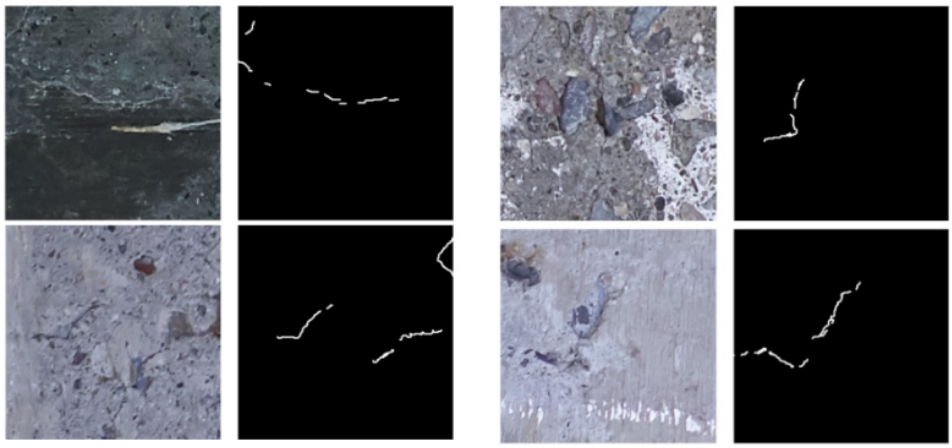

Figure 3: Example of a ground-truth mask generated for training.

2) Training Procedure

We trained a semantic segmentation model using patch–mask pairs, where the input is a patch image and the output is a pixel-wise crack/non-crack prediction. To ensure an unbiased evaluation, all patches derived from the six deck images reserved for testing were excluded from training. The remaining data were split into training and validation sets (8:2). In total, 24,457 pairs were used for training, 6,115 for validation, and 5,620 for testing.

Hyperparameters are summarized in Table 1.

Table 1: Hyperparameters used in training.

Item	Setting
Architecture	U-Net
Encoder	MiT-B5
Encoder Patch size	16
Max epochs	50
Optimizer	Adam
Loss function	$0.9\times DiceLoss_{crack}+0.1\times DiceLoss_{background}$

The loss function combines two Dice-based terms: one emphasizing correct segmentation of crack pixels and the other emphasizing correct

classification of background pixels. Because the application is intended for inspection support, the weighted sum (crack:background = 0.9:0.1) prioritizes suppression of missed detections, accepting a degree of over-detection consistent with a conservative "safety-filter" design. The checkpoint achieving minimum validation loss was selected for the final test evaluation.

3) Evaluation Method

We first applied an area-based post-processing procedure to the model predictions on the test images. Predicted crack pixels were grouped into connected components using 3 × 3 neighborhood connectivity, and components smaller than 4 mm^2 were removed. This filtering aims to reduce short, spurious predictions that could clutter attention guidance, while retaining slender line-like structures that reflect continuous cracks.

After filtering, we conducted qualitative and quantitative evaluations. Qualitative inspection used overlays of predictions on the original images, focusing on (i) tendencies of over-detection and missed detection, (ii) continuity of line-like crack structures, and (iii) whether the visualization plausibly supports attention allocation during close visual inspection.

Quantitatively, we computed Precision, Recall, and F1-score. Because cracks are thin structures, small spatial misalignments between prediction and ground truth can disproportionately affect pixel-wise metrics. In inspection support, however, approximate localization can still be sufficient to guide attention. Therefore, we adopted Relaxed Precision, Relaxed Recall, and Relaxed F1-score with a tolerance radius. We examined tolerance radii from 0.3 mm to 0.8 mm and confirmed that results were not drastically altered; consequently, we selected 0.3 mm (≈1.2 pixels) as a stricter setting to absorb minimal discretization effects while maintaining rigorous localization expectations.

(i) Relaxed Precision: A predicted crack pixel is correct if at least one ground-truth crack pixel exists within the tolerance radius.
(ii) Relaxed Recall: A ground-truth crack pixel is detected if at least one predicted crack pixel exists within the tolerance radius.
(iii) Relaxed F1-score: The harmonic mean of Relaxed Precision and Relaxed Recall.

Finally, suitability for inspection support was assessed comprehensively using three criteria: (1) suppression of missed detection of cracks as continuous line-like structures, (2) avoidance of excessive over-detection that would hinder attention guidance, and (3) consistency between quantitative metrics and qualitative observations.

RESULTS AND DISCUSSION

Table 2 summarizes quantitative results after post-processing for the six test images. Test Images 1–4 exhibit substantial variability in deck surface conditions, whereas Test Images 5–6 are characterized by efflorescence.

Table 2: Quantitative results for the six test images (after post-processing).

	R-Precision	R-Recall	R-F1-score
Test Image 1	0.46	0.62	0.52
Test Image 2	0.73	0.59	0.65
Test Image 3	0.62	0.70	0.66
Test Image 4	0.74	0.65	0.69
Test Image 5	0.49	0.53	0.51
Test Image 6	0.34	0.55	0.42

Figures 4–9 visualize the prediction results for the six test images.

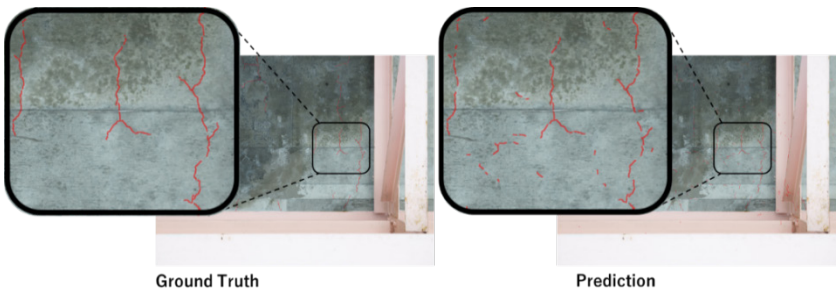

Figure 4: Visualization results for test image 1.

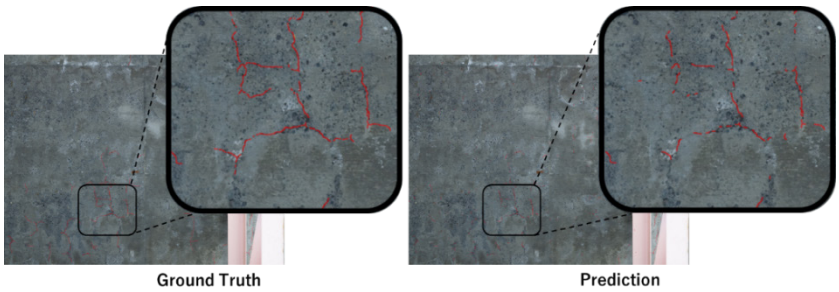

Figure 5: Visualization results for test image 2.

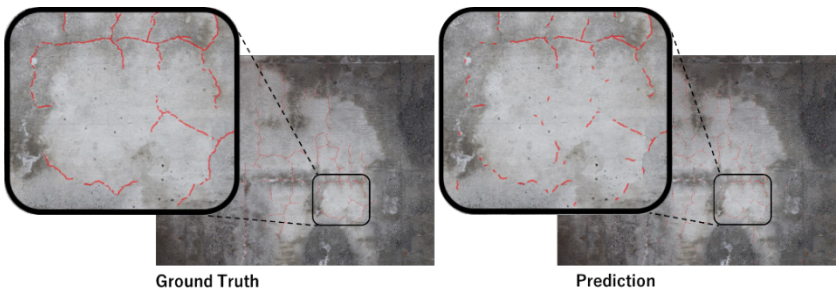

Figure 6: Visualization results for test image 3.

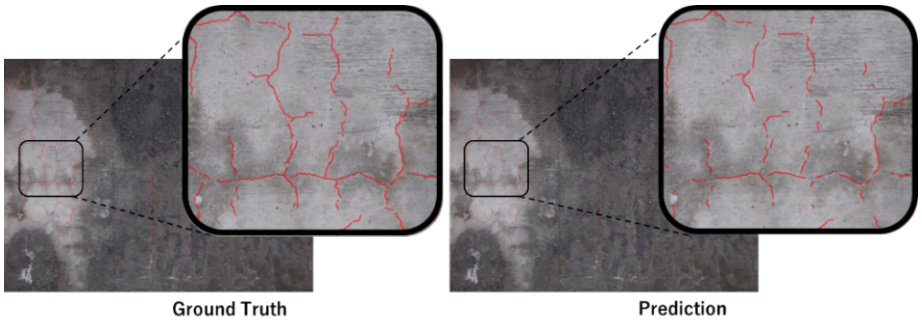

Figure 7: Visualization results for test image 4.

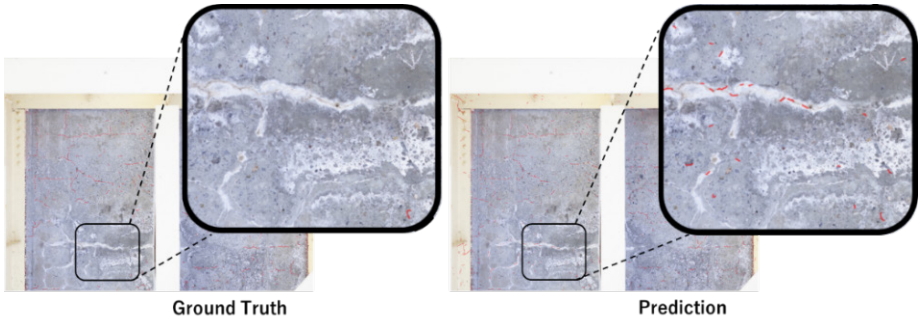

Figure 8: Visualization results for test image 5.

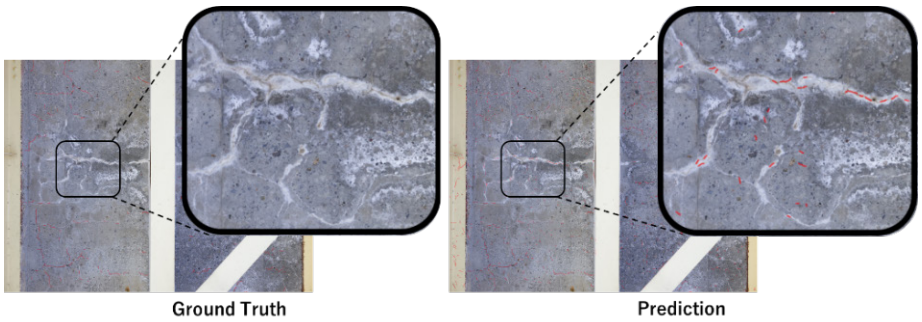

Figure 9: Visualization results for test image 6.

1) Performance Analysis for Inspection Support

For Test Images 1–4 (Figures 4–7), the predicted outputs captured line-like crack structures corresponding to the ground-truth masks. Some intermittency was observed; however, the outputs remained interpretable as candidates for subsequent expert verification, and the observed pattern does not appear to undermine the primary support objective—namely, reducing the risk of missing continuous cracks. Overall, the combination of semantic

segmentation and post-processing suggests practical potential for attention guidance under diverse surface appearances.

For Test Images 5–6 (Figures 8–9), quantitative metrics were lower than for the other images. One plausible contributor is that efflorescence and discoloration can produce visual patterns that are ambiguous even to human annotators, which may increase apparent false positives when ground truth does not label certain ambiguous regions as cracks. Within an inspection-support scope, such conservative flagging may be an acceptable trade-off: it surfaces ambiguous areas for human verification, prioritizing miss avoidance over aggressive noise suppression. Nevertheless, because false alarms can increase verification workload, future field evaluation should explicitly assess the workload implications and usability thresholds under efflorescence-dominant conditions.

2) Contribution to Human Reliability

Beyond numerical accuracy, the framework's validity depends on how the AI output reshapes human work in a safety-critical context. The central premise is complementarity between typical human and AI error tendencies:

Error management (misses vs. false alarms): In prolonged, repetitive visual search, humans are susceptible to vigilance decrement and lapse-type errors (misses). In contrast, a conservative segmentation model may produce false alarms. The proposed workflow deliberately shifts the inspector's task from exhaustive search to adjudicating candidates. In many inspection contexts, filtering false alarms is operationally safer than recovering from misses, provided that false-alarm density remains within a manageable range.

Attention Guidance and cognitive economy: By converting large-area scanning into prioritized verification, the AI output can reduce the cognitive demands of continuous search and support more strategic allocation of limited attentional resources to high-suspicion regions.

Authority and accountability (trust calibration): Because the AI provides visualization rather than a categorical pass/fail decision, inspectors retain authority over interpretation and final judgment. This design aims to reduce automation bias and support calibrated trust: the AI functions as a cueing aid, while responsibility boundaries remain explicit.

Taken together, the framework can be interpreted as a reliability-oriented redesign of the inspection workflow, in which AI functions as a conservative attention-shaping mechanism rather than a decision-maker.

CONCLUSION

Figure 10 summarizes the proposed human–AI collaborative inspection workflow and the processing pipeline of the crack-candidate visualization application developed in this study.

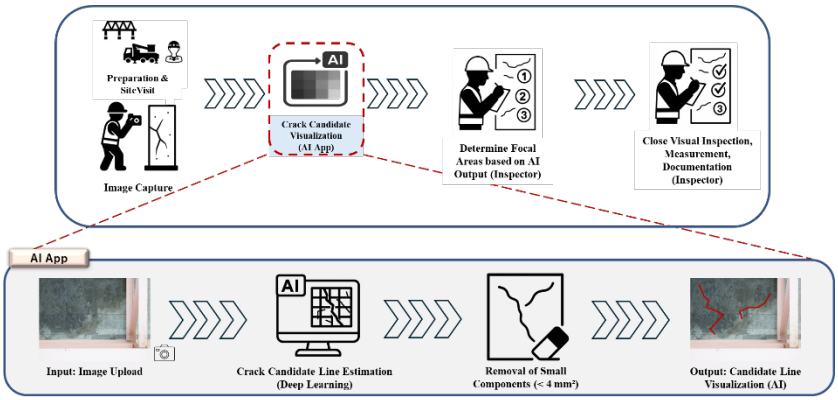

Figure 10: Proposed inspection workflow and AI application pipeline in this study.

This study proposed a human–AI collaborative bridge-deck inspection framework in which AI supports inspectors by visualizing crack candidates while inspectors retain responsibility for close visual inspection, measurement, and final assessment. A semantic-segmentation–based prototype application was developed to operationalize this framework. Evaluation results indicate that missed detections of cracks as continuous line-like structures were rare in our test set, including images with diverse surface appearances. Lower quantitative scores in efflorescence-dominant images may reflect conservative sensitivity to ambiguous surface anomalies; within an inspection-support scope, such behavior can function as a fail-safe cue for verification, although its workload impact must be assessed.

Overall, the findings suggest that inspection-support systems should not be evaluated solely by detection accuracy, but also by how they restructure attention allocation, responsibility boundaries, and human reliability in safety-critical workflows. Future work will conduct field-based evaluations with practicing inspectors to quantify efficiency gains, usability thresholds, and risk-reduction effects in operational settings.

REFERENCES

Chen, J. et al. (2022) The Improvement of Automated Crack Segmentation on Concrete Pavement with Graph Network Yong Zhang (ed.). *Journal of advanced transportation*. 2022, 2238095.

Flotzinger, J. et al. (2024) dacl1k: Real-world bridge damage dataset putting open-source data to the test. Engineering applications of artificial intelligence. 137, 109106.

Hussain, T. et al. (2025) Pixel-level crack segmentation and quantification enabled by multi-modality cross-fusion of RGB and depth images. Construction & building materials. 487, 141961.

Liu, H. et al. (2023) CrackFormer Network for Pavement Crack Segmentation. IEEE transactions on intelligent transportation systems. 24 (9), 9240–9252.

Ma, K. et al. (2024) Coarse–Fine Combined Bridge Crack Detection Based on Deep Learning. Applied sciences. 14 (12), 5004.

Rashid, T. et al. (2025) Cross-dataset evaluation of deep learning models for crack classification in structural surfaces. Journal of mechanical behaviour of materials. 34 (1), 20250074.
Shim, S. (2025) Semantic segmentation for crack detection via generative knowledge distillation. *Automation in construction*. 175, 106201.
Zaheer, Q. et al. (2025) Intelligent Multitasking Framework for Boundary-Preserving Semantic Segmentation, Width Estimation, and Propagation Modeling of Concrete Cracks. *Journal of infrastructure systems*. 31 (3), 04025009.
Zhang, J. et al. (2022) Automated bridge surface crack detection and segmentation using computer vision-based deep learning model. *Engineering applications of artificial intelligence*. 115, 105225.