Structural Semantic Segmentation and Fine-Grained Surface Feature Extraction in Heritage Buildings

Jingxuan Li [1, 2], Jian Bi [3], Vincent J.L. Gan [1, 2], Michael Yit Lin Chew [1]

[1] Department of the Built Environment, National University of Singapore, Singapore
[2] Centre of Digital Building Technology, National University of Singapore, Singapore
[3] College of Mechanical and Vehicle Engineering, Hunan University, China

Abstract

Heritage buildings exhibit complex and irregular legacy environments that pose significant challenges for conservation. This study introduces a framework grounded on high-precision terrestrial laser scan, which serves as input for both semantic segmentation of structural components and fine-grained surface feature extraction. Leveraging the TLS scans, the proposed approach supports 3D scene understanding of heritage buildings, including structural segmentation of walls, floors and ceilings, as well as extracting surface-level details (e.g., cracks) that are layered onto the main structural components. The framework (1) integrates 3D point cloud-based geometric representation of heritage buildings and (2) displays the primary building 3D geometry with updated surface attributes (crack spatial pattern) to uncover new insights. This method improves spatial interpretation in complex heritage environments, offering a platform for analysing potential heritage conservation issues.

Keywords –
Heritage Building; Scene Understanding; Semantic Segmentation; Fine-Grained Surface Feature

1 Introduction

Heritage buildings serve as vital touchstones of cultural identity; yet their conservation faces challenges such as weathering, material aging, and urban development pressures that accelerate deterioration. The shophouse is a predominant heritage building typology in Southeast Asia and Southern China [1, 2]. Singapore is one of the most significant centres for shophouse development [2]. Figure 1 shows a typical shophouse (i.e., NUS Baba House), featuring a ground floor designated for retail or commercial use and upper floors reserved for residence. Compared to modern commercial complexes, shophouse architecture is a less spatially efficient typology for commercial operation. Furthermore, it is characterised by fragile building materials that present unique vulnerabilities to age-related degradation. Such surface conditions may occur independently on specific floors or manifest in combination within particular areas, making the intensive monitoring of shophouse conditions increasingly important to preserve and sustain their cultural and historical significance.

Figure 1. Sample point clouds of shophouse obtained from terrestrial laser scanning

In Singapore, there are over 6,000 conserved shophouses required on-site monitoring and surveying which is a tremendous work [3]. Nowadays, reality capture technologies such as LiDAR or photogrammetry have been adopted to provide a digital archive of the shophouse. But these methods require intensive human labour and time to accurately scan and document the building information and geometry in digital format [4], followed by evaluating the surface conditions and recording relevant conservation issues.

To solve this problem, many studies have attempted to integrate scanning to support as-built data capture as well as automated structural segmentation and fine-grained surface feature extraction (material change, surface pattern etc.). Applying 3D point clouds for building surface feature extraction and assessment (such as cracks) has been attempted by many studies [5-7]. The studies could achieve defect identification accuracy from 42.5% to 99.7%. To achieve this level of accuracy, close-range scanning of the defect area and high-resolution scanned data are simultaneously required [5, 6]. However, digital documentation at such level-of-detail is impracticable for thousands of heritage buildings. For example, a typical 3-storey shophouse has 8 to 12 rooms, which require 16 to 36 scans to generate comprehensive point clouds. Manual labelling of point clouds in this level-of-detail across different structural vulnerabilities like cracks, spalling, moisture infiltration could be tedious and time consuming. More importantly, methods that directly use 3D point cloud data to undertake semantic segmentation or other related tasks are restricted by inherent quality limitations, including the inability to capture local structural details, sensitivity to non-uniform sampling densities, and a lack of translation invariance, which limits their generalisability to complex, irregular geometries often found in heritage sites [8, 9].

As such, more studies nowadays intend to leverage image-based sensing to address the semantic segmentation in heritage buildings. Image segmentation offers several advantages. First, computational efficiency is improved as they require less processing power and memory compared to direct processing of point clouds, making them more efficient and light-weighted for practical applications. Secondly, image-based methods are more flexible as the collection and labelling of images is inherently more straightforward and cost-effective than processing 3D point clouds [10-13]. Additionally, the abundance of open-source dataset, pretrained models and well-established networks provides a robust foundation for transfer learning specific to heritage buildings.

Other studies explored image semantic segmentation with foundational models introducing robust capabilities for segmenting objects in images and videos [14, 15]. Further advancements have focused on open-vocabulary segmentation [11, 12, 16]. Since heritage buildings present more complicated architectural and ornamental elements than modern architectures, adopting open vocabulary models allows the network to understand the various unseen or arbitrary elements present in heritage building environments.

While image-based methods significantly improve efficiency for rapid data acquisition and are easier to deploy at scale, they still face challenges in fully capturing the heterogeneous materials and diverse surface conditions present in historical architecture. Compared with LiDAR-based scanning, image-based methods often exhibit lower geometric fidelity and weaker spatial representation, making it difficult to integrate fine-grained surface details into a precise spatial context, particularly for depth-related features. This limitation reduces their ability to quantify defect dimensions or reveal underlying spatial patterns and their potential impacts for heritage buildings.

This study introduces a framework that utilises high-precision terrestrial laser scanning (TLS) as the primary input for both structural semantic segmentation and fine-grained surface feature extraction. By leveraging TLS, the approach enables comprehensive building 3D scene understanding, encompassing semantic segmentation of structural components like walls, floors, and ceilings alongside fine-grained surface feature extraction that layers onto the structural components. The framework achieves this by (1) integrating point cloud-based 3D geometric representation of heritage buildings and (2) displaying primary building geometry with fine-grained surface attributes. The 3D integration enhances interpretability and enables spatial-level structural analysis by providing a quantified crack dimensions and localization within the building's coordinate system.

2 Methodology

The methodology of this work is shown in Figure 2. For each individual scan, the TLS generates a panoramic colormap for colour point clouds. The colormap contains abundant information about the scene and can be used for 3D scene understanding.

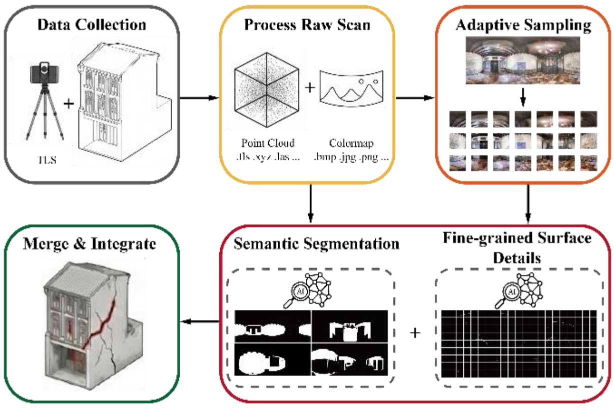

Figure 2. Overview of proposed method

This study uses two pipelines dedicated to processing the colormap to segment structural walls, floors, and ceilings alongside fine-grained surface detail extraction that is layered onto the main structural segmentation results. The colormap overlay embeds surface-level labels, allowing each point to carry both the structural component type and its surface condition through updated surface attributes. The resulting masks from both pipelines are integrated to provide a unified point cloud-based 3D geometric representation of the heritage building, updated with surface attributes on the segmented structural components such as cracks on walls etc. to reveal their spatial patterns. This method can be extended to other surface features such as openings, material transitions and surface patterns, thereby enhancing 3D scene understanding tasks.

2.1 Dataset Collection and Processing

To enable automatic crack identification and semantic segmentation of structural components on panoramic colormap, experiments are conducted on a shophouse to collect its point cloud using TLS. The dataset consists of the following: point clouds, panorama colormaps that are extracted from point clouds, and ground truth colormaps. They are processed to align the requirement of machine learning pipelines.

2.1.1 On-site Scanning

The point cloud dataset is acquired by scanning a heritage shophouse, the NUS Baba House. This building was built around 1895 and undergone several renovations [17]. To capture a comprehensive dataset, the shophouse is scanned from 24 positions (shown in Figure 3) to cover all interior surfaces and façade elevation.

A FARO® TLS was used to acquire the point cloud data. Scanner parameters were set at indoor 10 meters for interior scenes and outdoor 20 meters for exterior and semi-exterior scenes. The acquired raw point clouds are processed and the colormap of each scene is exported. Figure 4 shows the registered point cloud and on-site scanning operation photo.

2.1.2 TLS Panoramic Colormap

The machine learning pipeline for fine-grained surface feature extraction exhibits varying constraints regarding input resolution; while some architectures support native resolution processing via internal slicing, others require fixed dimensions or down-sampling. However, relying on down-sampling is problematic because pixel density is a critical determinant of prediction accuracy; reducing colormap quality inevitably leads to information loss and compromises feature extraction results. Furthermore, preserving the exact original resolution is essential for the subsequent accurate integration of 2D predictions back into the 3D point cloud. Consequently, to bypass the resolution constraints of specific networks and standardize testing across diverse pipelines without sacrificing fidelity, a systematic slicing and tiling approach was implemented. This process partitions the original panoramic colormap into a grid of unified patches. A significant challenge in standard tiling is the "boundary effect," where semantic information is corrupted because object instances (e.g., cracks) traversing tile borders are severed, leading to incomplete feature extraction.

To mitigate this issue, an overlap sampling mechanism is applied. This principle introduces a defined overlapping area between adjacent tiles, ensuring that objects located at boundaries are captured fully within the receptive field of at least one tile, thereby increasing model sensitivity. The tiling logic is mathematically defined as follows: for a colormap of dimensions $w \times h$, with tile size T, overlap distance d, and step size $s = T - d$, each tile is indexed by its column x and row y. Given these parameters, the total number of tiles required to cover this colormap is computed as $Nx = \lceil (w - d) / (T - d) \rceil$ along the horizontal direction and $N_y = \lceil (h - d) / (T - d) \rceil s$ along the vertical direction.

2.1.3 Test Scenes

Indoor scans 07 (1F atrium), 13 (2F front hall), 16 (2F corridor), 19 (3F front hall), 21 (3F staircase), 22 (3F corridor), 23 (3F atrium), 24 (3F rear hall) shown in Figure 4 are selected to test the defect feature extraction methods' performance. These scenes are selected considering the complexity of background context. The simplest scene 19 features one significant crack on an unobstructed wall. Scene 23 contains multiple cracks on multiple walls, while the background context is increased but the defect area is relatively clean. Scene 07 includes many defect types and is captured in higher resolution, bringing challenge to the computing. Scene 24 features insignificant crack and spalling, the background context is extremely complex, introducing many interferences to the scene.

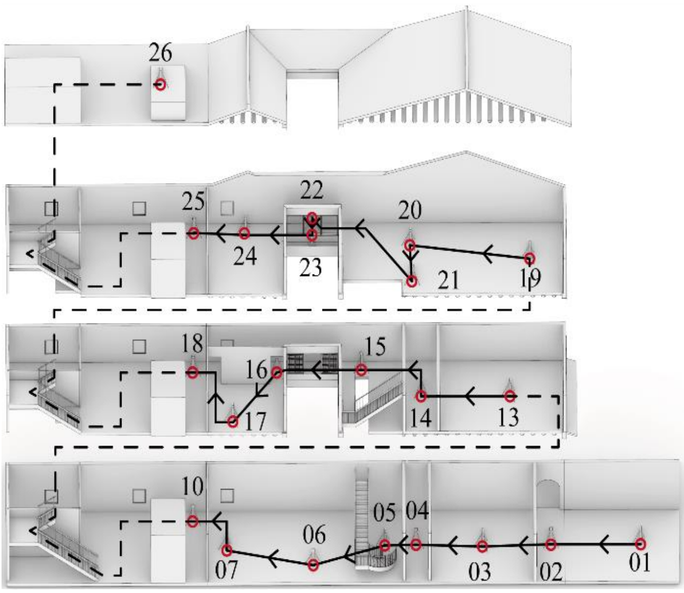

Figure 3. On-site scanning in NUS Baba House

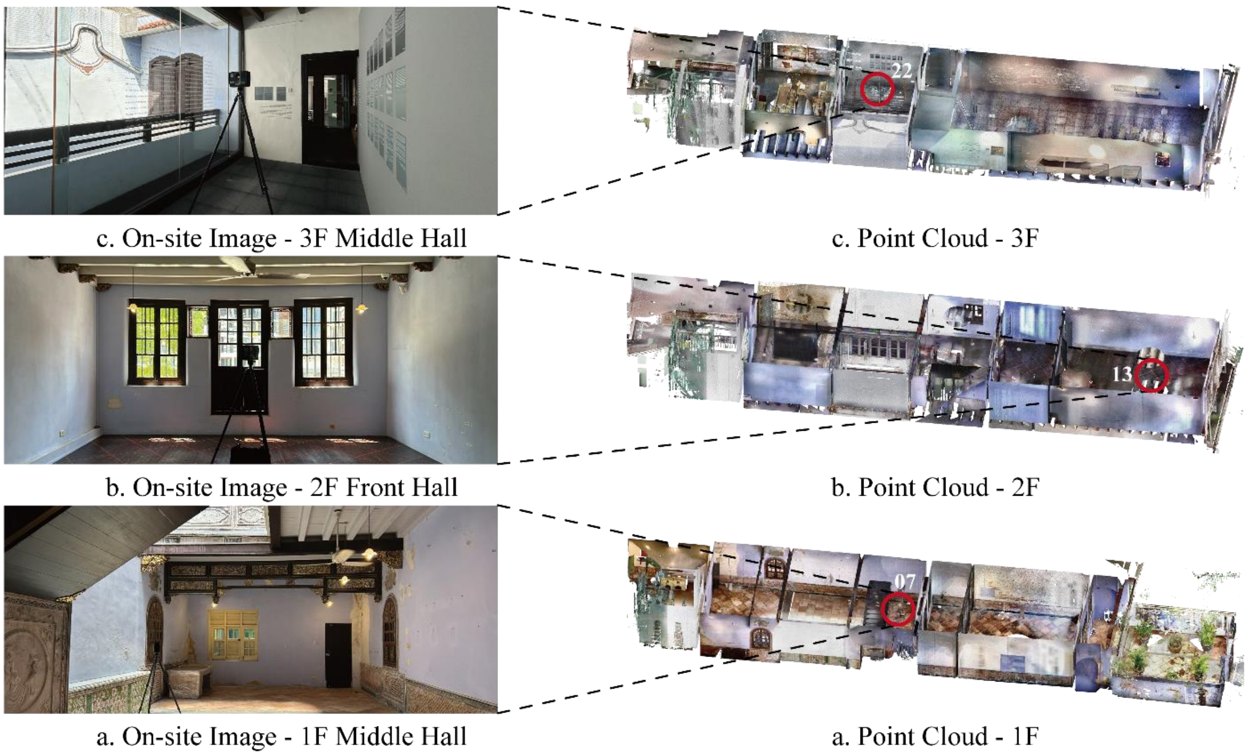

Figure 4. On-site scanning photo and corresponding registered point cloud

2.2 Fine-Grained Surface Feature Extraction

Identification of fine-grained surface features has achieved remarkable performance in structural health monitoring domains [18-22]. Advanced architectures like Vision Transformers (ViTs) significantly outperform traditional CNNs in capturing fine-grained crack details. While widely applied for precise segmentation, these methods remain restricted by application limitations, as existing models often exhibit limited generalisability, struggling to maintain accuracy when facing domain shifts caused by diverse lighting conditions, complex surface textures, or unseen structural materials.

In this study, crack identification is undertaken using multiple pipelines, each applied to tiled colormap. The pipelines generate complementary crack masks that highlight varying crack widths, orientations, and intensities across different surface conditions. The results are analysed using Precision (Pr), Recall (Re), F1-score (F1) and IoU to compare and evaluate all tested pipelines under all scenes and test settings, the metrics are defined in Equations (1)-(4).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{4}$$

2.3 Segmentation of Structural Components

When applying crack identification methods to the panoramic colormaps together with 3D point clouds, numerous false positive predictions are generated due to the complexity and scale of the scanned scenes. Successfully identifying the true positive predictions from the false alarms is critical for increasing the accuracy of crack identification. Incorporating the segmentation of structural components-such as walls and columns-facilitates the removal of false predictions detected in the non-structural area. In this study, the 3D scene is semantically interpreted using a semantic segmentation masking network, whereby structural components are extracted as combined masks to exclude false predictions that are detected on irrelevant elements.

Structural component segmentation is achieved using semantic masking generated by state-of-the-art segmentation networks [16, 23]. These masks restrict the crack identification process to structural regions, effectively suppressing noise and reducing false positives. This approach enables more accurate crack identification across complex indoor environments characterized by occlusions and elements geometrically similar to cracks.

2.4 Crack Identification in Point Clouds

Mapping crack extraction results in 3D point cloud representation provides an intuitive visualisation of the surface conditions, revealing the crack's spatial distribution patterns and their potential structural implications. To achieve this, detected defects are labelled on the panoramic colormap, which replaces the original colormap. Since the colormap is inherently pixel-to-point aligned with the XYZ coordinates, this replacement automatically colourizes the corresponding 3D points to highlight defects. These colour labelled scenes are subsequently registered into a holistic 3D model of the building. Following registration, redundant low-resolution overlapping regions are removed to preserve only the high-resolution in the final point cloud.

3 Results and Discussion

3.1 Crack Identification

Tables 1 and 2 show the Recall and IoU of crack identification [19-22]. Figure 5 shows the results which reveal insight into the proposed tile and merge method. Initial testing of the crack identification revealed a dichotomy in performance. While high accuracy and specificity were observed in 75% of the raw colormap, the models struggled significantly in scenes where cracks pixels are minute relative to total pixels. Despite the successful identification of certain cracks, the overall quantitative metrics for Precision, Recall, and IoU are severely limited by scene complexity. Across all unmasked test scenes, the baseline maximum precision peaked at only 25.44%, with a maximum IoU of 20.75%. These limited results are primarily attributed to an excessive amount of false positive pixels generated from background noise and non-structural elements. Therefore, a robust mechanism is required to filter the noise and retain only valid predictions on relevant building surfaces for subsequent 3D integration. While overlap sampling is applied to preserve the cracks on tile boundaries, its contribution to the correct identification of true defects was limited.

3.2 Structural Wall Segmentation

To address the limitations identified in the baseline crack extraction, a two-stage optimisation strategy involving semantic masking and overlap sampling is implemented. Semantic segmentation networks [16, 23] are utilised to generate semantic masks, as illustrated in Figure 6, based on specific structural prompts, including "wall," "brick wall," "plaster wall," "column," and

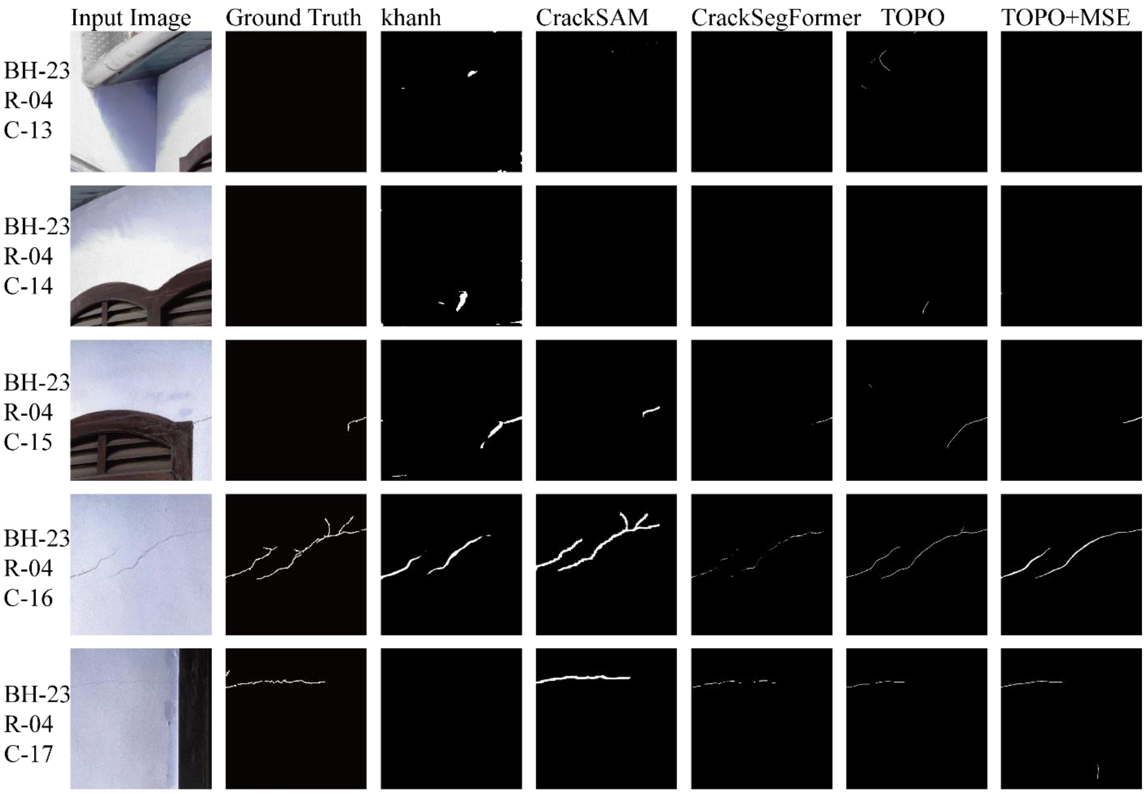

Figure 5 Examples for crack extraction using overlap sampling

"building surface." Among the evaluated networks, masks generated by GroundedSAM2 [16] were selected for its superior stability and prediction accuracy.

Experiment results show that semantic masking and overlap sampling combined approach enhances robustness in prediction. The introduction of semantic masking layer could significantly improve the overall mean precision across all methods from 4.55% to 21.18% and the mean IoU from 3.48% to 10.88%.

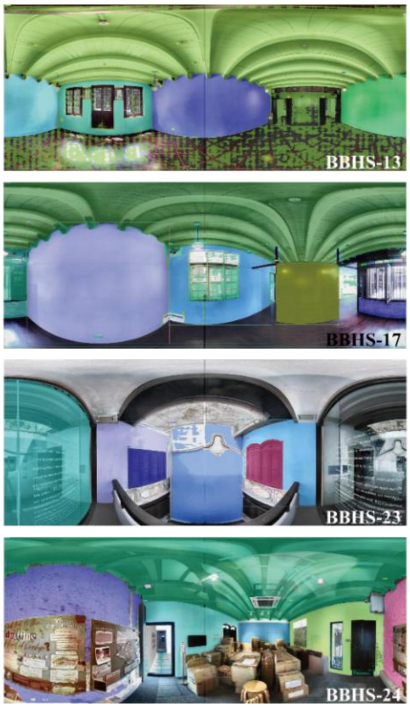

Figure 6. Semantic masks of structural components

The efficiency of this optimisation approach is most evident in complex scenes where baseline prediction generates more false positives. For instance, in test scene bbhs-22 using the CrackSegFormer method, the application of semantic masking combined with overlap sampling raised precision from 8.25% to 63.68%, marking a substantial recovery from the noise-dominated baseline. Similarly, CrackSAM in scene bbhs-21 increased IoU by 37.20% from a negligible 0.33% to 37.53%. These results confirm that introducing semantic masks effectively eliminates false positives on non-structural elements, fundamentally improved the precision and efficacy of the crack extraction pipeline. However, the improvement in Precision and IoU involves a trade-off in Recall. Because the semantic masks are isolating the predicted structural elements, some true defect pixels located at the boundaries of these surfaces could be excluded. While this leads to reductions in Recall depending on the architectural complexity of the scene, the trade-off is considered justifiable for structural health monitoring. Rather than pursuing pixel level accurate detection, the introduction of semantic masks significantly reduces background noise to improve the global detection reliability.

3.3 Surface Attributes in Point Clouds

Figure 7 displays representative results from the pipeline, illustrating fine-grained crack extraction within the segmented structural walls. Beyond visual inspection, the integration of these detections into a 3D point cloud environment enables several critical downstream tasks.

First, it allows quantification of defect length and area, providing a basis for severity evaluation. Second, visualizing surface defects in 3D reveals their spatial

Table 1 Defect extraction metrics - Recall

Detection Method	Test Scene							
	bbhs-07	bbhs-13	bbhs-16	bbhs-19	bbhs-21	bbhs-22	bbhs-23	bbhs-24
No Semantic Segmentation								
CrackSAM	84.18%	45.25%	26.66%	88.51%	79.82%	85.05%	75.42%	33.75%
CrackSegFormer	19.40%	4.23%	0.00%	3.89%	0.00%	26.41%	7.38%	8.37%
khan	48.11%	29.44%	2.42%	86.03%	21.69%	60.16%	15.46%	28.72%
TOPO	22.15%	10.86%	1.18%	24.96%	31.93%	23.23%	16.00%	13.23%
TOPO+MES	29.42%	10.50%	4.89%	50.25%	36.14%	34.00%	20.01%	17.20%
Semantic Segmentation								
CrackSAM	82.49%	44.83%	25.76%	88.51%	74.40%	84.33%	58.92%	25.58%
CrackSegFormer	19.16%	4.23%	0.00%	3.89%	0.00%	26.29%	6.64%	5.15%
khan	47.93%	29.44%	2.25%	86.03%	21.69%	60.03%	14.27%	18.21%
TOPO	22.14%	10.70%	1.18%	24.96%	31.93%	23.15%	14.65%	8.28%
TOPO+MES	29.42%	10.50%	3.32%	50.25%	36.14%	33.96%	19.01%	11.07%
Semantic Segmentation + Overlap Sampling								
CrackSAM	81.24%	47.65%	11.19%	99.83%	79.52%	89.45%	57.85%	27.48%
CrackSegFormer	21.24%	5.82%	0.51%	4.23%	0.00%	28.17%	10.16%	3.90%
khan	44.69%	31.39%	4.22%	86.31%	18.98%	59.15%	14.21%	18.04%
TOPO	23.37%	11.31%	1.01%	36.79%	31.93%	25.39%	15.76%	8.38%
TOPO+MES	30.49%	10.50%	8.10%	54.37%	36.14%	36.07%	20.24%	11.16%

Table 2 Defect extraction metrics - IoU

Detection Method	Test Scene							
	bbhs-07	bbhs-13	bbhs-16	bbhs-19	bbhs-21	bbhs-22	bbhs-23	bbhs-24
No Semantic Segmentation								
CrackSAM	0.80%	1.95%	0.36%	1.62%	0.33%	20.75%	13.44%	3.58%
CrackSegFormer	0.30%	0.26%	0.00%	0.29%	0.00%	6.95%	3.67%	2.59%
khan	0.21%	0.24%	0.01%	0.75%	0.02%	4.47%	0.48%	1.93%
TOPO	1.71%	0.34%	0.04%	2.42%	0.70%	10.71%	7.57%	4.89%
TOPO+MES	2.13%	0.65%	0.14%	4.70%	0.74%	17.03%	10.55%	6.90%
Semantic Segmentation								
CrackSAM	4.35%	22.04%	1.07%	9.18%	37.54%	24.75%	23.79%	13.04%
CrackSegFormer	5.53%	3.60%	0.00%	3.26%	0.00%	22.08%	6.27%	4.06%
khan	4.07%	10.89%	0.04%	16.76%	1.94%	9.24%	12.12%	4.69%
TOPO	12.35%	9.04%	0.18%	17.19%	9.49%	14.70%	12.89%	6.72%
TOPO+MES	11.05%	7.68%	0.38%	29.18%	12.86%	22.41%	16.28%	8.11%
Semantic Segmentation + Overlap Sampling								
CrackSAM	4.66%	24.62%	0.41%	14.01%	31.02%	25.92%	24.70%	12.89%
CrackSegFormer	4.41%	4.57%	0.05%	2.94%	0.00%	24.27%	9.53%	3.34%
khan	3.54%	10.72%	0.07%	14.97%	1.54%	8.02%	11.58%	4.41%
TOPO	12.30%	9.44%	0.14%	23.90%	8.85%	15.53%	13.75%	6.75%
TOPO+MES	10.52%	7.79%	0.77%	28.33%	11.93%	23.06%	16.82%	7.60%

distribution across the building, extending analysis from isolated scenarios to a comprehensive assessment. Ultimately, the 3D point cloud serves as foundation for BIM and digital-twin platforms, supporting efficient management and long-term building condition updating.

Figure 7. Point clouds of scan19 with cracks

4 Conclusions

This study presents a structural semantic segmentation and fine-grained surface feature extraction for heritage buildings. The proposed method addresses the inherent challenges of conducting structural health assessments on point cloud data captured in complex heritage environments. By incorporating semantic understanding of the scanned scenes, the method isolates structural components, refining crack identification on 3D point cloud data. The research was validated through experiments on a shophouse, benchmarking automated predictions against human visual inspections. Experimental results highlight the feasibility of TLS-derived point clouds for structural health assessment, confirming improvements in defect identification metrics in real-world heritage environment.

The proposed method has certain limitations for improvement. As the prediction was deployed using default parameters and generic datasets, fine-tuning these models on a dedicated heritage defect dataset could significantly bolster identification performance and robustness against domain-specific noise. Furthermore, current state-of-the-art networks predominantly focus on single-defect extraction (i.e., cracks), often yielding suboptimal results for multi-defect scenarios. Integrating diverse defect datasets with open-vocabulary semantic understanding models offers a promising pathway toward a holistic, multi-objective structural health assessment framework. Finally, while the high-resolution data provided by TLS is essential for detecting minute cracks, it introduces a trade-off in computational efficiency, the inherent operational constraints of TLS also limit its scalability in rapid field deployments. To bridge the gap between precision and operational practicality, future research will adapt this framework for data derived from mobile SLAM. Successfully adapting this pipeline to mobile data would not only mitigate the operational limitations of data acquisition, but also ready the framework for efficient structural health monitoring.

Acknowledgements

The authors would like to thank NUS Museum, Baba House, and the team for granting access and supporting the 3D scanning work. The authors also thank Centre of Digital Building Technology for supporting the experiment. This research is supported by Singapore MOE AcRF Tier-1 Grant (Ref: A-8004446-00-00). Any opinions, findings and conclusions or recommendations expressed in the material are those of the author(s) and do not reflect the views of the grantors.

References

[1] W. Han and J. Beisi. Urban morphology of commercial port cities and shophouses in Southeast Asia. Procedia Engineering, 142:190–197, 2016.

[2] M. Tohiguchi and H. S. Chong. Shophouse: Asian urban composite housing. In Planning for a Better Urban Living Environment in Asia, pages 212–230, Routledge, London, 2000.

[3] Urban Redevelopment Authority. Master Plan 2019 SDCP Conserved Building layer (2024). Data.gov.sg. On-line: https://data.gov.sg/datasets/d_3d73d61e5f5fe3ce52ac5d86b395d0b2/view, Accessed: 04/12/2025.

[4] M. Buldo, L. Agustín-Hernández, and C. Verdoscia. Semantic enrichment of architectural heritage point clouds using artificial intelligence: The Palacio de Sástago in Zaragoza, Spain. Heritage, 7(12):6938–6965, 2024.

[5] M. E. Mohammadi, R. L. Wood, and C. E. Wittich. Non-temporal point cloud analysis for surface damage in civil structures. ISPRS International Journal of Geo-Information, 8(12):527, 2019.

[6] Y. Jing, J.-X. Zhong, B. Sheil, and S. Acikgoz. Anomaly detection of cracks in synthetic masonry arch bridge point clouds using fast point feature histograms and PatchCore. Automation in Construction, 168:105766, 2024.

[7] Y. Jing, W. Lin, B. Sheil, and S. Acikgoz. 3D multimodal feature for infrastructure anomaly detection. Automation in Construction, 178:106388, 2025.

[8] C. R. Qi, Y. Li, H. Su, and L. J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. arXiv preprint arXiv:1706.02413, 2017.

[9] R. Pierdicca, M. Paolanti, F. Matrone, M. Martini, C. Morbidoni, E. S. Malinverni, E. Frontoni, and A. M. Lingua. Point cloud semantic segmentation using a deep learning framework for cultural heritage. Remote Sensing, 12(6):1005, 2020.

[10] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin. PointLLM: Empowering large language models to understand point clouds. arXiv preprint arXiv:2308.16911, 2024.

[11] P. Nguyen, T. D. Ngo, E. Kalogerakis, C. Gan, A. Tran, C. Pham, and K. Nguyen. Open3DIS: Open-vocabulary 3D instance segmentation with 2D mask guidance. Open3DIS: Open-vocabulary 3D instance segmentation with 2D mask guidance. arXiv preprint arXiv:2312.10671, 2024.

[12] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 815–824, Vancouver, Canada, 2023.

[13] Y. Yang, X. Wu, T. He, H. Zhao, and X. Liu. SAM3D: Segment anything in 3D scenes. arXiv preprint arXiv:2306.03908, 2023.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021.

[15] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. K. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. B. Girshick, P. Dollár, and C. Feichtenhofer. SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024.

[16] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang. Grounded SAM: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159, 2024.

[17] NUS Museum. About NUS Baba House. On-line: https://babahouse.nus.edu.sg/explore/about-baba-house/, Accessed: 04/12/2025.

[18] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li. DeepCrack: A deep hierarchical feature learning architecture for crack segmentation. Neurocomputing, 338:139–153, 2019.

[19] C. Benz, P. Debus, H. K. Ha, and V. Rodehorst. Crack segmentation on UAS-based imagery using transfer learning. In Proceedings of the 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ), pages 1–6, Dunedin, New Zealand, 2019.

[20] K. Ge, C. Wang, Y. T. Guo, Y. S. Tang, Z. Z. Hu, and H. B. Chen. Fine-tuning vision foundation model for crack segmentation in civil infrastructures. Construction and Building Materials, 431:136573, 2024.

[21] B. G. Pantoja-Rosero, D. Oner, M. Kozinski, R. Achanta, P. Fua, F. Perez-Cruz, and K. Beyer. TOPO-Loss for continuity-preserving crack detection using deep learning. Construction and Building Materials, 344:128264, 2022.

[22] H. Li, H. Zhang, H. Zhu, K. Gao, H. Liang, and J. Yang. Automatic crack detection on concrete and asphalt surfaces using semantic segmentation network with hierarchical Transformer. Engineering Structures, 307:117903, 2024.

[23] L. Medeiros. Language Segment-Anything. GitHub Repository. On-line: https://github.com/luca-medeiros/lang-segment-anything, Accessed: 04/12/2025