A Structure-Aware Deep Learning Framework for Automated Bridge Inspection Integrating SegFormer-Based Structural Member Segmentation and YOLOv8 Damage Detection

Sushama De Silva * and Pang-jo Chun

Posted Date: 20 May 2026

doi: 10.20944/preprints202605.1350.v1

Keywords: bridge inspection; deep learning; structural member segmentation; damage detection; SegFormer; YOLOv8; Segment Anything Model; structure-aware analysis; semantic segmentation; spatial damage mapping; damage-to-member association

 doi:10.20944/preprints202605.1350.v1

Article

A Structure-Aware Deep Learning Framework for Automated Bridge Inspection Integrating SegFormer-Based Structural Member Segmentation and YOLOv8 Damage Detection

Sushama De Silva and Pang-jo Chun *

Institute of Engineering Innovation, School of Engineering, The University of Tokyo, Tokyo, Japan

* Correspondence: sushama@g.ecc.u-tokyo.ac.jp

Highlights

This study presents a structure-aware deep learning framework that integrates transformer-based segmentation and object detection to automatically associate bridge damage with specific structural members, producing outputs such as "crack on main girder" directly usable for maintenance prioritization.

What are the main findings?

- The SegFormer-based segmentation model achieved a test mIoU of 0.851 for three structural member classes (main girder, deck slab, abutment), with SAM mask-prompt refinement outperforming bounding-box prompting.
- The integrated pipeline achieved 70.0% fully correct damage detection and 62.0% fully correct member assignment on 100 real bridge inspection images, with the main girder class reaching 90.9% and 93.9% respectively.

What are the implications of the main findings?

- Structure-aware inspection outputs that explicitly link damage type to structural location provide more actionable information for maintenance decision-making than conventional damage-only detection approaches.
- The framework demonstrates that combining foundation model refinement (SAM) with supervised segmentation and one-stage detection is a viable and practical direction for automated infrastructure monitoring.

Abstract

Aging bridge infrastructure and limited inspection resources have created an urgent need for automated and reliable bridge condition assessment systems. Most existing deep learning-based inspection approaches detect damage types from images without considering the structural member on which the damage occurs, limiting their practical utility for maintenance decision-making. This study proposes a structure-aware deep learning framework for automated bridge inspection that integrates structural member segmentation, multiclass damage detection, and spatial damage-to-member association within a unified pipeline. A SegFormer-based semantic segmentation model was trained on a custom bridge inspection dataset comprising 1,339 images to identify three primary structural member classes — main girder, deck slab, and abutment — achieving a test mean Intersection over Union (mIoU) of 0.851. Boundary refinement using the Segment Anything Model (SAM) in mask-prompt mode was applied to improve mask precision during training data preparation. A YOLOv8s object detection model was trained on a custom bridge damage dataset of 6,531 images to detect two damage classes — crack and corrosion — achieving a mean Average Precision (mAP50) of 0.445 at a confidence threshold of 0.30. The framework associates detected damage with segmented structural members using a region-based spatial assignment strategy, enabling structure-aware outputs such as "crack on main girder" and "corrosion on deck slab."

 doi:10.20944/preprints202605.1350.v1

Manual evaluation on 100 bridge inspection images demonstrated a damage detection accuracy of 70.0% fully correct and 84.0% fully or partially correct, and a member assignment accuracy of 62.0% fully correct and 87.0% fully or partially correct. The main girder class achieved the highest combined accuracy for both damage detection (90.9%) and member assignment (93.9%). These results demonstrate the potential of the proposed framework for practical automated bridge inspection and infrastructure monitoring applications.

Keywords: bridge inspection; deep learning; structural member segmentation; damage detection; SegFormer; YOLOv8; Segment Anything Model; structure-aware analysis; semantic segmentation; spatial damage mapping; damage-to-member association

1. Introduction

Bridges are critical components of transportation infrastructure and play a fundamental role in supporting economic activity and social connectivity. However, many bridge systems worldwide are aging and increasingly require inspection, maintenance, and rehabilitation to ensure their structural safety and serviceability. In the United States, recent infrastructure assessments report that there are more than 623,000 bridges, with approximately 45% already exceeding their original 50-year design life, while more than 221,000 bridges require repair or replacement and over 63,000 are subject to load restrictions [1]. Similar concerns exist in Japan, where bridge deterioration has become a major challenge for infrastructure management. According to the Ministry of Land, Infrastructure, Transport and Tourism (MLIT), Japan has approximately 700,000 road bridges, of which about 70%–75% are managed by municipalities, and the proportion of bridges older than 50 years increased from 18% in 2013 to 43% in 2023 [2]. In addition, the number of municipal bridges with traffic restrictions has more than doubled in recent years, indicating that deterioration is already affecting serviceability and safety [2].

The challenge of aging bridges is further compounded by limitations in maintenance capacity and inspection resources. In Japan, many municipalities are responsible for a large number of bridges but often lack sufficient engineering personnel for bridge maintenance. MLIT data indicate that 50% of towns and 70% of villages do not have civil engineering technicians dedicated to bridge maintenance, while some municipal inspection practices still rely on distant visual inspection, which may lead to unreliable condition assessment [2]. Similar infrastructure management challenges are also observed in developing countries. In Sri Lanka, for example, more than 4,200 bridges are managed under the national road authority, and many older bridges have been identified as requiring rehabilitation, reconstruction, or load capacity enhancement [3]. These trends demonstrate that aging bridge infrastructure is not only a national issue but also a global engineering problem.

Traditionally, bridge condition assessment has relied primarily on manual visual inspection. Although visual inspection remains the most common method because of its simplicity and practical applicability, it is labor-intensive, costly, time-consuming, and strongly dependent on the experience and judgment of inspectors [4,5]. In some cases, bridge inspectors must access high or difficult-to-reach locations using rope access systems or high-elevation work platforms, which increases both cost and safety risk [5]. Moreover, inadequate inspection and condition assessment may lead to serious failures. For example, the collapse of the I-35W Highway Bridge in Minneapolis in 2007 resulted in 13 fatalities and 145 injuries, and investigation reports highlighted shortcomings in inspection guidance and condition assessment practices [4]. Such cases emphasize the importance of more reliable and efficient inspection technologies.

To address these limitations, computer vision, machine learning, and deep learning techniques have been increasingly investigated for bridge inspection and structural health monitoring. Early studies demonstrated that computer vision-based methods could support automated defect detection and condition assessment of civil infrastructure [4]. More recent advances in deep learning, especially convolutional neural networks (CNNs), have significantly improved automated crack detection and

 doi:10.20944/preprints202605.1350.v1

damage recognition performance [6,7]. In addition, the rapid development of unmanned aerial vehicles (UAVs) has enabled inspectors to collect high-resolution visual data from difficult-to-access bridge components, supporting safer and more efficient non-contact inspection workflows [5,8]. Recent research has further explored hybrid frameworks combining UAV imagery and deep learning models to improve structural health monitoring under real-world field conditions [8]. More recently, transformer-based segmentation architectures such as SegFormer [9] and foundation models such as the Segment Anything Model (SAM) [10] have opened new opportunities for precise structural component delineation and boundary refinement in complex visual scenes.

Despite these advancements, most existing studies still focus primarily on detecting damage types from images without sufficiently considering the structural member on which the damage occurs. In practical bridge maintenance, engineers must interpret not only the type of damage but also its structural location, since the same damage type may have different implications depending on whether it appears on a main girder, deck slab, abutment, or other bridge component [11]. Some recent studies have begun to address bridge components and damage jointly. For example, multilevel structural component detection and segmentation methods have been proposed for computer vision-based bridge inspection [12], and bridge damaged-object detection frameworks using bridge member models have also been reported [13]. In addition, recent multiclass bridge surface damage detection methods have demonstrated promising results for identifying and segmenting various damage categories across bridge components [11]. However, the integration of structural member segmentation, multiclass damage detection, and spatial damage-to-member association within a unified framework for bridge inspection remains limited.

Motivated by this gap, this study proposes a structure-aware deep learning framework for automated bridge inspection that integrates SegFormer-based structural member segmentation, YOLOv8-based multiclass damage detection, and a spatial damage mapping strategy. In the proposed framework, structural members are first identified using a transformer-based segmentation model refined with SAM-assisted mask preparation. Damage types are then detected from bridge inspection images using an efficient one-stage detection model, and the detected damage regions are spatially associated with the corresponding structural members. By combining these stages, the framework produces structure-aware outputs that provide more informative and actionable results for practical bridge condition assessment and maintenance decision-making.

The main contributions of this study are summarized as follows:

- Development of a structure-aware bridge inspection framework that integrates transformer-based structural member segmentation with one-stage damage detection and region-based spatial assignment, producing labeled outputs such as “crack on main girder” and “corrosion on deck slab” that are directly usable for maintenance prioritization — a capability not demonstrated by existing damage-only or segmentation-only approaches.
- Application of SAM mask-prompt refinement to improve ground-truth mask quality for structural member segmentation training, achieving a test mIoU of 0.851. Notably, SAM mask-prompt mode outperformed SAM bounding-box prompt mode (0.851 vs. 0.550), providing a practical guideline for applying foundation models to specialized structural inspection datasets.
- Implementation of a spatial damage mapping strategy that associates detected damage with specific structural members, enabling structure-aware inspection outputs.
- Systematic manual evaluation of the integrated pipeline on 100 real bridge inspection images with per-member-class accuracy breakdown, demonstrating damage detection accuracy of 70.0% fully correct (84.0% fully or partially correct) and member assignment accuracy of 62.0% fully correct (87.0% fully or partially correct), providing a replicable benchmark for structure-aware inspection frameworks.

 doi:10.20944/preprints202605.1350.v1

2. Literature Review

2.1. Bridge inspection and infrastructure deterioration

Bridges are essential components of transportation infrastructure and play a vital role in maintaining connectivity within transportation networks. However, many bridges worldwide are approaching or exceeding their design service life, which increases the risk of structural deterioration and potential safety issues. Structural deterioration may occur due to environmental exposure, repeated loading, corrosion, and natural hazards, making regular inspection and maintenance essential for ensuring structural reliability and public safety [1–3].

Traditionally, bridge inspection relies primarily on manual visual inspection performed by trained engineers. Although visual inspection is widely used due to its practicality, it is often labor-intensive, time-consuming, and highly dependent on the experience and judgment of inspectors. In many cases, inspectors must access hazardous locations such as high bridge decks or underside components using specialized equipment, which increases inspection cost and risk [4,5].

To address these limitations, researchers have explored advanced inspection technologies such as unmanned aerial vehicles (UAVs), remote sensing systems, and automated monitoring techniques. UAV-based bridge inspection systems enable efficient image acquisition from difficult-to-access locations and significantly improve inspection safety and efficiency. The integration of UAV platforms with artificial intelligence and remote sensing technologies has therefore been increasingly investigated for infrastructure inspection applications [5,8].

2.2. Deep learning for damage detection

Recent advances in deep learning have significantly improved automated damage detection in civil infrastructure systems. Convolutional neural networks (CNNs) have demonstrated strong capability in extracting complex image features and identifying structural defects such as cracks, corrosion, and surface deterioration in bridge components [6,15].

Earlier studies on crack detection relied mainly on traditional image processing techniques such as edge detection, thresholding, and morphological filtering. However, these methods often struggled to achieve reliable performance under varying lighting conditions, surface textures, and environmental noise [16].

With the development of deep learning methods, CNN-based models have shown improved performance in detecting cracks and other structural defects from image data. For example, hierarchical convolutional networks have been proposed to learn multi-scale crack features and accurately detect crack patterns in structural images [6]. Similarly, deep learning frameworks have been developed to detect multiple types of structural damage in infrastructure inspection images using advanced convolutional architectures [15,17]. More recent studies have further extended these approaches to multiclass bridge surface damage detection, demonstrating the ability to identify and localize multiple damage categories simultaneously across bridge components [11].

Despite these advances, many existing deep learning-based approaches focus mainly on identifying damage types within images without explicitly considering the structural component where the damage occurs. In practical bridge inspection, engineers must interpret both the type of damage and its structural location in order to determine appropriate maintenance strategies.

2.3. Structural member segmentation in infrastructure inspection

In addition to damage detection, identifying structural components within bridge images is important for accurate structural assessment. Bridge structures consist of multiple structural members such as girders, decks, piers, bearings, and bracing systems, and each component plays a different role in structural performance and maintenance planning.

Recent developments in computer vision have enabled automated structural component recognition using deep learning-based segmentation techniques. These approaches allow image

 doi:10.20944/preprints202605.1350.v1

analysis systems to identify and classify structural members within complex bridge scenes, enabling more detailed infrastructure inspection and analysis [12]. Furthermore, deep learning-based frameworks that combine structural member recognition with damage detection have been proposed to improve the practical applicability of automated inspection systems [13].

More recently, transformer-based architectures have demonstrated superior performance in semantic segmentation tasks compared to conventional CNN-based methods. SegFormer, proposed by Xie et al., combines a hierarchical transformer encoder with a lightweight multilayer perceptron decoder, enabling efficient multi-scale feature extraction without positional encoding, which allows the model to generalize across varying image resolutions [9]. These advances have motivated the application of transformer-based segmentation models to infrastructure inspection tasks, where accurate delineation of structural components from diverse viewpoints and imaging conditions is required.

In parallel, foundation models for image segmentation have emerged as powerful tools for boundary refinement and general-purpose segmentation. The Segment Anything Model (SAM), introduced by Kirillov et al., demonstrated strong zero-shot segmentation capability across diverse image domains through a promptable architecture that accepts point, box, and mask inputs [10]. Recent studies have explored the application of SAM in specialized inspection and remote sensing contexts, leveraging its boundary precision to refine coarse segmentation outputs. However, the integration of SAM with supervised structural segmentation models for bridge member delineation remains underexplored. In particular, no prior study has systematically evaluated different SAM prompting strategies — specifically bounding-box versus mask-prompt modes — in the context of structural member segmentation for bridge inspection, nor has the comparative effect of these prompting strategies on downstream damage-to-member association accuracy been investigated.

2.4. Research gap

Although significant progress has been made in automated bridge inspection using deep learning techniques, several limitations remain in existing research. Most current studies focus primarily on detecting damage types such as cracks or corrosion using image-based deep learning models. However, these approaches generally analyze damage independently of the structural component where it occurs.

In practical bridge inspection, engineers must evaluate both the type of damage and the structural member affected by the damage, as different structural components require different inspection criteria and maintenance strategies. For example, damage occurring on a primary load-carrying member may have different structural implications compared to similar damage on a secondary component.

Furthermore, existing segmentation approaches for bridge structural members often produce coarse pixel-level boundaries, particularly in regions with shadow, occlusion, or low texture contrast. Traditional semantic segmentation models trained on limited inspection datasets may struggle to precisely delineate structural member boundaries under such conditions. This limitation can reduce the reliability of downstream damage-to-member association, as imprecise masks may lead to incorrect structural assignments. A boundary refinement mechanism capable of improving mask precision without requiring additional manual annotation is therefore needed to support accurate structure-aware inspection.

Therefore, there is a clear need for an integrated framework that combines structural member segmentation, multiclass damage detection, and spatial damage-to-member association within a unified deep learning pipeline. Such a framework would allow automated inspection systems to produce structure-aware outputs — such as “crack on main girder” or “corrosion on deck slab” — that directly support maintenance prioritisation based on both damage type and structural location. Critically, no existing study has demonstrated such an end-to-end integrated pipeline for bridge inspection, nor systematically evaluated its member assignment accuracy across all primary structural member classes using real field inspection imagery [11–14].

 doi:10.20944/preprints202605.1350.v1

3. Methodology

Automated bridge inspection requires accurate identification of structural components and reliable detection of structural damage within complex inspection images. To address this challenge, this study proposes a deep learning-based framework that integrates structural member segmentation, damage detection, and structure-aware damage analysis within a unified inspection pipeline.

The overall workflow of the proposed framework is illustrated in Figure 1. The framework processes bridge inspection images through a sequence of stages designed to extract structural information and associate detected damage with specific bridge components.

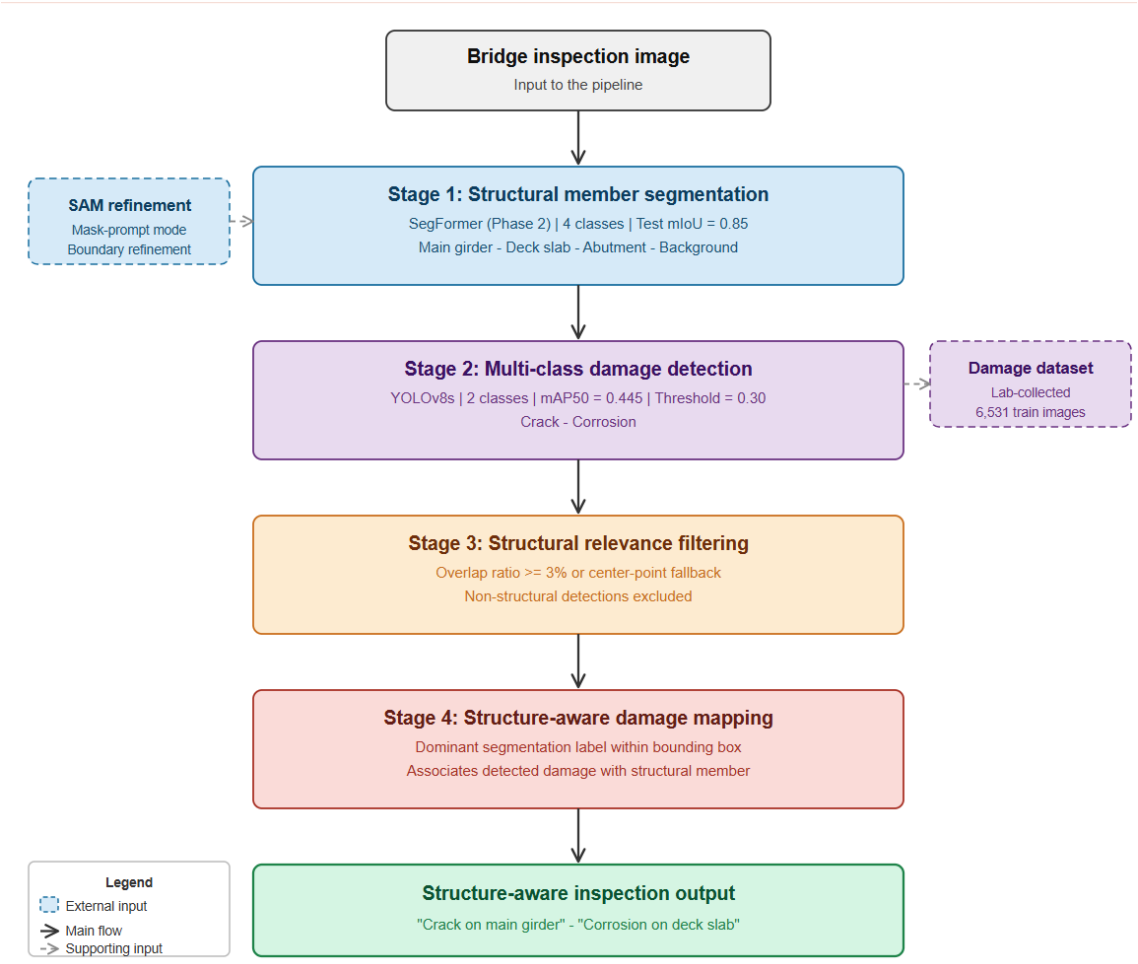


Figure 1. Overall workflow of the proposed bridge inspection framework integrating structural member segmentation, damage detection, and structure-aware damage mapping.

First, bridge inspection images are processed to identify the structural members present in the scene. A semantic segmentation model based on the SegFormer architecture is used to segment major bridge components such as main girders, deck slabs, and abutments. This stage provides the structural context of the bridge and enables the system to understand the spatial relationships between different structural elements.

Next, a damage detection model is applied to identify different types of structural damage within the inspection images. The detection stage locates potential damage regions, such as cracks and corrosion, using a deep learning-based object detection approach.

To ensure structural relevance, the detected damage regions are evaluated with respect to the segmented structural members. Detections that do not correspond to the target structural components are excluded from further analysis.

 doi:10.20944/preprints202605.1350.v1

Finally, the valid damage detections are associated with the segmented bridge structural members. This structure-aware damage mapping step allows the framework to determine which specific bridge component contains the detected damage, enabling more informative and practical inspection outputs for infrastructure monitoring. Based on this workflow, the proposed methodology is organized into four main stages:

1. **Structural Member Segmentation (SegFormer):** Extraction of bridge structural components from inspection images using a transformer-based semantic segmentation model. Major structural members, including main girders, deck slabs, and abutments, are identified to provide structural context.
2. **Multi-Class Damage Detection:** Detection of different types of structural damage present in bridge inspection images using a deep learning-based object detection model. Damage types considered include cracks and corrosion.
3. **Structural Relevance Filtering:** Filtering of detected damage regions based on their spatial correspondence with segmented structural members. Detections associated with non-structural elements are excluded to ensure structure-focused analysis.
4. **Structure-Aware Damage Mapping:** Association of validated damage detections with specific bridge structural members to support structure-based inspection and assessment.

The detailed methodology for each stage of the proposed framework is described in the following sections.

3.1. Structural Member Segmentation Using SegFormer

The first stage of the proposed framework focuses on identifying bridge structural members from inspection images using a semantic segmentation approach. Accurate segmentation of structural components is essential for providing structural context in subsequent damage detection and analysis stages.

In this study, a semantic segmentation model based on the SegFormer architecture [9] was employed to extract major bridge structural components from inspection images. SegFormer is a transformer-based segmentation model that combines hierarchical transformer encoders with a lightweight multilayer perceptron decoder, enabling efficient extraction of both local and global contextual features in complex visual scenes.

3.1.1. Bridge Inspection Image Dataset

Bridge inspection images used in this study were obtained from a custom dataset collected during routine bridge inspection activities. These images represent real inspection environments and include structural components such as main girders, deck slabs, abutments, and other bridge elements. Images were captured using standard inspection cameras, and the dataset reflects the visual diversity typically encountered in field inspection conditions.

To ensure high-quality annotations for segmentation tasks, images were carefully selected based on several criteria. Images containing people, ultraviolet inspection markings, or excessive darkness were excluded to avoid interference with the annotation process. Images with clear visibility of structural members were prioritized in order to facilitate accurate pixel-level labeling. Using this selection process, a total of 804 bridge inspection images were prepared for annotation.

3.1.2. Pixel-Level Annotation and SAM-Assisted Boundary Refinement

To generate ground-truth segmentation masks, the selected bridge inspection images were manually annotated using the Computer Vision Annotation Tool (CVAT). CVAT is an open-source annotation platform widely used for preparing datasets for computer vision and deep learning applications. The tool provides polygon-based annotation capabilities that allow precise delineation of object boundaries within images.

 doi:10.20944/preprints202605.1350.v1

During the annotation process, each image was labeled with polygon masks corresponding to bridge structural components and surrounding contextual elements. Among the 804 images prepared for annotation, 743 images contained identifiable structural components and were successfully annotated, while 61 images were excluded due to insufficient structural information. The annotation process produced a total of 2,036 segmentation masks across multiple structural and non-structural classes. Structural classes included STR_main_girder, STR_deck_slab, STR_abutment, and STR_cross_beam. In addition, contextual elements such as guard rails, drainage pipes, and parapets were annotated as non-structural classes.

To further improve mask boundary quality, the Segment Anything Model (SAM) was applied as a post-processing refinement step on the initial segmentation masks. SAM was used in mask-prompt mode, where the coarse segmentation predictions served as spatial priors to guide SAM in generating refined pixel-level boundaries. This hybrid approach enabled more precise delineation of structural member boundaries, particularly for large continuous members such as main girders and deck slabs, without requiring additional manual re-annotation. The refined masks were used as the final ground-truth annotations for model training.

3.1.3. Dataset Integration and Preparation

To improve training robustness and increase the diversity of structural samples, the annotated dataset was integrated with an additional bridge member mask dataset containing previously generated segmentation masks. Before integration, class labels of the two datasets were harmonized to ensure consistency with the structural annotation scheme adopted in this study. Only classes relevant to bridge structural components were retained for segmentation training. The integrated dataset contains segmentation masks for three primary bridge structural members: STR_main_girder, STR_deck_slab, and STR_abutment.

The combined dataset was divided into training, validation, and testing subsets to enable reliable model training and evaluation. The primary dataset was split using a 70%–15%–15% ratio, resulting in 520 training images, 111 validation images, and 112 test images. The supplementary dataset followed an 80%–10%–10% split, producing 476 training images, 59 validation images, and 61 test images. After integration, the final dataset used for segmentation training consisted of 1,339 bridge inspection images, comprising 996 training images, 170 validation images, and 173 test images.

3.1.4. SegFormer Model Configuration and Training

The SegFormer model training was conducted in two phases. In Phase 1, five classes were used, including background, STR_main_girder, STR_deck_slab, STR_abutment, and STR_bearing. However, the bearing class consistently achieved zero IoU throughout training due to its small spatial extent and limited representation in the dataset. Based on this observation, the bearing class was excluded in Phase 2 to reduce class imbalance and improve overall training stability.

In Phase 2, the model was trained on four classes: background, STR_main_girder, STR_deck_slab, and STR_abutment. This configuration enables the model to focus on the primary load-bearing structural components most relevant to bridge inspection and structural condition assessment. By leveraging the transformer-based architecture of SegFormer, the model effectively captures spatial relationships between structural members and surrounding regions in complex bridge inspection scenes.

The final Phase 2 model achieved a test mean Intersection over Union (mIoU) of 0.85. Per-class IoU values on the test set were 0.92 for background, 0.84 for STR_main_girder, 0.80 for STR_deck_slab, and 0.84 for STR_abutment. Table 1 summarizes the segmentation performance across the two training phases and SAM refinement configurations evaluated during model development.

 doi:10.20944/preprints202605.1350.v1

Table 1. SegFormer segmentation performance across training phases.

Phase	Configuration	mIoU	Background	Main girder	Deck slab	Abutment
Phase 1	SegFormer only	0.611	0.869	0.757	0.686	0.741
Phase 1	+ SAM bbox refinement	0.550	0.789	0.698	0.644	0.617
Phase 1	+ SAM mask refinement	0.614	0.869	0.762	0.694	0.743
Phase 2	**SAM mask refinement (bearing removed)**	**0.851**	**0.917**	**0.843**	**0.802**	**0.843**

A stability check evaluating the model across training, validation, and test splits confirmed consistent generalization behavior, with training mIoU of 0.93 and validation and test mIoU values of 0.78 and 0.85 respectively. The close agreement between validation and test performance indicates no evidence of data leakage or overfitting. The segmentation outputs generated by this stage provide the structural context required for the subsequent damage detection and structure-aware damage analysis stages of the proposed framework.

3.2. Multi-Class Bridge Damage Detection

The second stage of the proposed framework focuses on identifying structural damage within bridge inspection images using an object detection approach. Accurate detection of damage regions is essential for assessing structural condition and supporting maintenance decision-making. In this study, a deep learning-based object detection model based on the YOLOv8 architecture [18] was employed to detect and localize structural damage. YOLOv8 is a one-stage detection framework that performs object localization and classification simultaneously, enabling efficient and real-time performance. Its capability to process entire images in a single forward pass makes it well suited for large-scale bridge inspection applications.

3.2.1. Damage Detection Dataset

The damage detection model was trained using a custom bridge damage dataset collected under real inspection conditions, containing annotated images of bridge surfaces exhibiting various types of deterioration. Three model configurations were evaluated during development: a three-class baseline model detecting crack, corrosion, and leakage; a three-class model with augmented leakage training data; and a final simplified two-class model detecting crack and corrosion only.

The leakage class consistently showed the weakest detection performance across all configurations. Even after targeted data augmentation that expanded leakage training samples from 321 to approximately 1,000 images, the leakage mAP50 remained low at 0.145, compared to 0.313 in the baseline. This result indicates that the detection difficulty for leakage is primarily due to intrinsic visual variability — including irregular boundaries, low contrast, and high similarity to background staining — rather than insufficient training data. Based on this analysis, the leakage class was excluded from the final model configuration to improve detection reliability and training stability.

The final two-class dataset was organized into training, validation, and testing subsets as follows: 6,531 training images, 1,740 validation images, and 871 test images. This dataset provides diverse examples of structural damage under varying lighting conditions, viewpoints, and surface textures, enabling robust model training.

3.2.2. Annotation Format and Data Preparation

Damage annotations were provided in YOLO format, where each image is associated with a corresponding text file containing normalized bounding box coordinates. Each annotation specifies the damage class and the spatial location of the damage region. Prior to training, the dataset was verified to ensure consistency between images and labels. Corrupted samples and mismatched annotations were removed to improve data quality and training reliability.

 doi:10.20944/preprints202605.1350.v1

3.2.3. YOLOv8 Model Configuration and Training

The YOLOv8s model was selected as the base architecture for damage detection due to its balance between accuracy and computational efficiency. The model was trained using the following configuration: input image size of 640 × 640 pixels, batch size of 8, 40 training epochs, and 2 data loading workers. The model was optimized to detect crack and corrosion instances by learning both spatial and visual characteristics of damage patterns. During training, the model minimizes localization and classification errors to improve detection accuracy.

3.2.4. Damage Detection Performance

The final two-class YOLOv8s model achieved a Precision of 0.536, Recall of 0.427, mAP50 of 0.445, and mAP50–95 of 0.235 on the test dataset. The F1 score, computed as the harmonic mean of Precision and Recall, was 0.474 overall (crack: 0.523; corrosion: 0.387), reflecting moderate detection balance between the two classes. Class-wise mAP50 was 0.552 for crack and 0.339 for corrosion. Table 2 summarizes the comparative performance across the three model configurations evaluated during development.

Table 2. Comparison of damage detection model configurations.

Model	Classes	Precision	Recall	mAP50	mAP50-95
3-class baseline	Crack, Corrosion, Leakage	0.482	0.423	0.409	0.212
3-class + augmented leakage	Crack, Corrosion, Leakage	0.503	0.400	0.397	0.204
2-class (final)	**Crack, Corrosion**	**0.536**	**0.427**	**0.445**	**0.235**

The removal of the leakage class improved overall Precision and mAP50, while crack detection performance remained stable across configurations (mAP50 of 0.556 in the baseline versus 0.552 in the final model). A confidence threshold of 0.30 was selected for inference based on qualitative comparison at thresholds of 0.15, 0.30, and 0.50. A threshold of 0.15 produced excessive false positives, while 0.50 resulted in missed detections of smaller or lower-contrast damage regions. A threshold of 0.30 provided the best balance between detection sensitivity and reliability and was therefore adopted for all subsequent analysis.

3.2.5. Damage Detection Output

The trained model outputs bounding boxes, class labels, and confidence scores for each detected damage instance. These outputs provide spatial localization of damage regions within bridge inspection images. Although bounding boxes represent approximate damage regions rather than precise pixel-level boundaries, they provide sufficient spatial information for subsequent structure-aware analysis. The detected damage regions are further processed in the following stages to ensure structural relevance and enable mapping to specific bridge components.

3.3. Structural Relevance Filtering

The detected damage regions are further processed to ensure their relevance to structural components. Since object detection models may identify damage on both structural and non-structural elements, a filtering mechanism is introduced to retain only structurally meaningful detections. In this study, the spatial relationship between detected damage regions and segmented structural members is evaluated. Damage detections that do not sufficiently correspond to the target structural components — main girders, deck slabs, and abutments — are classified as non-structural and excluded from further analysis. This filtering step improves the reliability of the proposed framework by ensuring that only damage associated with key load-bearing structural members is considered in the subsequent stage. This step is particularly important given that the YOLOv8s model

 doi:10.20944/preprints202605.1350.v1

may detect damage-like patterns on non-structural elements such as signage, vegetation, or equipment visible in field inspection images. The structural relevance filter acts as a gating mechanism that improves the precision of structure-aware outputs by suppressing detections that, while potentially valid as damage instances, are not associated with the primary load-bearing members under inspection.

3.4. Structure-Aware Damage Analysis

The final stage of the proposed framework integrates the outputs from structural member segmentation and damage detection to perform structure-aware damage analysis. This step enables the association of detected damage with specific bridge structural components, thereby improving the interpretability and practical relevance of the inspection results.

For each detected damage instance, represented by a bounding box obtained from the YOLOv8 model (Section 3.2), its spatial relationship with the segmented structural member masks produced by the SegFormer model (Section 3.1) is evaluated using a two-step spatial assignment strategy.

In the first step, the distribution of segmentation labels within the damage bounding box region is analyzed. The proportion of pixels belonging to each structural class within the bounding box is computed, and if the dominant structural class occupies at least 3% of the bounding box area, that class is assigned as the corresponding structural member for the detected damage. This region-based approach is conceptually related to overlap-based methods such as Intersection over Union (IoU), but is specifically adapted for the integration of object detection bounding boxes with dense semantic segmentation masks.

In cases where no structural class meets the minimum overlap threshold, a center-point fallback strategy is applied. The structural label at the geometric center of the damage bounding box is retrieved from the segmentation mask and used for member assignment. This fallback improves robustness in cases where the damage region spans multiple structural and non-structural areas, or where the segmented member occupies only a small portion of the bounding box.

Detections for which neither the overlap criterion nor the center-point fallback yields a valid structural member assignment are classified as belonging to an unknown or non-structural region and excluded from the final inspection output.

This process enables the generation of structure-aware inspection results, such as “crack on main girder” or “corrosion on deck slab”, which provide more meaningful and actionable information compared to conventional damage detection approaches that report damage type alone without considering structural context.

4. Results

4.1. Structural Member Segmentation Results

The SegFormer-based structural member segmentation model was evaluated across two training phases to identify the optimal configuration for the proposed framework. Table 1 summarizes the segmentation performance across all evaluated configurations.

In Phase 1, the model was trained on five classes including STR_bearing. The baseline SegFormer model achieved a test mIoU of 0.611. Application of SAM in bounding-box prompt mode decreased overall performance to 0.550, indicating that bounding-box prompts alone were insufficient to capture precise structural boundaries. SAM mask-prompt refinement partially recovered performance to 0.614; however, the STR_bearing class consistently achieved an IoU of 0.000 across all Phase 1 configurations due to its small spatial extent and limited representation in the training data.

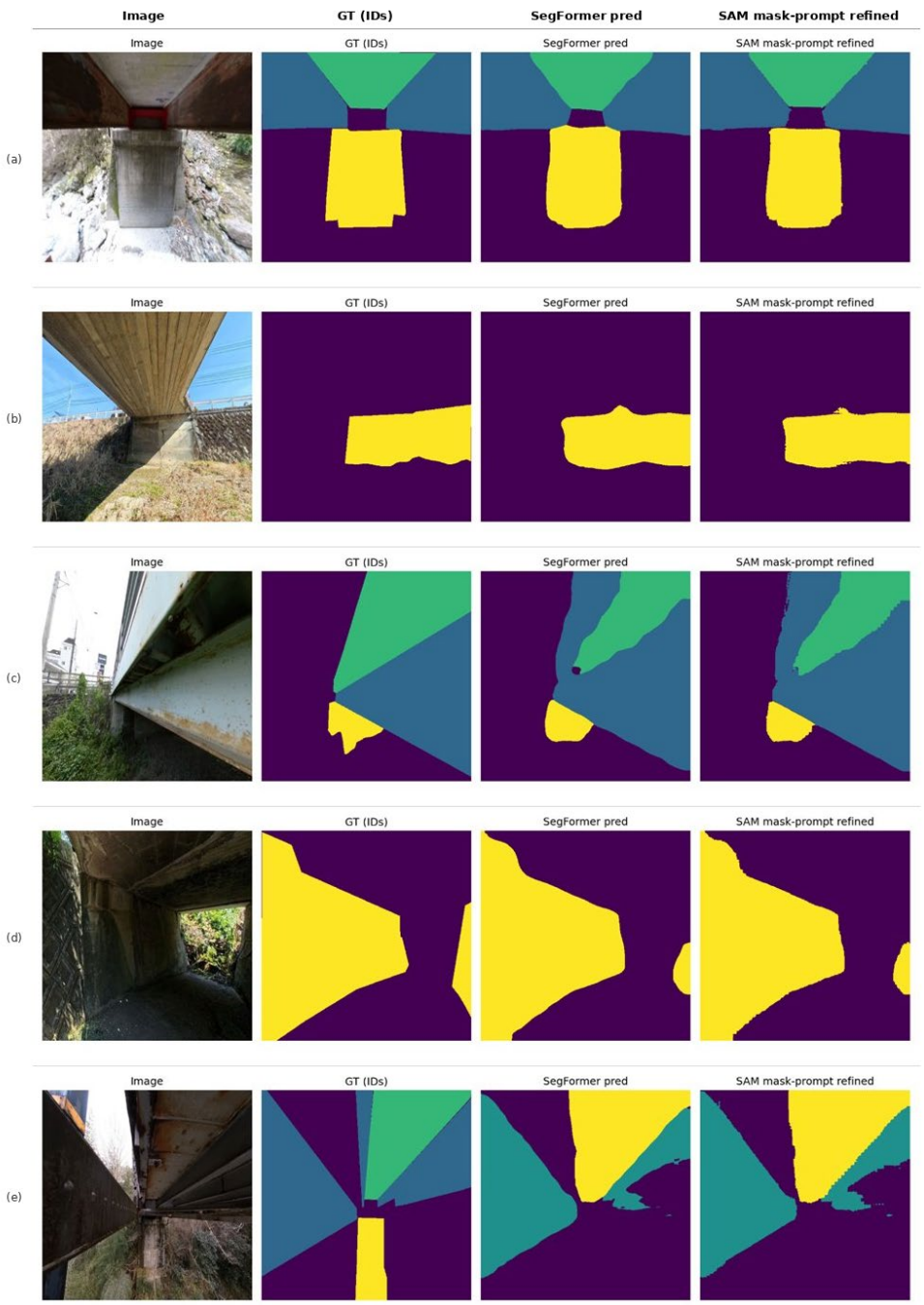
Based on these observations, the bearing class was excluded in Phase 2, and the model was retrained using SAM mask-prompt refined annotations on four classes. This configuration produced substantially improved results, achieving a validation mIoU of 0.822 and a test mIoU of 0.851. Per-class test IoU values were 0.917 for background, 0.843 for STR_main_girder, 0.802 for STR_deck_slab, and 0.843 for STR_abutment. The significant improvement from Phase 1 (0.611) to Phase 2 (0.851)

 doi:10.20944/preprints202605.1350.v1

demonstrates the combined benefit of removing the problematic bearing class and applying SAM mask-prompt boundary refinement during training data preparation.

The comparative segmentation performance across all evaluated configurations is summarized in Table 1 (Section 3.1.4). A stability check evaluating train, validation, and test performance confirmed consistent generalization behavior, with training mIoU of 0.931, validation mIoU of 0.784, and test mIoU of 0.851. The close agreement between validation and test mIoU indicates no evidence of data leakage or overfitting, and the train-validation gap is expected given the smaller held-out set size and higher structural and viewpoint variability in the held-out data.

Figure 2 illustrates representative qualitative results from the Phase 2 segmentation model, showing the original bridge inspection image, the ground-truth mask, the SegFormer prediction, and the SAM mask-prompt refined output for selected test samples. The model correctly delineates major structural members including main girders, deck slabs, and abutments across diverse viewpoints and imaging conditions. Minor discrepancies are primarily observed at structural boundaries and in shadowed or occluded regions, which is consistent with the quantitative results.



 doi:10.20944/preprints202605.1350.v1

Figure 2. Representative segmentation results from the Phase 2 SegFormer model on test images. Each row shows the original bridge inspection image, ground-truth annotation (GT), SegFormer prediction, and SAM mask-prompt refined output. Colors represent: purple = background, yellow = STR_main_girder, teal = STR_deck_slab, green = STR_abutment.

4.2. Damage Detection Results

The YOLOv8s damage detection model was evaluated across three configurations to determine the optimal class structure and training strategy. The comparative results are summarized in Table 2.

The three-class baseline model, trained to detect crack, corrosion, and leakage, achieved a Precision of 0.482, Recall of 0.423, and mAP50 of 0.409. Class-wise evaluation showed strong performance for crack detection (mAP50 = 0.556), moderate performance for corrosion (mAP50 = 0.358), and consistently weak performance for leakage (mAP50 = 0.313). To address leakage detection limitations, a targeted augmentation strategy was applied, expanding leakage training samples from 321 to approximately 1,000 images through brightness adjustment, slight rotation, blur, and mild scaling. Despite this augmentation, leakage detection performance further decreased to an mAP50 of 0.145, confirming that the difficulty is attributable to the intrinsic visual complexity of leakage patterns — including irregular boundaries, diffuse staining, and high visual similarity to background surfaces — rather than insufficient training data.

Based on this analysis, the leakage class was removed and a simplified two-class model was trained on crack and corrosion only. The final model achieved a Precision of 0.536, Recall of 0.427, mAP50 of 0.445, and mAP50-95 of 0.235. The corresponding F1 score was 0.474 overall (crack: 0.523; corrosion: 0.387). Class-wise mAP50 was 0.552 for crack and 0.339 for corrosion. The improvement in overall precision and mAP50 compared to the three-class baseline confirms that removing the problematic leakage class reduced class imbalance and improved detection stability.

The comparative damage detection performance across the three model configurations is summarized in Table 2. A qualitative evaluation of detection outputs under confidence thresholds of 0.15, 0.30, and 0.50 revealed that a threshold of 0.30 provided the best balance between detection sensitivity and false positive suppression. A threshold of 0.15 produced excessive false positives, while a threshold of 0.50 resulted in missed detections, particularly for smaller or lower-contrast damage instances. Based on this analysis, a confidence threshold of 0.30 was adopted for all subsequent inference.

Figure 3 shows representative detection outputs from the final two-class model under different confidence thresholds across multiple bridge inspection images, illustrating the effect of threshold selection on detection behavior.

4.3. Structure-Aware Damage Mapping Results

The final stage of the proposed framework evaluated the ability of the integrated pipeline to associate detected damage with specific bridge structural members. The spatial assignment strategy, combining region-based overlap analysis with a center-point fallback mechanism, was applied to bridge inspection images using the Phase 2 SegFormer segmentation model and the final two-class YOLOv8s detection model.

The integrated pipeline was evaluated on a manually reviewed sample of 100 bridge inspection images selected proportionally across all structure-aware label categories. For each image, the predicted structural member and damage type were verified through visual inspection of the pipeline output. Overall, the pipeline achieved a damage detection accuracy of 70.0% (fully correct) and 84.0% (fully or partially correct), and a member assignment accuracy of 62.0% (fully correct) and 87.0% (fully or partially correct). Performance varied across member classes — the main girder class achieved the highest combined accuracy for both damage detection (90.9%) and member assignment (93.9%), while the deck slab class showed the lowest performance (75.0% and 78.1% respectively). The lower deck slab performance is attributed to the frequent presence of secondary bridge

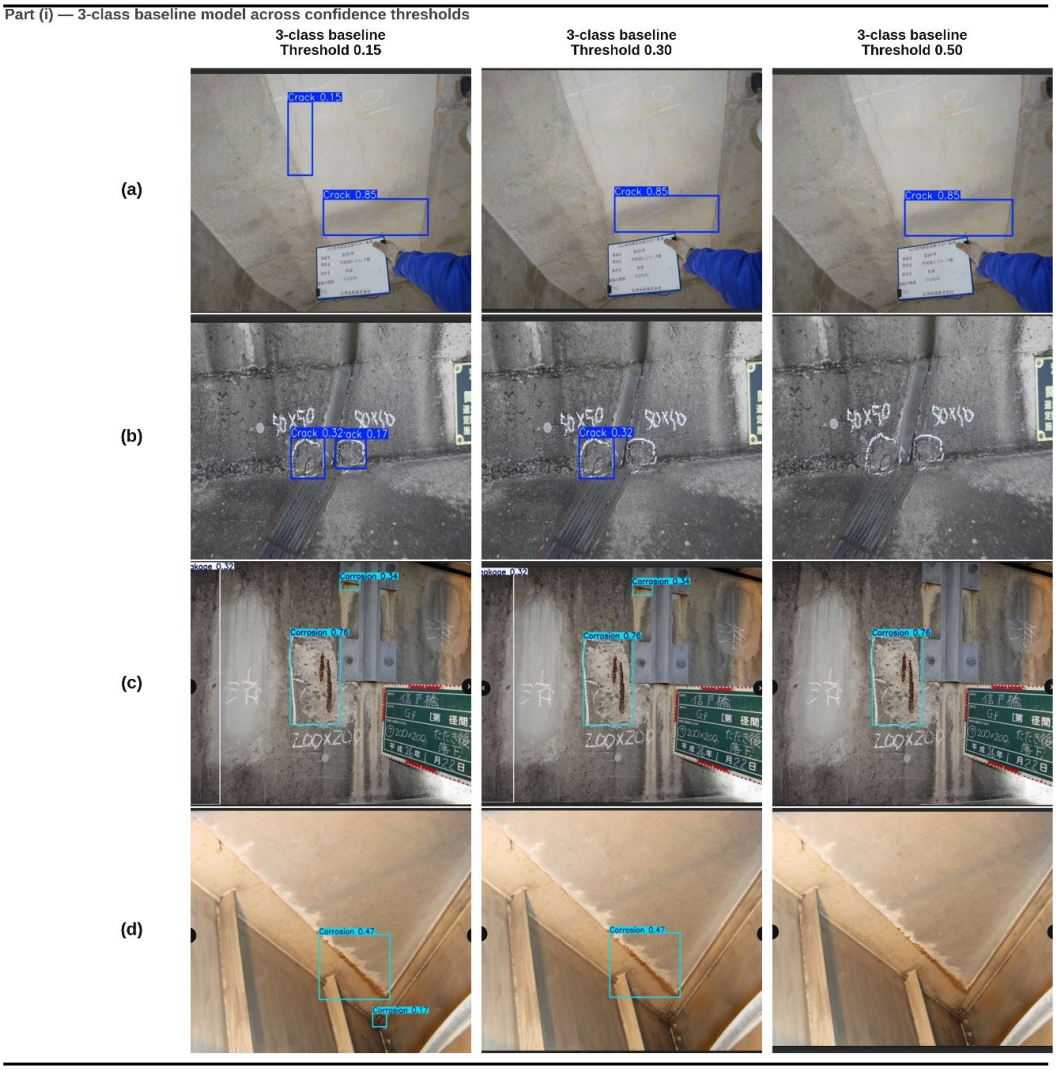
 doi:10.20944/preprints202605.1350.v1

components such as drainage pipes, railings, bearing assemblies, and parapet walls in the foreground of deck slab images, which the segmentation model tends to classify as deck slab due to background dominance. To address this limitation, future work will explore multi-scale segmentation refinement and dedicated negative-sample augmentation strategies for deck slab regions to improve the model's ability to distinguish structural slab surfaces from foreground attachments. These results demonstrate that the proposed framework provides useful structure-aware inspection outputs across all three structural member classes, with particularly strong performance for main girder damage scenarios. The results are summarized in Table 3.

Table 3. Manual evaluation of the structure-aware integrated pipeline on 100 bridge inspection images. Damage Correct and Member Correct refer to the proportion of images in which the predicted damage type and structural member assignment were judged fully correct, partially correct, or incorrect upon visual inspection.

Member Class	n	Damage Correct (%)	Damage Partial (%)	Damage Incorrect (%)	Member Correct (%)	Member Partial (%)	Member Incorrect (%)
Main girder	33	75.8	15.2	9.1	63.6	30.3	6.1
Deck slab	32	62.5	12.5	25.0	53.1	25.0	21.9
Abutment	35	71.4	14.3	14.3	68.6	20.0	11.4
Overall	**100**	**70.0**	**14.0**	**16.0**	**62.0**	**25.0**	**13.0**

Yes+Partial combined accuracy: Damage = 84%; Member = 87%.



 doi:10.20944/preprints202605.1350.v1

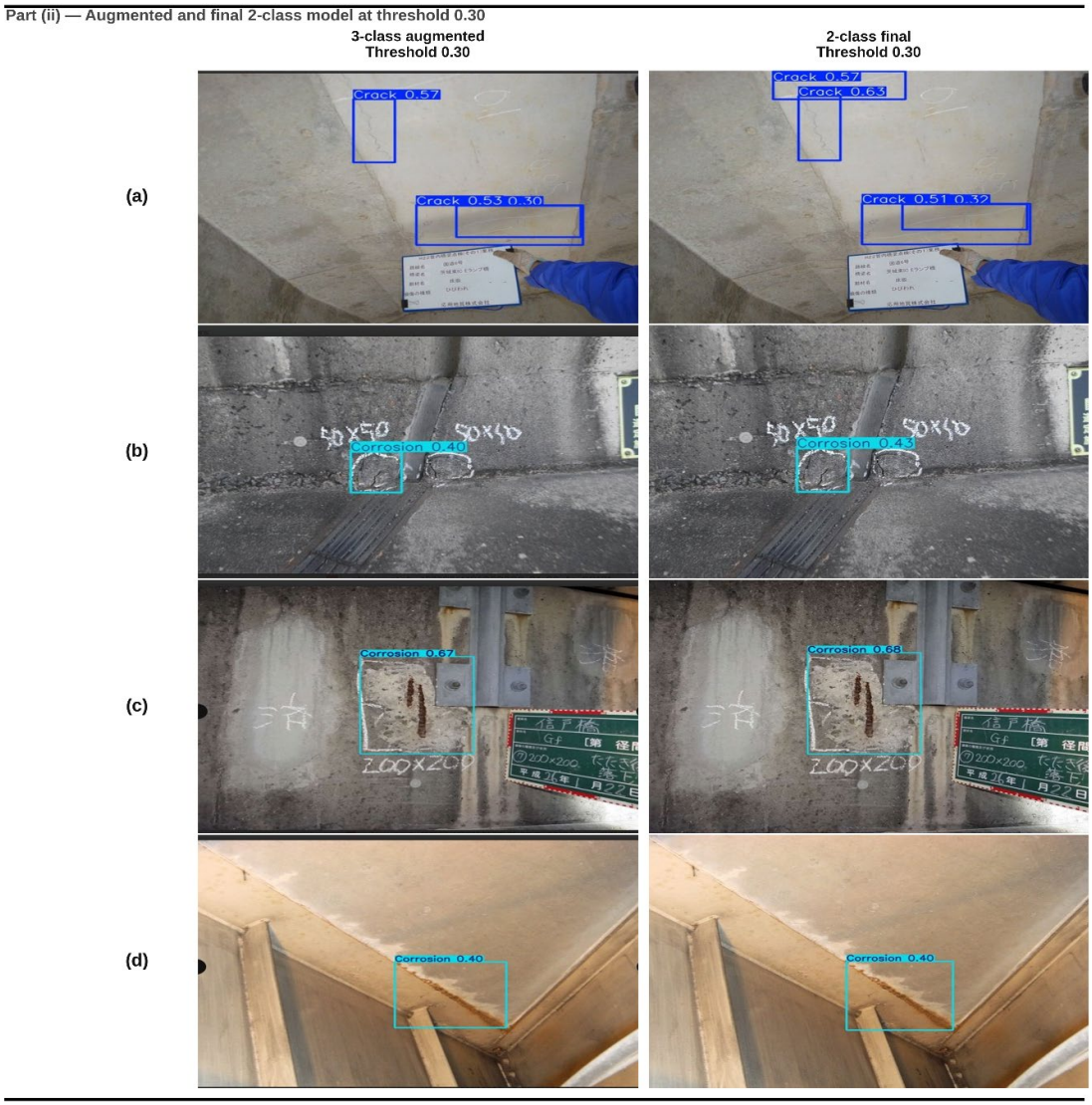


Figure 3. Qualitative comparison of damage detection outputs across model configurations and confidence thresholds. Part (i) — Columns 1–3 show the three-class baseline model at thresholds of 0.15, 0.30, and 0.50. Part (ii) — Column 4 shows the three-class model with augmented leakage data at threshold 0.30; Column 5 shows the final two-class model at threshold 0.30. Row labels: (a) crack detection on concrete surface; (b) small crack detection; (c) corrosion detection on structural components; (d) corrosion detection in bearing area. Blue bounding boxes indicate crack detections; cyan bounding boxes indicate corrosion detections.

Figure 4 illustrates representative structure-aware inspection outputs from the integrated pipeline across all three structural member classes and both damage types. Panels (a) and (b) demonstrate corrosion detection correctly assigned to the main girder class, which achieved the highest combined accuracy in the evaluation. Panels (c) and (d) show correct member assignment for deck slab and abutment corrosion cases respectively, with panel (d) achieving the highest single detection confidence of 0.86 in the illustrated examples. Panels (e), (f), and (g) demonstrate crack detection across all three member classes, with panel (e) showing multiple crack detections correctly assigned to the same abutment member. Panel (h) illustrates the framework's composite damage labeling capability, where both corrosion and crack are simultaneously detected on a single abutment member, generating a structure-aware output of "Corrosion + Crack on abutment." These outputs confirm that the proposed framework successfully associates detected damage with specific structural members across diverse bridge inspection imaging conditions, viewpoints, and structural configurations, providing actionable inspection information that extends beyond simple damage type classification.

 doi:10.20944/preprints202605.1350.v1

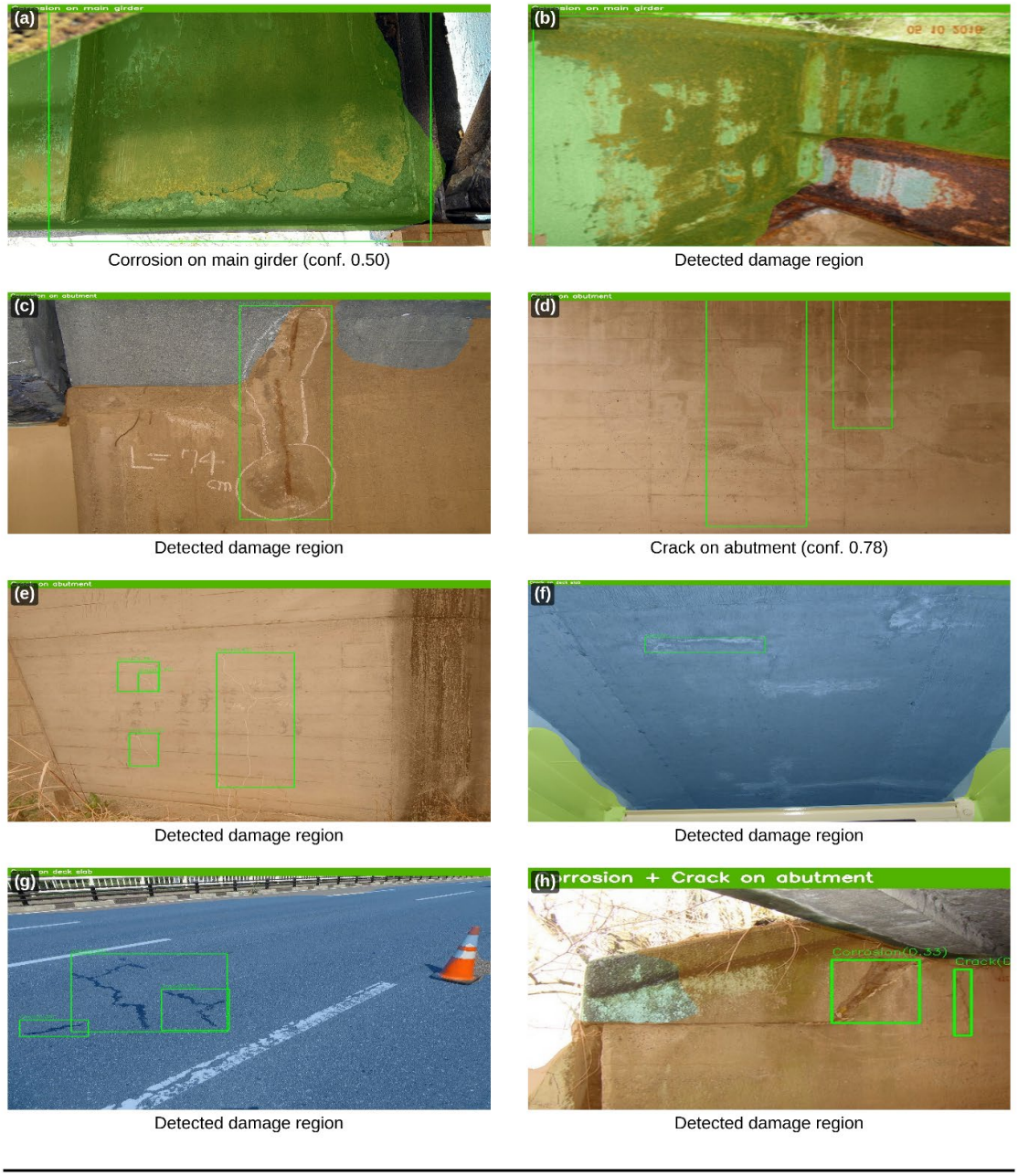


Figure 4. Representative structure-aware inspection outputs from the integrated pipeline across all three structural member classes and both damage types. Green bounding boxes indicate detected damage regions with predicted structural member labels and confidence scores. Panel labels: (a) corrosion on main girder (confidence 0.50) — single corrosion detection correctly assigned to steel main girder; (b) corrosion on main girder (confidence 0.73) — multiple corrosion detections on steel girder surface; (c) corrosion on deck slab (confidence 0.44) — corrosion correctly assigned to deck slab member; (d) corrosion on abutment (confidence 0.86) — high-confidence single detection correctly assigned to abutment; (e) crack on abutment (confidence 0.78, 0.51) — multiple crack detections correctly assigned to the same abutment member; (f) crack on deck slab (confidence 0.76) — multiple high-confidence crack detections on deck slab; (g) crack on main girder (confidence 0.62) — crack detection correctly assigned to main girder; (h) corrosion and crack on abutment — composite damage detection demonstrating the framework's multi-class output capability on a single structural member.

 doi:10.20944/preprints202605.1350.v1

5. Discussion

5.1. Segmentation performance and the role of SAM refinement

The Phase 2 SegFormer model achieved a test mIoU of 0.851, representing a substantial improvement over the Phase 1 baseline of 0.611 [12]. This improvement was driven by two factors: the removal of the bearing class, which introduced class imbalance and zero-IoU noise throughout Phase 1 training, and the application of SAM mask-prompt refinement to improve boundary quality in the training annotations. These results are consistent with findings reported by Yu and Nishio [12], who demonstrated that accurate structural component segmentation requires sufficient class representation and precise pixel-level annotations. The substantial performance gain achieved through SAM-assisted annotation preparation highlights the value of foundation model integration in specialized dataset construction, as previously noted in studies applying SAM to domain-specific segmentation tasks [10].

The finding that SAM bounding-box prompting decreased performance (from 0.611 to 0.550) while mask-prompt refinement maintained or improved it is noteworthy. Bounding-box prompts provide only coarse spatial constraints and may produce ambiguous masks in complex structural scenes [10]. In contrast, mask-prompt mode uses the SegFormer coarse prediction as a spatial prior, guiding SAM to refine boundaries within already-identified regions. This hybrid approach leverages the complementary strengths of supervised semantic segmentation and foundation-model-based promptable segmentation [9,10], and is consistent with the broader trend of combining task-specific models with general-purpose foundation models in specialized domains. To the best of our knowledge, this is the first study to systematically compare SAM bounding-box and mask-prompt modes for structural member segmentation in bridge inspection, demonstrating that mask-prompt mode using coarse supervised predictions as spatial priors outperforms bounding-box prompting in this domain. This finding provides a practical guideline for researchers applying foundation models to specialized infrastructure inspection datasets where precise boundary delineation is critical and manual re-annotation is costly.

The deck slab class showed the largest improvement from Phase 1 to Phase 2, with IoU increasing from 0.686 to 0.802. This is likely attributable to deck slabs typically occupying large, continuous image regions that benefit most from precise boundary delineation. Main girder and abutment classes also showed consistent improvements. These findings align with observations in previous structural segmentation studies, where larger structural members with more consistent visual appearance tend to achieve higher segmentation accuracy [12].

5.2. Damage detection performance and class complexity

The damage detection results reveal an important trade-off between model completeness and detection reliability in multi-class bridge damage detection. The leakage class consistently demonstrated the weakest performance across all evaluated configurations, even after targeted augmentation. This behavior is consistent with previous studies on bridge surface damage detection, which have noted that diffuse and visually variable damage patterns — including water-related staining and leakage — are inherently more difficult to detect using bounding-box-based approaches compared to structurally compact damage types such as cracks [11].

The decision to adopt a two-class model reflects a practical engineering judgment supported by the experimental evidence: a reliable model detecting two damage types provides more consistent and actionable outputs than an unreliable three-class model. The final two-class YOLOv8s model achieved a crack mAP50 of 0.552 and an overall mAP50 of 0.445. While these values are modest compared to enhanced YOLOv8 variants specifically designed for bridge defect detection — such as BD-YOLOv8s, which reported an mAP50 of 0.862 on a dedicated bridge defect dataset [19] — direct numerical comparison is not meaningful, as BD-YOLOv8s and similar enhanced architectures were evaluated on purpose-built, curated benchmark datasets under controlled conditions. The current dataset reflects real-world inspection variability including mixed viewpoints, partial occlusion, and

 doi:10.20944/preprints202605.1350.v1

inconsistent lighting, which inherently reduces detection scores. The reported mAP50 of 0.445 is therefore consistent with multi-class damage detection performance on field inspection imagery [11,20]. Furthermore, the primary contribution of the present framework lies not in detection performance alone, but in the integration of detection with structural member segmentation to produce structure-aware outputs that extend beyond the capabilities of detection-only approaches.

A confidence threshold of 0.30 was identified as optimal through qualitative threshold analysis. This finding is consistent with the general practice in object detection for structural inspection, where lower confidence thresholds are preferred to maximize detection recall at the cost of some precision, particularly when the downstream task involves further filtering — as in the structural relevance filtering stage of the proposed framework [18].

5.3. Structure-aware damage mapping and framework integration

The integrated pipeline demonstrated the ability to generate structure-aware inspection outputs, correctly associating detected damage with structural members in 62.0% of evaluated cases (fully correct) and 87.0% (fully or partially correct) across 100 bridge inspection images. Damage detection accuracy reached 70.0% fully correct and 84.0% fully or partially correct. These results demonstrate the practical feasibility of linking damage detection outputs with structural segmentation masks for automated bridge inspection, extending beyond the capabilities of existing approaches that report damage type alone without structural context [11–13]. The per-member-class breakdown in Table 3 reveals that the main girder class achieves the strongest performance, while deck slab presents the greatest challenge due to foreground occlusion by secondary components — a finding that motivates targeted improvements in future work.

Member assignment errors were primarily attributable to two factors: viewpoint-induced ambiguity and domain shift between training datasets. These challenges are well-documented in the bridge inspection literature, where image variability across inspection conditions, viewpoints, and equipment types frequently degrades model generalization [4,5]. The use of a center-point fallback strategy partially mitigated these effects by providing a secondary assignment mechanism when region-based overlap was insufficient.

5.4. Practical implications and limitations

The proposed framework addresses a genuine gap in existing bridge inspection research by providing structure-aware outputs that directly support maintenance decision-making. The ability to generate outputs such as “crack on main girder” or “corrosion on deck slab” — rather than simply “crack detected” — provides inspectors with immediately actionable information, consistent with the practical requirements identified in bridge maintenance guidelines [2,4]. This capability represents a meaningful step toward AI-assisted inspection systems that can support engineers in prioritizing maintenance interventions based on both damage type and structural location.

However, several limitations should be acknowledged. The current framework detects only two damage types, and extending coverage to additional categories such as spalling, delamination, and leakage would increase its practical scope, as these damage types are commonly observed in real bridge inspections [11,15]. Although leakage detection proved challenging in the current study due to its intrinsic visual variability, future work targeting improved leakage-specific architectures and annotation strategies, alongside expansion to spalling and delamination, would increase practical scope. Furthermore, while the current evaluation on 100 images provides a meaningful initial benchmark, a larger-scale quantitative evaluation on a fully held-out test set is needed to more comprehensively characterize framework performance across diverse bridge types and inspection conditions. Furthermore, performance degradation under domain shift conditions highlights the importance of collecting diverse multi-domain training data, as noted in recent structural health monitoring reviews [14].

 doi:10.20944/preprints202605.1350.v1

6. Conclusions

This study demonstrated that integrating transformer-based structural member segmentation, one-stage object detection, and spatial damage mapping within a unified deep learning pipeline can produce structure-aware bridge inspection outputs with 70.0% fully correct and 84.0% partially-or-fully correct damage identification, and 62.0% fully correct and 87.0% partially-or-fully correct member assignment, on 100 real bridge inspection images. These results address a key limitation of existing bridge inspection approaches, which report damage type without structural context, by explicitly linking detected damage to specific structural components and thereby providing more actionable information for maintenance decision-making. The following key findings were established.

The following conclusions are drawn from the experimental results:

The SegFormer-based structural member segmentation model, trained using SAM mask-prompt refined annotations, achieved a test mIoU of 0.851 on three primary structural member classes — main girder, deck slab, and abutment. The substantial improvement from Phase 1 (mIoU = 0.611) to Phase 2 (mIoU = 0.851) demonstrates that targeted dataset preparation using a foundation model for boundary refinement can significantly enhance segmentation performance without requiring additional manual annotation. The finding that SAM mask-prompt mode outperforms bounding-box prompt mode in structural segmentation contexts provides a practical guideline for applying foundation models to specialized inspection datasets.

The YOLOv8s damage detection model trained on a simplified two-class configuration achieved a mAP50 of 0.445 for crack and corrosion detection. The comparative evaluation across three model configurations revealed that leakage detection represents a fundamentally more challenging problem due to its intrinsic visual variability, and that reducing class complexity by removing the leakage class improved overall detection stability and precision. A confidence threshold of 0.30 was identified as optimal through qualitative threshold analysis, providing the best balance between detection sensitivity and false positive suppression.

The integrated structure-aware damage mapping pipeline successfully generated labeled outputs associating detected damage with identified structural members, achieving a member assignment accuracy of 62.0% fully correct and 87.0% fully or partially correct on 100 bridge inspection images. The spatial assignment strategy combining region-based overlap analysis with a center-point fallback mechanism proved effective in handling real inspection imagery where damage regions partially overlap multiple structural areas.

Several limitations were identified that motivate future research directions. The current framework detects only two damage types; extending coverage to spalling, delamination, and leakage remains an important next step. Although leakage detection proved challenging in the current study due to its intrinsic visual variability, future work targeting improved leakage-specific architectures and annotation strategies may improve coverage. While the evaluation on 100 images provides a meaningful initial benchmark with per-member-class breakdown, a larger-scale quantitative evaluation across diverse bridge types and inspection environments is needed to fully characterise framework performance. Furthermore, performance degradation under domain shift conditions — where the segmentation and detection models were trained on images from different inspection contexts — highlights the importance of collecting diverse multi-domain training data for robust deployment.

Future work will focus on expanding the training dataset to cover a wider range of bridge types, structural conditions, and imaging environments; incorporating additional damage categories through improved annotation strategies and class-balanced training; and developing a systematic quantitative evaluation protocol for the integrated pipeline including member assignment accuracy, damage localization precision, and framework-level performance metrics. In particular, the detection challenges associated with visually ambiguous classes such as leakage and the structurally small bearing class motivate the application of generative data augmentation strategies. Generative adversarial network (GAN)-based image synthesis and self-training with domain adaptation, which

 doi:10.20944/preprints202605.1350.v1

have demonstrated effectiveness in expanding training data for rare or visually complex categories in related infrastructure inspection tasks [21,22], represent promising directions for improving detection coverage of these difficult classes without requiring extensive additional manual annotation. The integration of temporal inspection data and comparison with inspector-assigned condition ratings would further validate the practical utility of the proposed framework for real-world infrastructure monitoring applications.

Overall, the results demonstrate that the integration of transformer-based segmentation, efficient object detection, and spatial reasoning within a unified pipeline is a promising direction for automated bridge inspection, and that structure-aware damage reporting represents a meaningful advancement toward more practical and informative AI-assisted infrastructure assessment.

Author Contributions: Conceptualization, S.D.S. and P.-J.C.; methodology, S.D.S.; software, S.D.S.; validation, S.D.S. and P.-J.C.; formal analysis, S.D.S.; investigation, S.D.S.; resources, P.-J.C.; data curation, S.D.S.; writing—original draft preparation, S.D.S.; writing—review and editing, P.-J.C.; visualization, S.D.S.; supervision, P.-J.C.; project administration, P.-J.C. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The part of this research was conducted as commissioned research with the Council for Science, Technology and Innovation (CSTI), Cross-Ministerial Strategic Innovation Promotion Program (SIP), the 3rd period of SIP "Smart Infrastructure Management System" Grant Number JPJ012187 (Funding agency: Public Works Research Institute) and JSPS Grant-in-Aid for Scientific Research Grant Numbers 25K01302, 25K22110 and 23H00198. We express our gratitude for this support.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. American Society of Civil Engineers. 2025 Infrastructure Report Card: Bridges; ASCE: Reston, VA, USA, 2025. Available online: https://www.infrastructurereportcard.org (accessed on 1 March 2025).
2. Ministry of Land, Infrastructure, Transport and Tourism (MLIT). Road Maintenance in Japan: Problems and Solutions; MLIT: Tokyo, Japan, 2023. Available online: https://www.mlit.go.jp (accessed on 1 March 2025).
3. Japan International Cooperation Agency (JICA). Study on Bridge Rehabilitation and Reconstruction in Sri Lanka; JICA: Tokyo, Japan, 2018.
4. Koch, C.; Georgieva, K.; Kasireddy, V.; Akinci, B.; Fieguth, P. A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure. Adv. Eng. Inform. 2015, 29, 196–210. https://doi.org/10.1016/j.aei.2015.01.008
5. Chun, P.-J.; Dang, J.; Hamasaki, S.; Yajima, R.; Kameda, T.; Wada, H.; Yamane, T.; Izumi, S.; Nagatani, K. Utilization of unmanned aerial vehicle, artificial intelligence, and remote measurement technology for bridge inspections. J. Robot. Mechatron. 2020, 32, 211–220. https://doi.org/10.20965/jrm.2020.p0211
6. Zou, Q.; Zhang, Z.; Li, Q.; Qi, X.; Wang, Q.; Wang, S. DeepCrack: Learning hierarchical convolutional features for crack detection. IEEE Trans. Image Process. 2019, 28, 1498–1512. https://doi.org/10.1109/TIP.2018.2876600
7. Deng, L.; Yuan, H.; Long, L.; Chun, P.-J.; Chen, W.; Chu, H. Cascade refinement extraction network with active boundary loss for segmentation of concrete cracks from high-resolution images. Autom. Constr. 2024, 162, 105410. https://doi.org/10.1016/j.autcon.2024.105410
8. Salehi, H.; Jamshidighadikolaei, N.; Gopu, V.; Alaywan, W. Lab-to-field integration in bridge monitoring: A hybrid structural health monitoring framework employing deep learning and unmanned aerial vehicle imagery. Mach. Learn. Appl. 2026, 24, 100872. https://doi.org/10.1016/j.mlwa.2026.100872
9. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 2021, 34, 12077–12090.
10. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; Dollár, P.; Girshick, R. Segment anything. In Proceedings of the IEEE/CVF International

 doi:10.20944/preprints202605.1350.v1

Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371

11. Huang, L.; Fan, G.; Li, J.; Hao, H. Deep learning for automated multiclass surface damage detection in bridge inspections. Autom. Constr. 2024, 166, 105601. https://doi.org/10.1016/j.autcon.2024.105601
12. Yu, W.; Nishio, M. Multilevel structural components detection and segmentation toward computer vision-based bridge inspection. Sensors 2022, 22, 3502. https://doi.org/10.3390/s22093502
13. Hwang, S.-S.; Hwang, C.-H.; Chung, S.-W.; Kim, B.-K. A deep-learning-based bridge damaged object automatic detection model using a bridge member model combination framework. Appl. Sci. 2022, 12, 12868. https://doi.org/10.3390/app122412868
14. Azimi, M.; Eslamlou, A.; Pekcan, G. Data-driven structural health monitoring and damage detection through deep learning: State-of-the-art review. Sensors 2020, 20, 2778. https://doi.org/10.3390/s20102778
15. Cha, Y.J.; Choi, W.; BüYüköztürk, O. Deep learning-based crack damage detection using convolutional neural networks. Comput.-Aided Civ. Infrastruct. Eng. 2017, 32, 361–378. https://doi.org/10.1111/mice.12263
16. Yeum, C.M.; Dyke, S.J. Vision-based automated crack detection for bridge inspection. Comput.-Aided Civ. Infrastruct. Eng. 2015, 30, 759–770. https://doi.org/10.1111/mice.12156
17. Cha, Y.J.; Choi, W.; Suh, G.; Mahmoudkhani, S.; BüYüköztürk, O. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. Comput.-Aided Civ. Infrastruct. Eng. 2018, 33, 731–747. https://doi.org/10.1111/mice.12334
18. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8; Ultralytics: Los Angeles, CA, USA, 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 1 March 2025).
19. Xu, W.; Li, X.; Ji, Y.; Li, S.; Cui, C. BD-YOLOv8s: Enhancing bridge defect detection with multidimensional attention and precision reconstruction. Sci. Rep. 2024, 14, 18673. https://doi.org/10.1038/s41598-024-69722-8
20. Cheng, Y.; Shi, Y.; Zhao, K.; Zhao, Y. A novel YOLOv8-GAM-Wise-IoU model for automated detection of bridge surface cracks. Constr. Build. Mater. 2024, 411, 134551. https://doi.org/10.1016/j.conbuildmat.2023.134551
21. Chun, P.-J.; Suzuki, M.; Kato, Y. Iterative application of generative adversarial networks for improved buried pipe detection from images obtained by ground-penetrating radar. Comput.-Aided Civ. Infrastruct. Eng. 2023, 38(17), 2472–2490. https://doi.org/10.1111/mice.12984
22. Chun, P.-J.; Kikuta, T. Self-training with Bayesian neural networks and spatial priors for unsupervised domain adaptation in crack segmentation. Comput.-Aided Civ. Infrastruct. Eng. 2024, 39(17), 2642–2661. https://doi.org/10.1111/mice.13292