# AUTOMATED IMAGE-BASED STRUCTURE ASSESSMENT USING SFM-VLM AND UAVS

**Phromphat Thansirichaisree[1,*] and Apichat Buatik[1]**

*[1] Thammasat Research Unit in Infrastructure Inspection and Monitoring, Repair and Strengthening (IIMRS), Faculty of Engineering, Thammasat School of Engineering, Thammasat University Rangsit, Klong Luang, Pathumthani 12120, Thailand*

*Corresponding author: tphromph@engr.tu.ac.th

## 1. INTRODUCTION

Structural health monitoring (SHM) plays a critical role in ensuring the safety, reliability, and longevity of infrastructure such as bridges, buildings, and industrial facilities. Conventional visual inspections are manual, time-consuming, and often hazardous, requiring skilled inspectors to access difficult or dangerous locations. This process is not only inefficient but also prone to human error and inconsistency.

Recent advances in Unmanned Aerial Vehicles (UAVs), Structure-from-Motion (SfM) photogrammetry, and deep learning have enabled automated, high-resolution 3D reconstruction of structural surfaces. However, while SfM provides accurate geometric mapping, it lacks semantic interpretation of defects. Conversely, AI-driven defect detection systems often work in 2D image space and struggle to integrate spatial context or align results over large structural areas. Misalignments, incomplete mapping, and inability to generate actionable 3D inspection reports remain common limitations.

In this work, we propose an integrated UAV–SfM–AI pipeline for Automated Image-Based Structure Assessment. The Stacking Convolutional Neural Network (S-CNN) is inspired by pixel-level crack mapping systems, but adapted to operate directly in the texture space of SfM-derived 3D models. This approach eliminates the common misalignment issues seen in multi-image defect mapping and produces accurate, scalable results suitable for engineering reports. As future work, we plan to extend this architecture with Vision-Language Models (VLMs) to provide automatic, context-aware defect descriptions. Figure 1 illustrates the system overview and contribution of this paper.

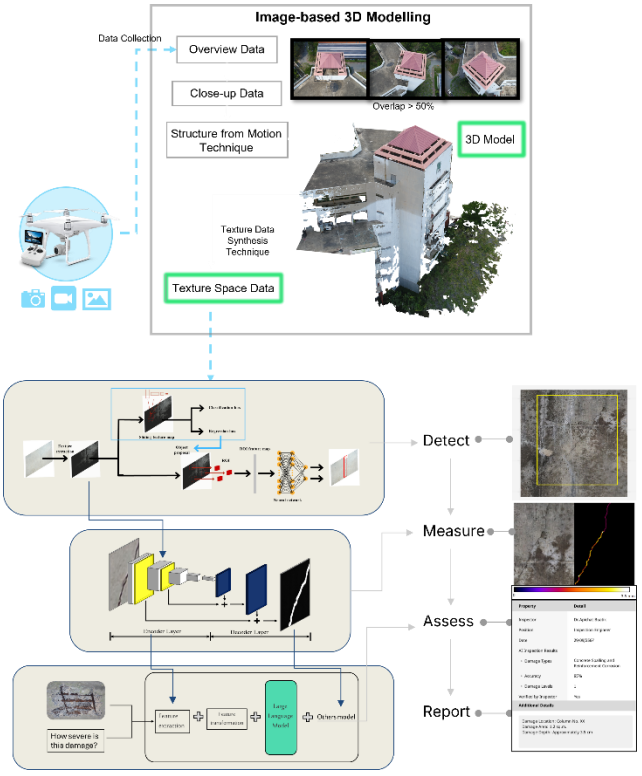


Figure 1. System overview and contribution of this paper

## 2. PROBLEM STATEMENT

Although recent research shows that deep learning can achieve high precision in defect detection, practical deployment for large-scale infrastructure faces three core challenges:

- Alignment Problems: Defect detections from multiple 2D images are difficult to stitch seamlessly into a large-area report due to perspective and geometric distortion.
- Limited 3D Integration: Many systems fail to take advantage of the full 3D model for mapping results, leading to context loss.
- Semantic Gap: While the detection is automated, engineering interpretation of the defects still requires manual effort.

The proposed S-CNN pipeline addresses the first two issues by detecting defects directly in the SfM texture space, generating coherent large-area crack maps without misalignment. The semantic interpretation gap will be addressed in the next research phase by extending S-CNN with VLMs for description and classification.

## 3. METHODOLOGY

3.1 Data Acquisition

The proposed S-CNN pipeline addresses the first two issues by detecting defects directly in the SfM texture space, generating coherent large-area crack maps without misalignment. The semantic interpretation gap will be addressed in the next research phase by extending S-CNN with VLMs for description and classification.

3.2 3D Reconstruction with SfM

The captured images are processed using a Structure-from-Motion (SfM) workflow, implemented in software such as Agisoft Metashape [1], to produce a geometrically accurate 3D model of the inspected structure. SfM algorithms detect and match feature points across overlapping images to estimate camera positions and reconstruct the scene geometry. The process generates a dense point cloud, which is then converted into a surface mesh representing the object's shape. High-resolution textures are mapped onto this mesh to preserve fine surface details, providing both spatial accuracy and visual fidelity [5-11]. RMS re-projection errors are monitored and maintained below one pixel to minimize distortion, ensuring reliable localization of defects in the subsequent texture-space analysis.

3.3 Texture-Space Defect Detection with S-CNN

Following the creation of the high-resolution textured 3D model, defect detection is performed directly within the texture space rather than on individual images. This approach eliminates the alignment issues often encountered when reprojecting detections from multiple viewpoints. The proposed Stacking Convolutional Neural Network (S-CNN) [2] operates in two sequential stages. First, a classification stage identifies image patches within the texture space that are likely to contain defects. Second, a segmentation stage processes only the positively classified patches to generate pixel-level defect maps. This staged process reduces false positives, improves computational efficiency, and maintains high precision. The resulting defect maps are inherently spatially coherent, enabling seamless integration with the 3D model for visualization and reporting.

3.4 Visualization & Reporting

The defect maps generated in the texture space are projected back onto the 3D model using equation 1, creating a spatially accurate and visually intuitive representation of the inspection results. This integrated model can be explored interactively as a digital twin, allowing inspectors to examine defect locations in their real-world context. The system also supports automated generation of certified inspection reports in PDF format, which include defect imagery, location data, and quantitative metrics such as defect size and extent. For integration into engineering workflows, the annotated 3D models can be exported in standard formats such as OBJ, IFC, or other BIM-compatible files, enabling seamless use within asset management and maintenance planning systems.

$$m' = K\,[\,R \mid t\,]\,M' \qquad (1)$$

where $M'$ is are the coordinates of a 3D point in the world coordinate space. $m'$ are the coordinates of the projection point in pixels. $K$ is a camera matrix or a matrix of intrinsic parameters. The rotation-translation matrix [ R | t ] is called an extrinsic matrix. It is used to describe the camera motion around a static scene.

## 4. RESULTS

Field tests on reinforced concrete structures demonstrated that the proposed SfM–S-CNN pipeline achieves high accuracy and efficiency in large-area defect detection. Pixel-level accuracy reached 99.88%, with a precision of 82.2%, recall of 90.2%, and an F1-score of 86.0%. The texture-space approach eliminated the stitching errors typically associated with multi-view projection methods, producing defect maps that were spatially coherent and seamlessly integrated into the 3D model, as show in Figure 2. With a Ground Sampling Distance of 1 mm/pixel, the system captured fine structural details, enabling reliable detection of narrow cracks and other small-scale defects. Processing time was reduced by more than 70% compared to traditional manual inspection workflows, highlighting the method's scalability for large structures. While the current work focuses on spatially accurate defect detection, the coherent output generated by S-CNN provides an ideal foundation for future integration with Vision-Language Models (VLMs). This planned extension will enable automatic classification, descriptive reporting, and context-aware interpretation of detected defects, further reducing the reliance on manual engineering assessment. Table 1 shows the results of crack segmentation on a footing compared between 3 systems, (1) the FCN (traditional segmentation using deep learning), (2) pre-trained DeepCrack [3-4], and (3) the proposed system (S-CNN).

Figure 2. The finalized crack map – the composition of crack pixels onto the texture space

Figure 3. Shows an example of the application of this research in real-world structural inspection scenarios

Table 1. Summary results of crack segmentation on a concrete footing, compared between different systems, (1) FCN, (2) DeepCrack, and (3) S-CNN

| Parameters | FCN | DeepCrack [[3], [4]] | S-CNN (proposed system) |
|---|---|---|---|
| Accuracy | 0.9737 | 0.9930 | 0.9988 |
| Precision | 0.1279 | 0.3384 | 0.8220 |
| Recall | 0.9116 | 0.7015 | 0.9020 |
| F1 Score | 0.2243 | 0.4565 | 0.8601 |

## 5. CONTRIBUTIONS

This paper advances the field of automated structural inspection by integrating UAV-based data acquisition, SfM-based 3D reconstruction, and texture-space deep learning into a robust, high-accuracy assessment pipeline. The proposed approach addresses persistent challenges such as misalignment in multi-view projections, difficulty in scaling to large structures, and the lack of semantic readiness for automated reporting. It is applied to real-world damage inspections of high-rise buildings, as illustrated in Figure 3. Specifically, this work introduces:

- S-CNN for Texture-Space Analysis: An adaptation of deep convolutional stacking for accurate, large-area, pixel-level defect detection directly in SfM texture space.
- Misalignment-Free Mapping: A unified detection space eliminates common projection errors in UAV-based inspection workflows.
- Foundation for Semantic Extension: S-CNN provides the spatially coherent defect maps required for future integration with VLMs.

## 6. FUTURE WORK

The next research phase will integrate Vision-Language Models with S-CNN outputs, enabling:

- Automatic defect classification into engineering categories (cracks, spalling, corrosion).
- Natural-language descriptions of defects with severity assessment.
- Structured outputs for asset management systems.

## 7. CONCLUSION

The proposed SfM–S-CNN pipeline enables accurate, efficient, and scalable automated image-based structural assessment. Operating directly in texture space avoids the misalignment issues of multi-view stitching, while S-CNN achieves high pixel-level accuracy. With the planned extension to VLMs, the system will evolve into a fully autonomous inspection assistant, capable of both detecting and describing defects in engineering terms.

## 3. REFERENCE

[1] Agisoft Metashape, Website, Access Date: 2 February 2022. https://www.agisoft.com.

[2] Chaiyasarn, K., Buatik, A., Mohamad, H., Zhou, M., Kongsilp, S., & Poovarodom, N. (2022). Integrated pixel-level CNN-FCN crack detection via photogrammetric 3D texture mapping of concrete structures. Automation in Construction, 140, 104388. https://doi.org/10.1016/j.autcon.2022.104388.

[3] Y. Liu, J. Yao, X. Lu, R. Xie, L. Li, DeepCrack: a deep hierarchical feature learning architecture for crack segmentation, Neurocomputing 338 (2019) 139–153, https://doi.org/10.1016/j.neucom.2019.01.036.
[4] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, Y. Liu, RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from highresolution remotely sensed images, IEEE Trans. Geosci. Remote Sens. 57 (4) (2018) 2043–2056, https://doi.org/10.1109/TGRS.2018.2870871.
[5] H. Hoppe, Poisson surface reconstruction and its applications, Proc. 2008 ACM Symp. Solid Phys. Model. (2008) 1–10, https://doi.org/10.1145/1364901.1364904.
[6]D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110, https://doi.org/10.1023/B:VISI.0000029664.99615.94.
[7] A.M. Andrew, Multiple View Geometry in Computer Vision, Kybernetes, 2001, https://doi.org/10.1108/k.2001.30.9_10.1333.2.
[8] D. Nist´er, An efficient solution to the five-point relative pose problem, IEEE Trans. Pattern Anal. Mach. Intell. 26 (6) (2004) 756–770, https://doi.org/10.1109/TPAMI.2004.17.
[9] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395, https://doi.org/10.1145/358669.358692.
[10] Y. Furukawa, B. Curless, S.M. Seitz, R. Szeliski, Towards internet-scale multi-view stereo, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. (2010) 1434–1441, https://doi.org/10.1109/CVPR.2010.5539802.
[11] H. Hoppe, Poisson surface reconstruction and its applications, Proc. 2008 ACM Symp. Solid Phys. Model. (2008) 1–10, https://doi.org/10.1145/1364901.1364904.