Automated Detection of Box-Girder Bridge Deterioration Using Cylindrical Projection from Multi-Camera 3D Reconstruction and Deep Learning

Ming-Yun Ou [1], Jyun-Ping Jhan [2], Chen-Kuang Lin [3], Shih-Syun Lin [4], Hsin-Chu Tsai [5], Tzu-Liang Chou [6], Chang-yu Chang [7]

[1] Graduate Institute of A.I. Cross-disciplinary Technology, NTUST, Taiwan – oumingyun@gmail.com
[2] Graduate Institute of A.I. Cross-disciplinary Technology, NTUST, Taiwan – jpjhan@mail.ntust.edu.tw
[3] Department of Mechanical and Materials Engineering, Tatung University, Taiwan – cglin@gm.ttu.edu.tw
[4] Graduate Institute of A.I. Cross-disciplinary Technology, NTUST, Taiwan – linss@mail.ntust.edu.tw
[5] Facility Technology Center, China Engineering Consultants, Inc., Taiwan – hsinchu@ceci.org.tw
[6] Facility Technology Center, China Engineering Consultants, Inc., Taiwan – julietchou@ceci.org.tw
[7] Facility Technology Center, China Engineering Consultants, Inc., Taiwan – daniel@ceci.org.tw

Keywords: Bridge Deterioration Inspection, Multi-Camera Imaging System, 3D Reconstruction, Semantic Segmentation

Abstract

As large-scale infrastructure gradually ages, hundreds of existing bridges require regular inspections to ensure structural safety. While many researchers have proposed deterioration detection methods based on computer vision and deep learning—which can detect deterioration at the image level—no effective approach has yet been developed that integrates 3D reconstruction technology to achieve spatial localization and area quantification. To address this, this study proposes a two-part automated inspection workflow for the classification, localization, and measurement of internal deterioration in box-girder bridges. In the first part, the camera system is calibrated using an indoor calibration scene, and images are captured inside the box-girder. A 3D model is constructed using Structure from Motion (SfM) algorithms, and a cylindrical projection unfolded map is generated. In the second part, a boundary-aware model—modified from DeepV3+—is used to perform pixel-level deterioration detection and classification on the unfolded map. Experimental results demonstrate that the system can generate scale-corrected cylindrical unfolded maps from 3D models with sub-millimeter scale accuracy (0.105 mm), effectively transforming complex 3D inspection tasks into measurable and analyzable 2D images. The model achieved an overall mean Intersection over Union (mIoU) of 65.11% across five classes (four deterioration types and the background), representing a 7.54 percentage point improvement over the original DeepV3+. The research results validate the effectiveness of the proposed workflow in enhancing detection efficiency and objectivity for box-girder bridge maintenance.

1 Introduction

The internal structure of box-girder bridges must be inspected regularly to maintain structural integrity. While previous studies have proposed using drones to assist with inspections, actual implementation still faces challenges due to the confined and narrow space inside box-girder bridges. Furthermore, even when deterioration is observed, it is difficult to precisely locate its spatial position and quantify its severity.

1.1. Research Motivation

Traditional bridge maintenance relies primarily on manual inspections, requiring inspectors to enter the interior of the box-girder to identify and document signs of deterioration. This process is not only time-consuming but also involves significant safety risks. Furthermore, the assessment of deterioration severity and type is largely subjective, often leading to inconsistent evaluation results among different inspectors. These issues highlight the urgent need for an automated inspection workflow capable of handling image acquisition, 3D spatial localization, and intelligent deterioration identification within box-girder bridges.

1.2. Related Work

Recent research has combined UAV imagery with deep learning for bridge deterioration detection. Researchers integrated UAV imagery with Faster R-CNN to achieve automatic detection of bridge cracks, demonstrating the feasibility of deep learning-based detection techniques in real-world applications (Li et al., 2023). Additionally, researchers have proposed a three-stage UAV-based crack sensing system that combines YOLO classification, U-Net segmentation, and 3D point cloud reconstruction techniques for crack quantification and visualization (Zhou et al., 2025). However, these methods are limited to external surface inspection, as conventional UAVs face challenges such as insufficient lighting and confined spaces when operating inside bridges (Panigati et al., 2025). Additionally, damage detection based on deep learning has been combined with photogrammetric 3D reconstruction to localize detected damage within reconstructed 3D models (Zhao et al., 2022). Overall, there are currently no existing methods capable of comprehensively addressing the dual challenges of 3D spatial localization and pixel-level damage segmentation within the confined interior of bridges.

1.3. Objectives

The objective of this study is to improve the efficiency of box-girder deterioration detection by developing a multi-camera imaging system that can be mounted on a rail vehicle or operated manually. The system acquires multi-perspective images and generates stitched panoramic images via 3D reconstruction and cylindrical projection. The locations and extents of deteriorated regions can be accurately determined from the panoramic images, while their types are identified using the proposed AI model.

Due to technical challenges in the data acquisition process, this study was conducted in two parts. First, the performance of the multi-camera system was evaluated through 3D reconstruction of a concrete bridge. Second, due to difficulties in obtaining access permits for steel bridges, the AI-based deterioration

detection method was validated using a steel bridge dataset provided by China Engineering Consultants, Inc. (CECI).

2 System Design and Data

This section presents the hardware configuration of the multi-camera imaging system, the box-girder field site, and the deterioration dataset used to train and validate the artificial intelligence model.

2.1. Multi-Camera System

As shown in Figure 1, the imaging system consists of five industrial cameras, an industrial computer, a mobile display, and a power supply. The system can be operated manually or mounted on a rail vehicle, as also shown in Figure 1. Table 1 summarizes the camera specifications. Four cameras are used to capture side-view images, while the fifth camera captures forward-facing views. Each camera is equipped with a 12-megapixel sensor and a 4.5 mm focal-length fisheye lens, achieving a ground sampling distance (GSD) of approximately 1.2 mm at a capture distance of 2 m. With a wide field of view of up to 128° × 106°, the overlap between adjacent cameras is approximately 23°, ensuring sufficient feature correspondences for reliable 3D reconstruction.

A control system was developed to enable real-time monitoring and synchronized image acquisition. In addition, high-brightness LED lights mounted on the brackets are used to compensate for the limited natural illumination inside the box-girder bridge.

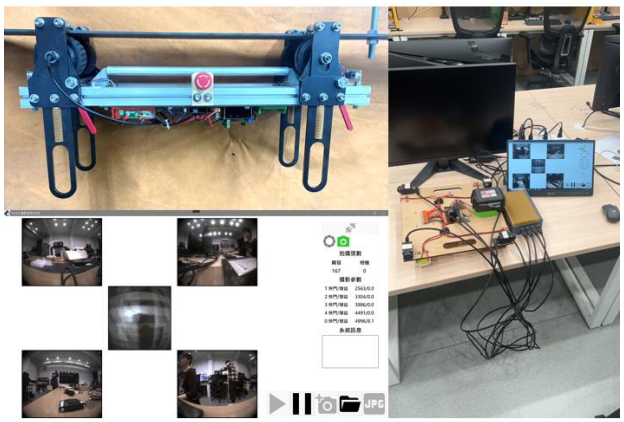

Figure 1. Multi-camera system and user interface.

Resolution	4096 × 3000 pixels
Pixel Size	2.74 µm
Focal Length	4.5 mm
Lens Type	Fisheye
Field of View (FOV)	128° × 106°
GSD @ 2 m Distance	1.2 mm

Table 1. Specifications of the camera sensors

2.2. Concrete Box-girder Field Site

In this study, the camera system was validated within the interior structure of a section of an elevated bridge on National Freeway No. 3 in Taiwan. The bridge is a concrete box-girder structure with an internal width of approximately 4.5 m and a height of approximately 4 m, making it suitable for evaluating the reliability of 3D reconstruction.

Since the rail system has not yet been installed, image acquisition was performed in a handheld manner. The camera system was moved along the longitudinal axis of the bridge, and images were captured at intervals of approximately 1 m, resulting in a total of 249 image sets (1,245 images). Figure 2 shows an example of the captured images from the concrete bridge.

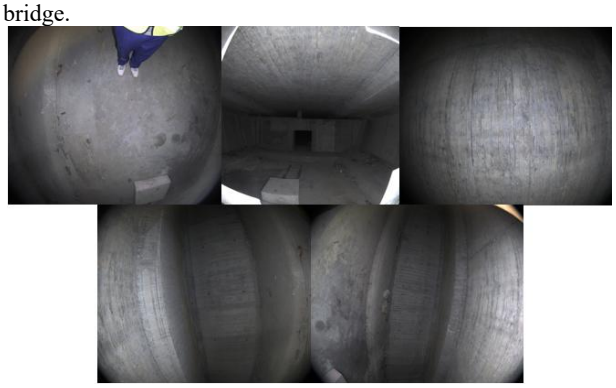

Figure 2. Sample images. Top row from left to right: bottom view, front view, top view. Bottom row: left view, right view.

2.3. Deterioration Dataset of Steel Bridge

Due to the insufficient number of deterioration cases captured at the test site, an external dataset comprising 594 images of deterioration was obtained from China Engineering Consultants, Inc. (CECI). These images were captured by personnel using handheld cameras or mobile phones during routine inspections of steel box-girder bridges, resulting in variations in shooting angle, resolution, and image quality.

The dataset covers the four most common types of internal deterioration in steel box-girder bridges: corrosion, water seepage, efflorescence, and paint damage. Prior to annotation, all images were pre-processed through cropping and rotation to remove extraneous elements such as foreign objects, people, and embedded timestamps, thereby retaining only the bridge surface and the areas of deterioration. These images were then annotated at the pixel level within Label Studio to create a ground-truth dataset for semantic segmentation, where blue, red, yellow, and green represent water seepage, rust, efflorescence, and paint damage, respectively. The dataset was split into training and validation sets in an 80:20 ratio, and data augmentation techniques—including random flipping—were applied to enhance sample diversity. Figure 3 shows representative images of deterioration and their corresponding annotations.

Figure 3. Sample images and ground truth.

3 Methodology

As shown in Figure 4, the proposed workflow comprises two parts. The first part involves calibrating the camera system, capturing images, and performing 3D reconstruction using the calibrated parameters to generate a cylindrical projected panoramic image. The second part involves enhancing the DeepV3+ model by incorporating a boundary-aware mechanism and evaluating its performance in steel bridge deterioration detection.

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-1085-2026

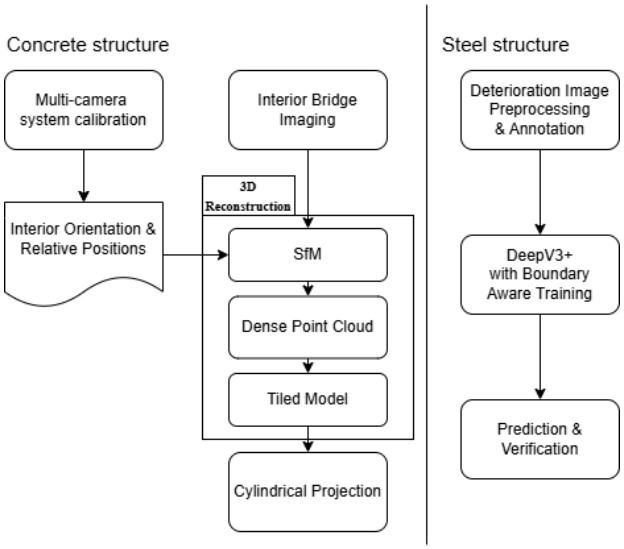

Figure 4. Overall workflow of the bridge inspection system.

3.1. Camera System Calibration

The multi-camera system used in this study requires not only the calibration of the interior orientation parameters (IOPs) of each camera, but also the estimation of the relative orientation parameters (ROPs) between the side-view and forward-view cameras. By integrating the relative orientation information, the system can provide scale information, thereby enabling model localization and scale measurement without the need for control points.

To achieve this objective, an indoor calibration scene was established, with multiple sets of control point pairs of known scale arranged on the surrounding walls (Figure 5). Image acquisition was performed at intervals of approximately 1 m, and after each image set was captured, the camera system was rotated by 90° to ensure sufficient overlap between images and to facilitate the observation of control points.

All images were processed using Metashape (Agisoft LLC, 2025), and conventional photogrammetric methods were applied to calibrate the IOPs and ROPs of each camera. The forward-facing camera was defined as the primary camera, while the side-view cameras were treated as auxiliary cameras.

Figure 5. (a) Indoor calibration field (b) Control point pairs with known scale.

3.2. 3D Reconstruction and Cylindrical Projection

The purpose of 3D reconstruction is to employ Structure from Motion (SfM) to estimate the camera positions and orientations at the time of image capture, thereby generating an accurate cylindrical projected panoramic image. By utilizing the IOPs and ROPs of the multi-camera system, the SfM reconstruction process can achieve absolute scale without relying on a large number of control points. However, to verify the reconstruction accuracy, control points with known coordinates were placed at the front, middle, and rear sections inside the bridge for validation.

Furthermore, to unfold the closed surface of the box-girder into a two-dimensional representation, a cylindrical projection method is applied after generating the three-dimensional point cloud and model. Using the longitudinal axis of the bridge as the central axis, the surrounding walls are projected onto a panoramic image. This panoramic image preserves the spatial correspondence with the bridge structure, allowing each deterioration region to be traced back to its corresponding location within the bridge and enabling direct measurement of the deteriorated area.

3.3. DeepV3+ with Boundary Aware

The original DeepV3+ (Chen et al., 2018) suffers from the loss of boundary information due to consecutive downsampling, making it unable to delineate deterioration boundaries or distinguish overlapping regions between different deterioration types. To address this issue, the method proposed in this study enhances DeepV3+ in three aspects: feature fusion, decoder design, and loss function. As shown in Figure 6, the modified architecture integrates the BA-CSF module (Boundary Detection, Boundary-Guided Attention, and Adaptive Fusion), BASPP (Boundary-Aware Atrous Spatial Pyramid Pooling), and BRB (Boundary Residual Block), and is trained end-to-end with the DRN backbone model. Details on the loss function design are provided in Section 3.4.

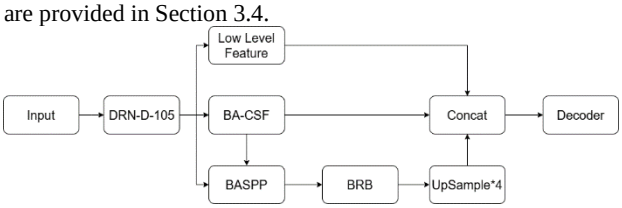

Figure 6. Architecture of the Boundary-Aware DeepV3+.

3.3.1. DRN-D-105 Backbone Network

The original DeepV3+ employs a ResNet backbone, where successive pooling reduces the deep feature resolution to 1/32, causing significant loss of boundary information. This study replaces it with DRN-D-105 (Yu et al., 2017), which substitutes standard downsampling with dilated convolutions in the later stages, maintaining the deep feature maps at 1/8 resolution. The network outputs a low-level feature map F_{low} (32 channels, 1/2 resolution) and a high-level feature map F_{high} (2048 channels, 1/8 resolution), retaining spatial boundary cues and semantic information respectively, and is initialized with ImageNet (Deng et al., 2009) pre-trained weights.

3.3.2. Boundary-Aware Cross-Scale Fusion (BA-CSF)

Since shallow features retain spatial details while deep features capture semantic context, effective fusion of both is critical for accurate deterioration boundary delineation. A boundary probability map is derived from F_{low} using a sigmoid activation:

$$B_{map} = \sigma(B_{logits}) \quad (1)$$

To prevent boundary information from being diluted during feature fusion, a boundary-guided attention mechanism enhances both shallow and deep projected features $F \in R^{B\times 256\times H\times W}$:

$$A = \sigma\left(Conv\left(\left[F; B_{map}\right]\right)\right) \quad (2)$$

$$F' = F \odot (1 + A) \quad (3)$$

The residual design (1+A) preserves original features while amplifying boundary-relevant responses, with the enhancement factor bounded within [1, 2]. The enhanced features are then fused via pixel-wise adaptive weighting, enabling the network to learn region-dependent fusion strategies:

$$\left[w_{low}, w_{high}\right] = Softmax\left(Conv\left(\left[F'_{low}; F'_{high}; B_{map}\right]\right)\right) \quad (4)$$

$$F_{fused} = w_{low} \odot F'_{low} + w_{high} \odot F'_{high} \quad (5)$$

This allows boundary pixels to favor shallow spatial information for precise localization, while interior pixels favor deep semantic features for accurate classification.

3.3.3. Boundary Atrous Spatial Pyramid Pooling (BASPP)

The deep feature F_{high} is processed through five parallel ASPP branches (global average pooling, 1×1 convolution, and three depthwise separable convolutions with dilation rates of 6, 12, 18). A boundary-guided reweighting mechanism dynamically adjusts each branch's contribution based on B_{map}:

$$W = Softmax\left(\frac{Conv\left(F_{high}, B_{map}\right)}{\tau}\right) \quad (6)$$

where τ is a learnable temperature parameter initialized at 2.0 and K=5 is the number of branches. The resulting weights are multiplied by K to maintain the expected scale. This encourages boundary pixels to rely on small-dilation branches to avoid cross-boundary feature mixing, while interior pixels favor large-dilation branches for broader context. The weighted outputs are fused via a 1×1 convolution:

$$F_{aspp} = Conv_{1\times1}(Cat(w_k \odot F_k)) \quad (7)$$

To further refine residual blurring at boundaries, a Boundary Detail Branch (BDB) applies a gated residual correction with a learnable scalar α initialized at zero, ensuring progressive introduction during training.

3.3.4. Decoder

The decoder concatenates and fuses three feature paths: F'_{aspp}, F_{low}, and F_{fused}, and outputs five-class segmentation results after convolutional refinement.

3.4. Loss Function Design

The training objective comprises four loss terms. In addition to the standard weighted cross-entropy loss L_{CE} and boundary loss L_B, two novel loss functions are proposed.

3.4.1. Hierarchical Boundary-Aware Cross-Entropy Loss

Reformulates the five-class prediction into a coarse binary classification (background vs. defect) by merging the four deterioration logits via log-sum-exp:

$$Z_{coarse} = \left[z_0, log\sum_{c=1}^{4} exp(z_c)\right] \quad (8)$$

$$L_{HBCE} = mean\left(w(x) \cdot CE(Z_{coarse}, y_{coarse})\right) \quad (9)$$

where $w(x) = 1 + 3.0 \times B_{map}(x)$, amplifying the loss at boundary pixels by approximately four times, forcing the model to pay particular attention to avoiding missed detections at boundaries.

3.4.2. Boundary-Aware Asymmetric Tversky Loss

For classes with lower IoU, the Tversky index imposes a stronger penalty on false negatives (FN), encouraging the model to expand its predictions to recover under-segmented boundary regions. All pixels are weighted by $w(x) = 1 + 3.0 \cdot B_{map}(x)$, giving boundary pixels approximately four times the influence of interior pixels. The overall training objective is defined as:

$$L = \lambda_1 \cdot L_{CE} + \lambda_2 \cdot L_{HBCE} + \lambda_3 \cdot L_{BAAT} + \lambda_4 \cdot L_B \quad (10)$$

4 Results and Analysis

This section presents the experimental results for each stage of the proposed workflow: camera calibration (Section 4.1), 3D reconstruction and cylindrical projection (Section 4.2), and deterioration recognition performance of the improved DeepV3+ with boundary aware model (Section 4.3).

4.1. Calibration Verification

Figure 7 shows the reconstruction results of the indoor calibration scene. A total of 33 images were taken, and 12,000 control points were used. Table 2 lists the ROPs and their standard deviations for each camera with respect to the reference camera. The relative distances are approximately 15 cm, with positional accuracy better than 1 mm. The rotation angles (ω, φ, κ) are close to 90°, with angular accuracy better than 0.1°. These results are consistent with the installation configuration, indicating the reliability of the calibration.

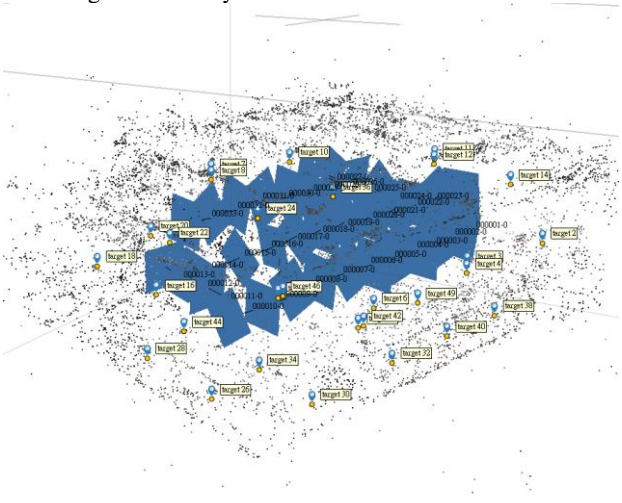

Figure 7. Orientation reconstruction results

ROP	X (m)	Y (m)	Z (m)	ω (°)	φ (°)	κ (°)
Cam 2	0.15	0.17	0.04	89.6	0.54	0.11
σ	0.02	0.02	0.01	0.08	0.07	0.05
Cam 3	-0.18	0.15	0.04	-90.1	88.8	179.7
σ	0.02	0.02	0.01	0.08	0.06	0.08
Cam 4	-0.15	-0.18	0.04	-89.4	-0.70	179.7
σ	0.02	0.02	0.01	0.07	0.06	0.05
Cam 5	0.17	-0.15	0.04	90.0	-89.6	-0.06
σ	0.02	0.02	0.01	0.08	0.06	0.08

Table 2. Relative orientation standard deviations.

4.2. 3D Reconstruction Results of Concrete Bridge

Figure 8 presents the 3D reconstruction results. A total of 1,237 images were aligned, generating 674,466 tie points. Using twelve sets of control points and six scale bars (reference distance 0.129 m), the total scale error was 0.105 mm,

confirming sub-millimeter accuracy. The aligned images produced a dense point cloud of approximately 132 million points, from which a tiled model was constructed.

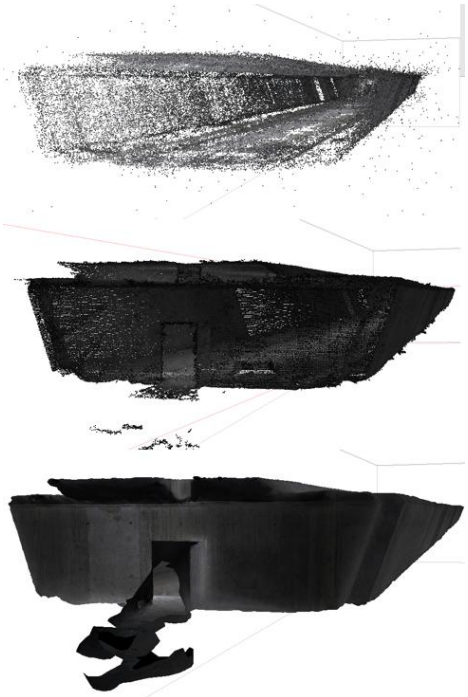

Figure 8. 3D reconstruction results. Top to bottom: (a) sparse point cloud, (b) dense point cloud, (c) tiled model.

A cylindrical projection was applied along the bridge longitudinal axis, producing an unfolded map of 62,857 × 109,171 pixels at 0.816 mm/pixel (Figure 9). Verification against field-surveyed control point distances confirmed millimeter-level metric accuracy. The unfolded map was segmented into 26,322 patches of 512 × 512 pixels, each covering approximately 418 mm × 418 mm, serving as input for the semantic segmentation model.

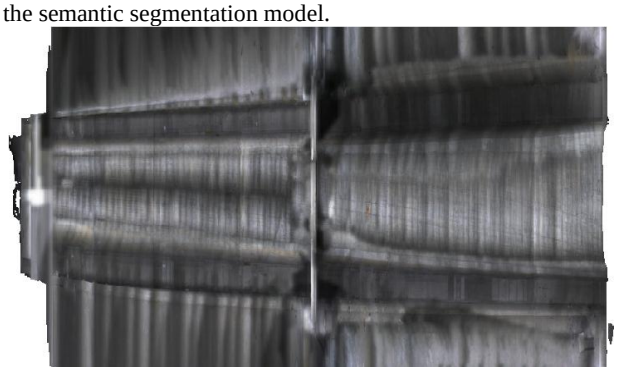

Figure 9. Projected view of the bridge interior.

4.3. Deterioration Detection

In this study, the proposed model was trained on a single NVIDIA RTX 4070 Ti Super GPU using the MMSegmentation framework with the SGD algorithm (learning rate = 0.005, momentum = 0.9, weight decay = 0.0005) for 10,000 iterations, with a batch size of 2, and an input resolution of 512 × 512 pixels. As shown in Table 3, the model achieved an overall mIoU of 65.11 %. The IoU for the seepage class was the lowest, primarily because its visual appearance varies greatly under different environmental conditions, making it inherently difficult to segment. Figure 10 displays representative segmentation results on the validation set.

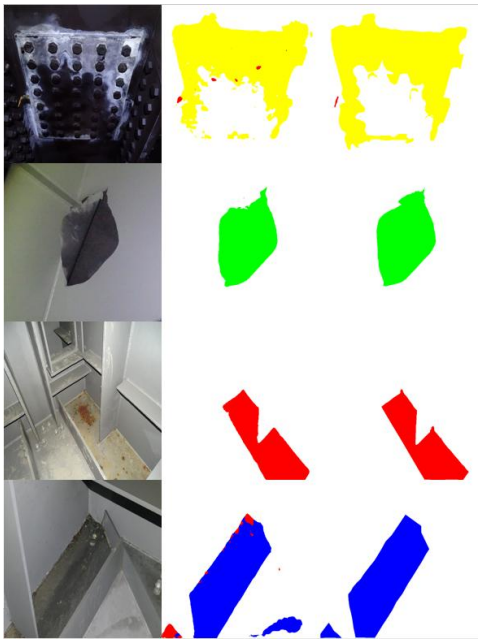

Figure 10. Inference results on the validation set.

Class	IoU(%)	Acc(%)	Precision(%)	Recall(%)
Background	93.95	96.53	97.23	96.53
Efflorescence	57.56	67.82	79.20	67.82
Paint damage	70.74	82.06	83.68	82.06
Corrosion	67.26	84.41	76.80	84.41
Water seepage	36.04	62.21	46.14	62.21
Overall	65.11	78.61	76.61	78.61

Table 3. Performance metrics of the Boundary-Aware DeepV3+

To evaluate the effectiveness of the proposed model, we used the original DeepV3+ as the baseline model and trained it under the same configuration. As shown in Table 4, the proposed model consistently outperforms the baseline model across all deterioration categories, both in terms of IoU and recall. The most significant improvements were observed in the efflorescence and seepage categories, where the boundary-aware attention mechanism and boundary-guided loss function effectively addressed the issue of boundary blurring and reduced false negatives. Figure 11 presents a comparison of the segmentation results between the two models on the test set.

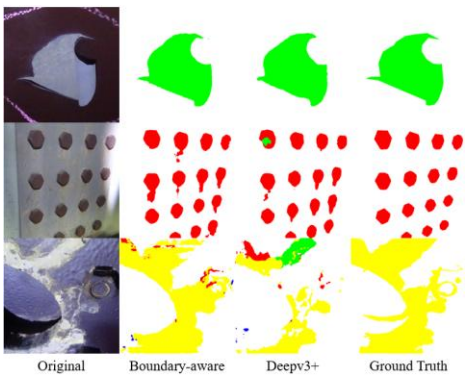

Figure 11. Inference results of the boundary-aware model and the baseline model

Class	Baseline IoU (%)	Proposed IoU (%)	Improvement
Background	93.33	93.95	+0.62
Efflorescence	42.72	57.56	+14.84

Paint damage	66.76	70.74	+3.98
Corrosion	60.86	67.26	+6.40
Water seepage	24.18	36.04	+11.86
Overall	57.57	65.11	+7.54

Table 4. Comparing Boundary-Aware Models with the Baseline

5 Conclusions

This study proposes an automated inspection workflow tailored to the internal inspection needs of box-girder bridges. The workflow integrates a camera system, 3D reconstruction based on Structure from Motion (SfM) and cylindrical projection techniques, and a boundary-aware DeepV3+ model. The system achieves sub-millimeter reconstruction accuracy (scale error of 0.105 mm), with a mean IoU (mIoU) of 65.11% for the semantic segmentation of four types of deterioration. Through the cylindrical projection, both the spatial localization and area quantification of deterioration can be performed simultaneously on a unified two-dimensional plane, effectively transforming complex 3D inspection tasks into two-dimensional analyses with precise measurements.

The current deterioration dataset used for model training consists of deterioration images provided by China Engineering Consultants, Inc., whose image features differ from those of cylindrical projection unfolded diagrams. In the future, this study will conduct on-site photography and 3D reconstruction of the interiors of more box-girder bridges, followed by segmentation and pixel-level annotation of the resulting cylindrical projection diagrams. We will then establish a corresponding deterioration dataset to validate the model’s segmentation performance.

References

Agisoft LLC, 2025. Metashape Professional. Software. agisoft.com (20 March 2026).

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Computer Vision – ECCV 2018, Springer, Cham, 833–851, doi.org/10.1007/978-3-030-01234-2_49.

Deng, J., Dong, W., Socher, R., Li, L. J., Kai, L., and Li, F.-F., 2009. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255, doi.org/10.1109/CVPR.2009.5206848.

Li, R., Yu, J., Li, F., Yang, R., Wang, Y., and Peng, Z., 2023. Automatic bridge crack detection using Unmanned aerial vehicle and Faster R-CNN. Construction and Building Materials, 362, 129659, doi.org/10.1016/j.conbuildmat.2022.129659.

Panigati, T., Zini, M., Striccoli, D., Giordano, P. F., Tonelli, D., Limongelli, M. P., and Zonta, D., 2025. Drone-based bridge inspections: Current practices and future directions. Automation in Construction, 173, 106101, doi.org/10.1016/j.autcon.2025.106101.

Yu, F., Koltun, V., and Funkhouser, T., 2017. Dilated Residual Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 636–644, doi.org/10.1109/CVPR.2017.75.

Zhao, S., Kang, F., and Li, J., 2022. Concrete dam damage detection and localisation based on YOLOv5s-HSC and photogrammetric 3D reconstruction. Automation in Construction, 143, 104555, doi.org/10.1016/j.autcon.2022.104555.

Zhou, L., Jiang, Y., Jia, H., Zhang, L., Xu, F., Tian, Y., Ma, Z., Liu, X., Guo, S., Wu, Y., Zhao, Z., and Zheng, H., 2025. UAV vision-based crack quantification and visualization of bridges: system design and engineering application. Structural Health Monitoring, 24, 1083–1100, doi.org/10.1177/14759217241251778.