Transformer-based Multi-resolution Fast 3D Reconstruction for Structural Damage Detection

Hui Zuo[1], Tao Sun[2], Hao Xie[1], Xiao Ma[3], Nima Shirzad-Ghaleroudkhani[1] and Qipei Mei[1]

[1]Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada
[2]Department of Civil Engineering, McGill University, Montréal, QC H3A 0C3, Canada
[3]Department of Biomedical Engineering, University of Alberta, Edmonton, AB T6G 2T4, Canada
polo.huizuo@ualberta.ca, tao.sun@mail.mcgill.ca, hxie9@ualberta.ca, xm11@ualberta.ca, shirzadg@ualberta.ca, qipei.mei@ualberta.ca

Abstract –

This study proposes a multi-resolution fast 3D reconstruction framework that integrates transformer-based damage detection with rapid 3D modeling to enhance bridge surface defect identification and spatial localization. The framework consists of three phases: (I) 3D reconstruction using Structure from Motion (SfM) to generate a structural model with sparse point cloud, (II) damage segmentation via a customized Swin UNETR model for precise defect detection, and (III) multi-resolution dense reconstruction that prioritizes high-resolution modeling of detected defects while reducing the resolution of non-critical areas to improve efficiency. Experimental validation on the High Level Bridge in Edmonton, Canada, demonstrated the framework's capability to accurately map surface defects onto a 3D model, providing an intuitive and detailed localization for structural assessment. This approach offers significant potential for efficient and accurate bridge inspection, supporting data-driven maintenance strategies.

Keywords –

Transformer; Point Cloud; 3D Reconstruction; Structure from Motion; Structural Damage Detection

1 Introduction

Bridges are critical components of civil infrastructure but are subjected to various adverse loads throughout their service life, accelerating structural deterioration [1]. The accumulation of these loads induces repeated stress and strain, leading to the formation of surface cracks that gradually compromise structural integrity over time [2]. Thus, regular inspections and timely repair based on damage assessment are essential to ensure bridge safety [3]. Recent advances in computer vision have greatly improved defect detection methods, enabling efficient analysis of large-scale image datasets for rapid damage assessment [4-5]. However, the widespread and varied nature of bridge damage presents challenges for existing techniques in simultaneously detecting and localizing surface defects [6]. Moreover, mapping surface damage onto a 3D model has proven beneficial for effective bridge maintenance, offering critical insights that support accurate decision-making by bridge owners [7]. Therefore, achieving precise detection and localization of bridge surface damage within a 3D model is of significant importance [8].

Currently, vision-based damage detection methods can be mainly divided into image classification, object detection, and semantic segmentation [9]. With advancements in deep learning, semantic segmentation has become a leading approach [10]. Various convolutional neural networks (CNNs) have been applied to damage segmentation, including U-Net [11], ResNet [12], DenseNet [13], Mask R-CNN [14], DeepLabV3+ [15], and hybrid models [16]. However, due to the limited receptive fields of CNN-based models, capturing global features in complex real-world scenarios remains challenging [17]. One promising solution is the use of transformer. Since Dosovitskiy et al. [18] introduced the encoder-based Vision Transformer (ViT), transformer has gained attraction in computer vision. Its embedding and self-attention mechanisms have shown superior accuracy compared to complex CNN models for image segmentation tasks [19]. Swin UNETR [20], which integrates the Swin Transformer [21] with an enhanced ViT, has demonstrated outstanding accuracy and efficiency across benchmark datasets [22]. Moreover, combining Unmanned Aerial Vehicle (UAV) with advanced algorithms offers significant potential for rapid inspection of high-risk areas, such as bridge piers and abutments [23].

To obtain the 3D spatial information of bridge damage, point cloud data is commonly used to generate a 3D bridge model. With the development in localization technology, stereo vision-based local positioning methods have been applied, addressing the reliance of

UAVs on the Global Positioning System (GPS) [24]. And Structure from Motion (SfM) is widely regarded as one of the common techniques for 3D reconstruction from a series of sparse images. This method enables the creation of a complete 3D model, integrating local detection results and reflecting the spatial position of objects [25]. Although semantic segmentation can classify damage at the pixel level, it faces challenges in consistently identifying and matching the same damage across sequential image frames, thus hindering the integration of damage segmentation with spatial localization [6]. Additionally, generating 3D bridge models with SfM presents several challenges. First, precise image capture and detailed trajectory planning are often required to avoid motion blur during data collection to ensure successful reconstruction. This time-consuming process significantly reduces efficiency [26]. Second, processing large volumes of images with SfM greatly extends computation time, further limiting reconstruction efficiency [27].

In response, this study integrates the high accuracy of image-based damage detection with the spatial localization capabilities of 3D reconstruction. A new method is proposed that combines transformer-based damage detection with rapid 3D reconstruction, enabling seamless and efficient mapping of surface damage on bridges onto the 3D model. This approach offers decision-makers an intuitive and comprehensive localization of structural assessment.

2 Methodology

The proposed multi-resolution 3D reconstruction framework comprises three phases: (I) 3D reconstruction, (II) transformer-based damage segmentation, and (III) multi-resolution dense reconstruction. Phase I generates a 3D structural model with data captured by UAV using SfM. Phase II employs a transformer model to detect and segment surface defects from feature images. Phase III performs multi-resolution dense reconstruction, producing a dense point cloud where defects are rendered in high resolution while the overall structural resolution is reduced, enabling fast and precise damage localization. The overall scheme is shown in Figure 1.

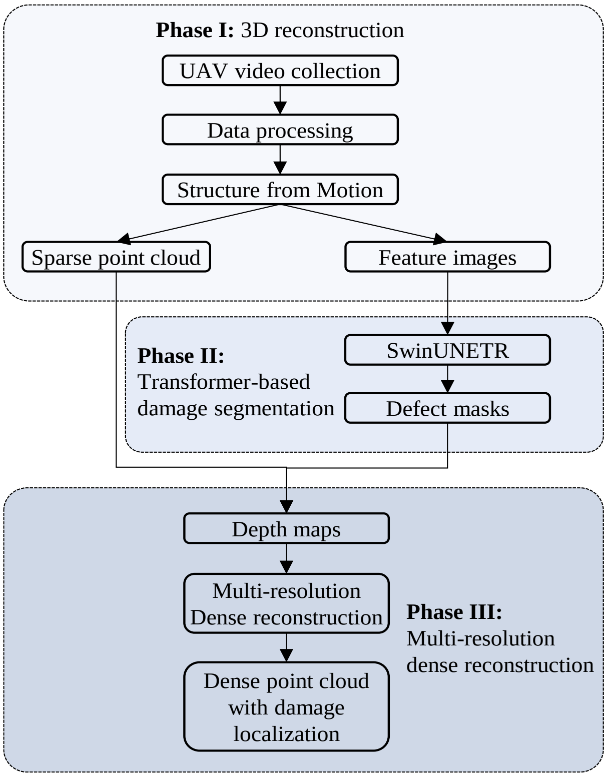
Phase I: 3D reconstruction
UAV video collection
Data processing
Structure from Motion
Sparse point cloud
Feature images
Phase II: Transformer-based damage segmentation
SwinUNETR
Defect masks
Depth maps
Multi-resolution Dense reconstruction
Dense point cloud with damage localization
Phase III: Multi-resolution dense reconstruction

Figure 1. Proposed Framework

2.1 3D reconstruction

In this phase, the incremental SfM [28] is used to process RGB images collected from UAV inspections, with each image associated with a local coordinate system (LCS) defined by the camera's field of view and orientation. The goal is to convert these LCS-based observations into a global coordinate system (GCS) to enable accurate 3D reconstruction. This process requires the extraction of intrinsic and extrinsic camera parameters and involves feature matching and incremental reconstruction.

To accurately represent the reconstructed model in a global coordinate system, it is essential to convert the image data from the LCS to the GCS [29]. The intrinsic camera matrix $\mathbf{K}$ (see Equation (1)) and the extrinsic rotation and translation matrix $\mathbf{R}\,|\,\mathbf{t}$ (see Equation (2)) are necessary for transforming coordinates from the LCS to the GCS. The intrinsic matrix $\mathbf{K}$ captures the camera's internal parameters, including the focal length (f_x, f_y), camera lens distortion s and the principal point coordinates (c_x, c_y).

$$\mathbf{K} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

While the extrinsic matrix $\mathbf{R}\,|\,\mathbf{t}$ defines the camera's rotation and translation in the global frame, including the camera position transformation matrix $r_{i,j}$, and the translation vector $\mathbf{t}=\left[t_x, t_y, t_z\right]^T$.

$$\mathbf{R}\,|\,\mathbf{t}=\begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

Finally, the relationship in Equation (3) can be used to transform a 3D point in the LCS (x_L, y_L, z_L) into the GCS (X_G, Y_G, Z_G) through the extrinsic parameters:

$$\begin{bmatrix} x_L \\ y_L \\ z_L \\ 1 \end{bmatrix} = [\mathbf{R}\,|\,\mathbf{t}] \begin{bmatrix} X_G \\ Y_G \\ Z_G \\ 1 \end{bmatrix} \tag{3}$$

where the rotation matrix $\mathbf{R}$ is further decomposed into rotations around the X, Y, and Z axes as shown below.

$$\mathbf{R}=\begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y \\ 0 & 1 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & -\sin\theta_x & \cos\theta_x \end{bmatrix} \tag{4}$$

2.2 Transformer-based damage segmentation

In this phase, the Swin UNETR, designed for 3D tumor segmentation, is customized for 2D structural damage segmentation. The model processes 2D structural images with two channels by dividing them into non-overlapping patches using a patch partition layer. These patches are transformed into windows suitable for self-attention operations. The encoded features from the Swin transformer are passed through skip connections at various resolutions to a CNN-based decoder. The final output is a segmentation map with two channels, identifying damaged areas in the structure. The architecture of the customized Swin UNETR is shown in Figure 2.

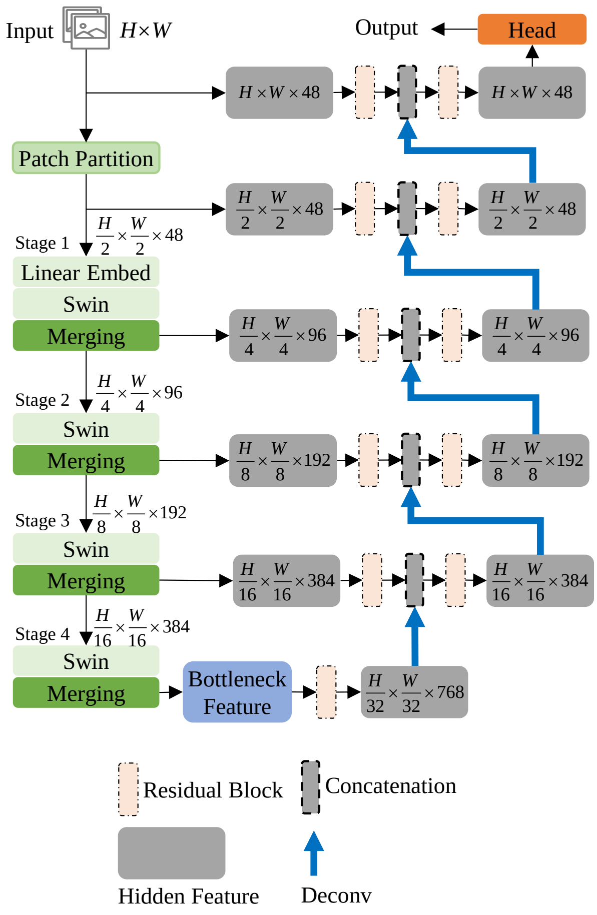

Figure 2. Architecture of the customized Swin UNETR

For the encoding part, the model input $\chi \in \mathbb{R}^{H \times W \times S}$ represents a token with a patch resolution of (H', W') and a dimension of $H' \times W' \times S$. Initially, the model uses a patch partitioning layer to segment the input image into fixed-size, non-overlapping patches with dimensions $\frac{H}{H'} \times \frac{W}{W'}$. These patches are then embedded into a feature space of dimension C. Self-attention is applied to these patches within non-overlapping windows to efficiently model relationships between tokens. Specifically, windows of size $M \times M$ partition the token into $\frac{H'}{M} \times \frac{W'}{M}$ regions at a given layer l in the transformer encoder. In the following layer, window positions are shifted by $\left(\frac{M}{2}, \frac{M}{2}\right)$ pixels to capture cross-region interactions. The encoder computes layer outputs using the following equations:

$$\begin{aligned}\tilde{z}^{l} &= W\text{-}MSA\left(LN\left(z^{l-1}\right)\right)+z^{l-1}\\ z^{l} &= MLP\left(LN\left(\tilde{z}^{l}\right)\right)+\tilde{z}^{l}\\ \tilde{z}^{l+1} &= SW\text{-}MSA\left(LN\left(z^{l}\right)\right)+z^{l}\\ z^{l+1} &= MLP\left(LN\left(\tilde{z}^{l+1}\right)\right)+\tilde{z}^{l+1}\end{aligned} \tag{5}$$

where W-MSA and SW-MSA represent standard and shifted window-based multi-head self-attention modules. LN and MLP denote layer normalization and multi-layer perceptrons. $\tilde{z}^{l}$ and $\tilde{z}^{l+1}$ indicate the output of W-MSA and SW-MSA, respectively.

For the 2D task, the encoder begins with a patch size of 2×2 and a feature dimension of 2×2×4=16. The embedding dimension C is set to 48. The encoder consists of four stages, each with two transformer blocks, totalling eight layers. A linear embedding layer reduces the spatial resolution by half at each stage, resulting in feature maps of size $\frac{H}{2}\times\frac{W}{2}$, $\frac{H}{4}\times\frac{W}{4}$, $\frac{H}{8}\times\frac{W}{8}$, $\frac{H}{16}\times\frac{W}{16}$.

The decoder adopts a U-shaped architecture where feature representations from the encoder are integrated through skip connections at corresponding resolutions. In each stage i, $i \in \{0,1,2,3,4\}$, the feature maps are reshaped to $\frac{H}{2^{i}}\times\frac{W}{2^{i}}$ and processed through residual blocks composed of two 3×3 convolutional layers, followed by instance normalization. To progressively restore spatial resolution, deconvolutional layers double the size of feature maps, and these upsampled features are concatenated with the outputs from the encoder. This process continues until the original resolution is recovered. Finally, a 1×1 convolutional layer and a sigmoid activation function are used to generate the segmentation mask, highlighting regions of structural damage.

The Focal Loss [30] was employed as the loss function, which can effectively address the class imbalance commonly present in damage detection tasks, where damaged regions occupy a much smaller area compared to the background. As displayed in Equation (6), it modifies the standard cross-entropy loss by adding a modulating factor that down-weights easy examples.

$$FL\left(p_{t}\right)=-\alpha_{t}\left(1-p_{t}\right)^{\gamma}\log\left(p_{t}\right) \tag{6}$$

where p_t is the predicted probability for the true class, and

$$p_{t}=\begin{cases} p, if\ y=1\\ 1-p, if\ y=0\end{cases} \tag{7}$$

where p is the predicted probability and y is the ground truth label. α_t is a balancing factor for class imbalance, γ is the focusing parameter that adjusts the rate at which easy examples are down-weighted.

For performance evaluation, several metrics are used to assess the segmentation quality, including Intersection over Union (IoU), Dice Coefficient, and Pixel-wise Accuracy:

$$IoU=\frac{TP}{TP+FP+FN} \tag{8}$$

$$Dice=\frac{2TP}{2TP+FP+FN} \tag{9}$$

$$Accuracy=\frac{TP}{TP+FN} \tag{10}$$

where TP, FP and FN stand for True Positive, False Positive, False Negative, respectively.

By leveraging the Swin transformer's ability to model multi-scale contextual information and long-range dependencies, the network can effectively capture complex patterns of structural defects.

2.3 Multi-resolution Dense Reconstruction

SfM provides essential camera parameters (intrinsic and extrinsic), sparse 3D points, and their 2D image correspondences. However, this sparse representation lacks the detail needed for comprehensive modeling. To overcome this, Multi-View Stereo (MVS) leverages the camera's internal and external parameters to perform stereo-matching and identify points in space that have photometric consistency [31], thus transforming sparse point clouds into dense models. In this phase, Open Multi-View Stereo (OpenMVS) [32], an open-source library, is adapted for dense 3D reconstruction.

OpenMVS primarily employs a depth map fusion-based MVS approach, which consists of several stages. Initially, for each input image, the most relevant neighboring images are selected to form stereo pairs. Depth maps are then estimated for each image by identifying photometrically consistent points across these pairs. Once individual depth maps are computed, they are fused into a unified and dense point cloud that captures fine surface details.

To optimize the dense 3D reconstruction process and prioritize defect areas over the whole structure, a significant modification is to ignore specified regions in the input corresponding images during the dense reconstruction stage. In this workflow, the structural components of the object are reconstructed at a reduced resolution, significantly accelerating the dense point cloud generation for the overall structure. Meanwhile, defect regions identified by Swin UNETR are preserved and processed at full resolution. By this way, the fine details of critical defect areas are captured with high fidelity, while the less critical structural areas are downsampled to minimize computational load. As a result, the overall dense reconstruction process becomes faster and more efficient compared to standard methods that process the entire structure uniformly.

3 Experiments and Results

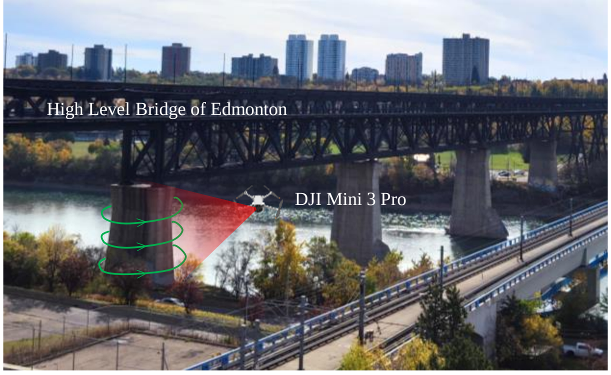

Figure 3. Data Collection

To evaluate the proposed multi-resolution 3D reconstruction framework, experiments were conducted on the High Level Bridge in Edmonton, Canada (see Figure 3). Opened in 1913, the bridge accommodates both Light Rail Transit (LRT) and vehicle traffic with concrete piers, making it a suitable structure for testing defect detection and 3D reconstruction methods.

A DJI Mini 3 Pro drone was employed to collect visual data of the first concrete pier following the approach bridge. The UAV was operated to fly around the pier, capturing continuous 4K video footage to ensure comprehensive coverage of the structure's surface. After the data collection, frames were extracted from the recorded video at a rate of one image per second.

The extracted images were processed using SfM to obtain a sparse 3D point cloud and corresponding camera poses. The SfM pipeline effectively reconstructed the pier's general geometry by identifying and matching feature points across images, followed by incremental camera pose estimation and 3D point triangulation. The resulting sparse point cloud and the associated camera positions are visualized in Figure 4.

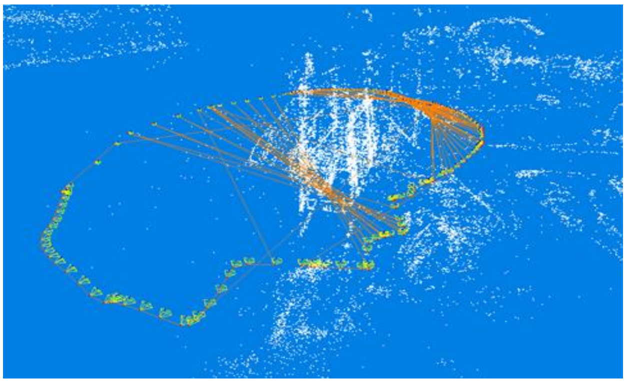

Figure 4. Sparse point cloud with camera pose

To achieve optimal performance in segmenting surface defects on the concrete pier, the Swin UNETR model was trained using domain-specific data. The public dataset Crack5769 [33] was selected for training. This dataset consists of 5,769 pixel-wise labeled concrete crack images, each with a resolution of 256×256 pixels, making it well-suited for the concrete pier context. The dataset was divided into training, validation, and testing sets in a ratio of 8:1:1 to ensure balanced evaluation and prevent overfitting. The model was trained on an Ubuntu 22.04 system equipped with an NVIDIA RTX A6000 GPU. The training was configured with batch size of 16, feature size of 48, learning rate of 0.0001, Adam optimizer and a total of 300 epochs.

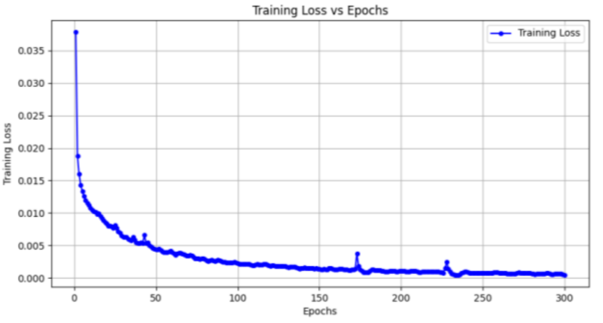

Figure 5. Training loss

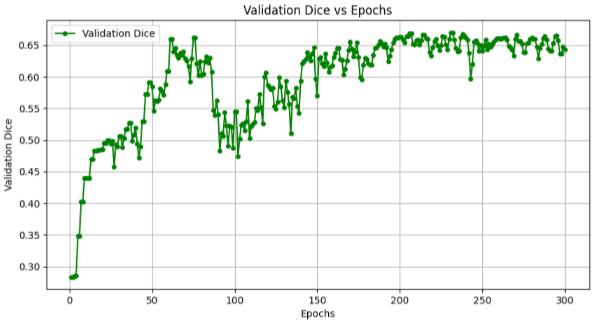

Figure 6. Validation Dice Coefficient

Throughout the training process, the model's performance was monitored by tracking both the training loss and validation Dice coefficient across epochs. Figure 5 illustrates the training loss over 300 epochs, showing a gradual and consistent decrease, which indicates effective model convergence and learning stability. Figure 6 presents the validation Dice coefficient across epochs, demonstrating steady improvement and stabilization in segmentation performance as the training progressed. These plots confirm that the model was learning meaningful features for crack detection without overfitting. Finally, the trained model achieved an average IoU of 0.54, an average Dice coefficient of 0.67, and an average pixel-wise accuracy of 0.69 on the testing set. These results suggest that the Swin UNETR model effectively learned to detect and segment cracks on concrete surfaces.

After achieving the optimal transformer model, the undistorted images from SfM were then passed through the customized Swin UNETR for damage segmentation. The segmentation results were subsequently integrated into the multi-resolution dense reconstruction process.

Using sparse point cloud generated by SfM and the segmentation masks from Swin UNETR, the OpenMVS pipeline was employed to perform multi-resolution dense reconstruction. The process began with depth map estimation, where depth information was calculated for each image to guide point cloud densification, as illustrated in Figure 7.

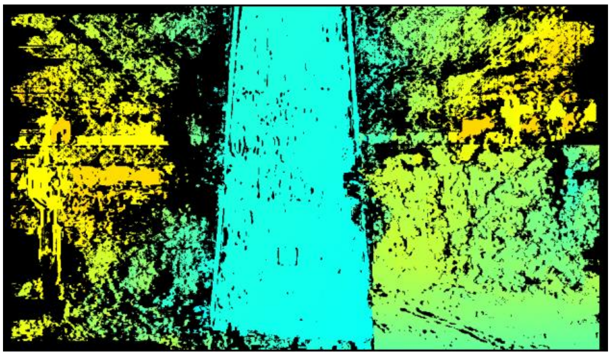

Figure 7. Depth map

The final output consisted of a dense 3D point cloud with varying levels of detail. Structural areas such as flat pier surfaces were reconstructed with 2 times reduced point density, whereas cracks maintained high-density point clouds, as shown in Figure 8.

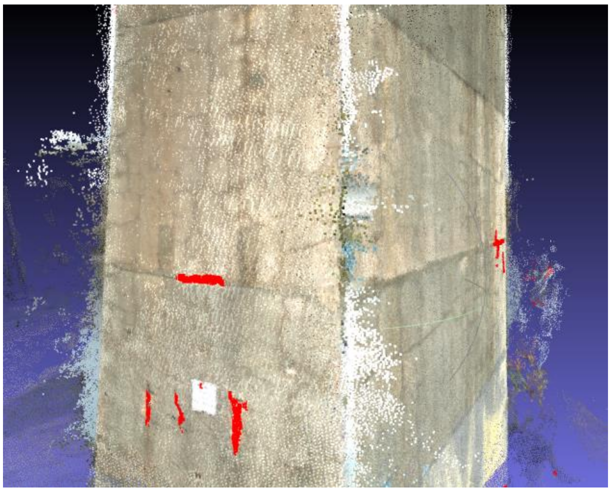

Figure 8. Dense point cloud

To evaluate the efficiency of the proposed method, a runtime comparison was conducted between the uniform-resolution reconstruction and the proposed multi-resolution approach. The results, summarized in Table 1, demonstrate a significant reduction in processing time while maintaining high detail in defect areas. Specifically, the multi-resolution approach achieved substantial reductions in processing time across all reconstruction stages, particularly in depth-maps estimation and point cloud densification. The multi-resolution approach reduced the total reconstruction time by around 64% and decreased memory consumption, while maintaining defect quality. This balance between efficiency and accuracy makes the proposed framework well-suited for large-scale structural health monitoring applications where both accuracy and speed are critical.

Table 1. Runtime Comparison

Action	Normal	Multi-Resolution	Improvement
Estimate Depth-maps	25 m 10 s 221 ms	8 m 3 s 478 ms	↓ 17.1 m
Geometric Consistency	6 m 22 s 885 ms	2 m 28s 319 ms	↓ 3.9 m
Dense fused Depth-maps	1 m 11 s 192 ms	51 s 758 ms	↓ 0.3 m
Densify Point Cloud	39 m 23 s 54 ms	14 m 19 s 721 ms	↓ 25.1 m

Overall, the experimental results validate the effectiveness of the proposed framework in achieving efficient and precise structural damage reconstruction.

4 Conclusion

This study introduces a multi-resolution fast 3D reconstruction framework that integrates UAV-based data collection, transformer-based damage segmentation, and multi-resolution dense reconstruction for efficient and precise structural damage detection. The use of SfM effectively captures the global geometry of the structure, while the customized Swin UNETR model accurately segments surface defects. By prioritizing high-resolution reconstruction in damaged areas and reducing detail in non-critical regions, the proposed framework significantly improves computational efficiency without compromising defect detection quality. Experimental results on the High Level Bridge demonstrated a 64% reduction in processing time and effective defect localization, validating the framework's applicability for structural health monitoring.

However, the current approach has limitations. The model's performance is constrained by the quality and diversity of the training dataset, potentially impacting defect detection under varying lighting and environmental conditions. Additionally, the focus on crack detection limits its generalizability to other defect types, such as spalling, mold or corrosion, etc.

Real-world implementation also presents challenges such as environmental factors affecting UAV inspections, the trade-off between computational efficiency and accuracy, and the need for robust generalization across different defect types and structural conditions. Addressing these limitations, future developments will explore adaptive flight planning, cloud-based computing for real-time processing, and multi-task learning strategies to extend defect detection capabilities. Expanding the framework to identify various defect types, integrating real-time processing and optimizing the system for large-scale infrastructure inspections will further enhance the framework's scalability and practical deployment for long-term structural health monitoring.

Acknowledgement

The authors gratefully acknowledge the support from the NSERC-Alberta Innovates Advance Grants under grant number ALLRP 576826 - 22.

References

[1] L. Sun, Z. Shang, Y. Xia, S. Bhowmick, and S. Nagarajaiah, "Review of Bridge Structural Health Monitoring Aided by Big Data and Artificial Intelligence: From Condition Assessment to Damage Detection," *Journal of Structural Engineering*, vol. 146, no. 5, p. 04020073, May 2020, doi: 10.1061/(ASCE)ST.1943-541X.0002535.

[2] R. Gomasa, V. Talakokula, S. K. R. Jyosyula, and T. Bansal, "Non-destructive Damage Identification of Blended Concrete Systems Using Embedded Piezo Sensors," in *Civil Structural Health Monitoring*, W. Abdullah, M. T. Chaudhary, H. Kamal, J. Parol, and A. Almutairi, Eds., Cham: Springer Nature Switzerland, 2024, pp. 52–61. doi: 10.1007/978-3-031-62253-3_5.

[3] S. Jiang, Y. Zhang, F. Wang, and Y. Xu, "Three-dimensional reconstruction and damage localization of bridge undersides based on close-range photography using UAV," *Measurement Science and Technology*, vol. 36, no. 1, p. 015423, Nov. 2024, doi: 10.1088/1361-6501/ad90fb.

[4] Q. Mei and M. Gül, "Multi-level feature fusion in densely connected deep-learning architecture and depth-first search for crack segmentation on images collected with smartphones," *Structural Health Monitoring*, vol. 19, no. 6, pp. 1726–1744, Nov. 2020, doi: 10.1177/1475921719896813.

[5] Q. Mei, M. Gül, and M. R. Azim, "Densely connected deep neural network considering connectivity of pixels for automatic crack detection," *Automation in Construction*, vol. 110, p. 103018, Feb. 2020, doi: 10.1016/j.autcon.2019.103018.

[6] Y. Ni, J. Mao, H. Wang, Z. Xi, and Z. Chen, "Surface Damage Detection and Localization for Bridge Visual Inspection Based on Deep Learning and 3D Reconstruction," *Structural Control and Health Monitoring*, vol. 2024, no. 1, p. 9988793, Jan. 2024, doi: 10.1155/2024/9988793.

[7] X. Gao and P. Pishdad-Bozorgi, "BIM-enabled facilities operation and maintenance: A review," *Advanced Engineering Informatics*, vol. 39, pp. 227–247, Jan. 2019, doi: 10.1016/j.aei.2019.01.005.

[8] K. Hattori, K. Oki, A. Sugita, T. Sugiyama, and P. Chun, "Deep learning-based corrosion inspection of long-span bridges with BIM integration," *Heliyon*, vol. 10, no. 15, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35308.

[9] X. Pan and T. Y. Yang, "Postdisaster image-based damage detection and repair cost estimation of reinforced concrete buildings using dual convolutional neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 35, no. 5, pp. 495–510, 2020, doi: 10.1111/mice.12549.

[10] C. Xiang, J. Guo, R. Cao, and L. Deng, "A crack-segmentation algorithm fusing transformers and convolutional neural networks for complex detection scenario," *Automation in Construction*, vol. 152, p. 104894, Aug. 2023, doi: 10.1016/j.autcon.2023.104894.

[11] G. Pan, Y. Zheng, S. Guo, and Y. Lv, "Automatic sewer pipe defect semantic segmentation based on improved U-Net," *Automation in Construction*, vol. 119, p. 103383, Nov. 2020, doi: 10.1016/j.autcon.2020.103383.

[12] Z. Fan, H. Lin, C. Li, J. Su, S. Bruno, and G. Loprencipe, "Use of Parallel ResNet for High-Performance Pavement Crack Detection and Measurement," *Sustainability*, vol. 14, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/su14031825.

[13] İ. Akgül, "Mobile-DenseNet: Detection of building concrete surface cracks using a new fusion technique based on deep learning," *Heliyon*, vol. 9, no. 10, Oct. 2023, doi: 10.1016/j.heliyon.2023.e21097.

[14] L. Deng, H. Zuo, W. Wang, C. Xiang, and H. Chu, "Internal Defect Detection of Structures Based on Infrared Thermography and Deep Learning," *KSCE Journal of Civil Engineering*, vol. 27, no. 3, pp. 1136–1149, Mar. 2023, doi: 10.1007/s12205-023-0391-7.

[15] Z. Pan *et al.*, "High-precision segmentation and quantification of tunnel lining crack using an improved DeepLabV3+," *Underground Space*, Dec. 2024, doi: 10.1016/j.undsp.2024.10.002.

[16] H. Chu, W. Wang, and L. Deng, "Tiny-Crack-Net: A multiscale feature fusion network with attention mechanisms for segmentation of tiny cracks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 14, pp. 1914–1931, 2022, doi: 10.1111/mice.12881.

[17] Y. Jin, D. Han, and H. Ko, "TrSeg: Transformer for semantic segmentation," *Pattern Recognition Letters*, vol. 148, pp. 29–35, Aug. 2021, doi: 10.1016/j.patrec.2021.04.024.

[18] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 03, 2021, *arXiv*: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

[19] X. Shen *et al.*, "Self-attentional microvessel segmentation via squeeze-excitation transformer Unet," *Computerized Medical Imaging and Graphics*, vol. 97, p. 102055, Apr. 2022, doi: 10.1016/j.compmedimag.2022.102055.

[20] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds., Cham: Springer International Publishing, 2022, pp. 272–284. doi: 10.1007/978-3-031-08999-2_22.

[21] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Aug. 17, 2021, *arXiv*: arXiv:2103.14030. doi: 10.48550/arXiv.2103.14030.

[22] R. Kakavand, P. Tahghighi, R. Ahmadi, W. B. Edwards, and A. Komeili, "Swin UNETR Segmentation with Automated Geometry Filtering for Biomechanical Modeling of Knee Joint Cartilage," *Annals of Biomedical Engineering*, Jan. 2025, doi: 10.1007/s10439-024-03675-x.

[23] Z. Yu, Y. Shen, and C. Shen, "A real-time detection approach for bridge cracks based on YOLOv4-FPM," *Automation in Construction*, vol. 122, p. 103514, Feb. 2021, doi: 10.1016/j.autcon.2020.103514.

[24] S. Jiang, Y. Cheng, and J. Zhang, "Vision-guided unmanned aerial system for rapid multiple-type damage detection and localization," *Structural Health Monitoring*, vol. 22, no. 1, pp. 319–337, Jan. 2023, doi: 10.1177/14759217221084878.

[25] L. Deng, T. Sun, L. Yang, and R. Cao, "Binocular video-based 3D reconstruction and length quantification of cracks in concrete structures," *Automation in Construction*, vol. 148, p. 104743, Apr. 2023, doi: 10.1016/j.autcon.2023.104743.

[26] C.-Q. Feng, B.-L. Li, Y.-F. Liu, F. Zhang, Y. Yue, and J.-S. Fan, "Crack assessment using multi-sensor fusion simultaneous localization and mapping (SLAM) and image super-resolution for bridge inspection," *Automation in Construction*, vol. 155, p. 105047, Nov. 2023, doi: 10.1016/j.autcon.2023.105047.

[27] J. L. Carrivick, M. W. Smith, and D. J. Quincey, *Structure from Motion in the Geosciences*. John Wiley & Sons, 2016.

[28] P. Moulon, P. Monasse, and R. Marlet, "Adaptive Structure from Motion with a Contrario Model Estimation," in *Computer Vision – ACCV 2012*, K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, Eds., Berlin, Heidelberg: Springer, 2013, pp. 257–270. doi: 10.1007/978-3-642-37447-0_20.

[29] G. Kim and Y. Cha, "3D Pixelwise damage mapping using a deep attention based modified Nerfacto," *Automation in Construction*, vol. 168, p. 105878, Dec. 2024, doi: 10.1016/j.autcon.2024.105878.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," Feb. 07, 2018, *arXiv*: arXiv:1708.02002. doi: 10.48550/arXiv.1708.02002.

[31] H. Ming, Q. Li, H. Xia, and P. Li, "Refined 3D Modeling of Complex Models Based on Stereo Vision," *IEEE Access*, vol. 12, pp. 93487–93501, 2024, doi: 10.1109/ACCESS.2024.3424293.

[32] D. Cernea, "OpenMVS: Multi-View Stereo Reconstruction Library," 2020. [Online]. Available: https://cdcseacave.github.io/openMVS

[33] X. W. Ye, T. Jin, Z. X. Li, S. Y. Ma, Y. Ding, and Y. H. Ou, "Structural Crack Detection from Benchmark Data Sets Using Pruned Fully Convolutional Networks," *Journal of Structural Engineering*, vol. 147, no. 11, p. 04721008, Nov. 2021, doi: 10.1061/(ASCE)ST.1943-541X.0003140.