Assessment of Rock and Stone Decay in Heritage Sites Using Machine Learning

Wendi Ying[1], Kourosh Khoshelham[2], Jonathan Kemp[3]

[1] School of Computing and Information Systems, The University of Melbourne, Melbourne, Victoria, Australia
[2] Department of Infrastructure Engineering, The University of Melbourne, Melbourne, Victoria, Australia
[3] Grimwade Centre for Cultural Materials Conservation, The University of Melbourne, Melbourne, Victoria, Australia
E-mails: wyying@student.unimelb.edu.au, (k.khoshelham, jonathan.kemp)@unimelb.edu.au

Keywords: SAM, Segmentation, 3D Reconstruction, MVS, Cultural Heritage.

Abstract

Cultural heritage sites face growing threats from environmental factors and human activities, highlighting the need for efficient techniques to monitor and preserve their structural integrity. While advanced machine learning models, such as Segment Anything Model (SAM), have shown success in areas such as healthcare, their potential for cultural heritage conservation remains largely unexplored. In this research, we propose an automatic decay detection and visualization framework by combining advanced segmentation techniques with 3D reconstruction methods. We fine-tune SAM and integrate it with You Only Look Once (YOLO) to create a fully automatic, real-time segmentation framework that offers strong generalization for identifying unseen decay types. By incorporating Structure from motion (SfM) and multi-view stereo (MVS), the framework produces 3D models that highlight decay regions, providing a robust tool for structural assessment and visualization. Through both quantitative and qualitative evaluations, we show that our approach outperforms several state-of-the-art models, demonstrating its effectiveness in identifying and visualizing stone decay. Our results contributes to heritage preservation by providing a novel, scalable solution for real-time monitoring of cultural heritage sites.

1. Introduction

Cultural heritage buildings and sites made of stone and rock represent invaluable resources that not only foster tourism and economic development but also preserve historical culture and customs from ancient to recent times (Timothy, 2014, Fusco Girard and Vecco, 2021). However, these sites face degradation due to human activities, natural disasters, climate change, and microbial deteriorations (De la Fuente et al., 2013, Spennemann and Graham, 2007, Liu et al., 2020), making their conservation a crucial and urgent work. Traditional conservation condition surveys rely on manual assessment, which is often costly and time-consuming (Agdas et al., 2016). Recently, unmanned aerial vehicles (UAVs) have enhanced the efficiency and safety of heritage inspection (Tan et al., 2021, Liu et al., 2021, Grosso et al., 2020), yet they primarily serve image collection rather than evaluation of the state of building or site preservation. This situation highlights the need for innovative approaches that enable effective and automated analysis of UAV-captured images.

A key task in heritage conservation is the identification of decay regions on stone and rock surfaces. While sensor-based techniques offer time savings over traditional visual inspection, they require high-cost and complex setups for equipment (Gopinath and Ramadoss, 2021). In contrast, deep learning-based image segmentation can utilize UAV-captured images to automatically generate different masks within digital imagery to represent areas of specific features such as stone decay areas, which improves both efficiency and safety (Minaee et al., 2021, Nooralishahi et al., 2022). Recent advancements in deep learning, including models like U-Net and Mask R-CNN, have achieved notable success in heritage conservation but often require extensive data for retraining when a new decay type is introduced, limiting their scalability (Chen et al., 2021, Hatır et al., 2021, Bruno et al., 2023, Hou et al., 2024). Moreover, the Segment Anything Model (SAM) demonstrates potential for cultural heritage applications due to its zero-shot ability and generalization capabilities in areas such as healthcare and remote sensing (Ma et al., 2024, Osco et al., 2023).

Furthermore, 3D reconstruction techniques like Multi-View Stereo (MVS), Neural Radiance Fields (NeRF), and 3D Gaussian Splatting (3D GS) provide detailed models of heritage structures, but they lack automated decay assessment features. Most existing approaches primarily focus on the geometry and texture of the building or rock heritage, hence further manual evaluations for defect regions are required (Ma and Liu, 2018, Welponer, 2019). Integrating segmentation results into these models could enable comprehensive visualizations of decay on 3D heritage models, facilitating long-term monitoring.

In this research, we propose a framework for real-time segmentation and 3D visualization of decay regions on built and rock heritage structures using a fine-tuned SAM and YOLOv8 model combined with 3D reconstruction through SfM and MVS, as shown in Figure 1. The captured images of a heritage site are processed to generate decay segmentation masks, which are then mapped onto a 3D model with decay areas highlighted in different colors. The main contribution of this study can be summarized as follows: (1) We introduce a deep learning-based framework for real-time decay segmentation with better accuracy and generalization ability, and (2) We develop a 3D reconstruction pipeline that visually highlights and distinguishes various decay types on the surface of historical buildings and rock heritage sites, aiding targeted conservation.

2. Related Works

In this section, we discuss related works in image segmentation techniques applied to cultural heritage datasets, with an emphasis on approaches that improve segmentation accuracy and

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-1019-2025

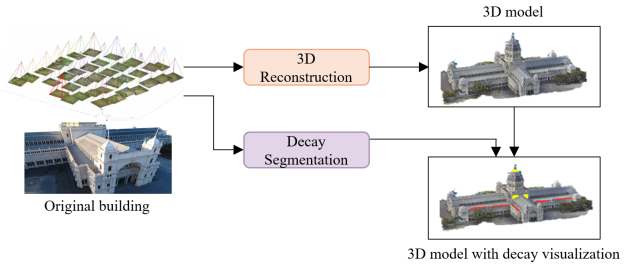


Figure 1. The proposed framework for segmentation and 3D visualization of decay on heritage sites.

adaptability. We also explore prior studies in 3D reconstruction methods tailored for heritage sites.

2.1 Image Segmentation for Heritage Conservation

Image segmentation represents a central challenge in computer vision, which primarily partitions the input image into meaningful regions (masks) for diverse applications including medical imaging, autonomous driving, and video surveillance (Minaee et al., 2021). Based on the classifying purpose, it can be divided into semantic segmentation, which focuses on categorizing each pixel in the image into the same classes, and instance segmentation, which further distinguishes each object within the class. Various deep learning models have been proposed, such as U-Net, U-Net++, and Mask R-CNN (Ronneberger et al., 2015, Zhou et al., 2018, He et al., 2017). These methods have shown effectiveness in detecting decay types within the cultural heritage, including cracks, scaling, and biological colonization (Chen et al., 2021, Hatır et al., 2021, Bruno et al., 2023). However, these models significantly depend on diverse training data for accurate results, presenting challenges when facing unseen decay types that require additional data for retraining. To address this limitation, the solutions could involve compiling a comprehensive dataset or developing methods with superior generalization and adaptability.

2.2 Segment Anything Model

The Segment Anything Model (SAM) (Kirillov et al., 2023) overcomes the limitation of previous deep learning segmentation methods by introducing a prompt-based approach that utilizes points, bounding boxes, or text to specify segmentation targets. SAM's zero-shot capability enables it to segment unseen objects by transferring knowledge obtained from training on a massive and diverse dataset, guided by prompts (Larochelle et al., 2008, Pourpanah et al., 2022). Optimized for real-time performance (around 50ms in a web browser), SAM ensures smooth and interactive prompting. Moreover, variants like FastSAM, MobileSAM, and SAM2, as well as domain-specific versions such as MedSAM for medical imaging and gazeSAM for eye tracking technologies, enhance accuracy in specialized applications (Zhao et al., 2023, Zhang et al., 2023, Ravi et al., 2024, Ma et al., 2024, Wang et al., 2023). These advances highlight SAM's potential in heritage preservation, especially for segmenting stone decays. Additionally, by using object detection models such as YOLOv8 (Jocher et al., 2023) to generate bounding box prompts for SAM, segmentation can focus precisely on specific decay regions, which reduces the need for manual input and enhances accuracy.

2.3 3D Reconstruction for Heritage Sites

Three-dimensional (3D) reconstruction is a core task in computer vision, involving the process of generating 3D models of objects from data collected by cameras, sensors, or other equipment (Ma and Liu, 2018). It has been applied in fields such as bone reconstruction in medical engineering and digital preservation of historical buildings in civil engineering. Approaches are broadly categorized as image-based or scanner-based (Verykokou and Ioannidis, 2023). Among image-based methods, techniques such as Multi-View Stereo (MVS), Neural Radiance Fields (NeRF), and 3D Gaussian Splatting (3D GS) have significantly improved the accuracy and quality of digital heritage preservation (Furukawa et al., 2015, Mildenhall et al., 2021, Kerbl et al., 2023). While MVS, NeRF, and 3D GS have been applied to cultural conservation, existing studies primarily focus on texture quality rather than condition monitoring (Mazzacca et al., 2023, Murtiyoso et al., 2024, Basso et al., 2024), suggesting a gap for more comprehensive analysis and frameworks aimed at long-term monitoring the condition of heritage buildings. This gap also motivates further research to explore how these 3D reconstruction techniques can support the preservation of cultural heritage sites.

3. Methodology

Our method consists of two stages: first, decay regions are segmented by the fine-tuned SAM using YOLOv8-generated prompts to guide segmentation specifically toward decay regions. Second, MVS is employed to construct a dense 3D model of the heritage site, where segmented decay areas are highlighted.

3.1 Fine-tuned SAM with YOLO Prompts

We fine-tune the recent SAM (Kirillov et al., 2023) model with ViT-B as its backbone on a specific stone decay dataset. Due to the limited size of our dataset, which is insufficient for a full re-training of SAM since deep learning models typically require large datasets (the original SAM was trained on 11 million images), fine-tuning serves as an effective transfer learning technique. To streamline resources, we freeze the image and prompt encoders and iteratively fine-tune only the mask decoder, as shown in Figure 2. This strategy enhances the model's ability to accurately identify different types of stone decay.

We initially used grid-point prompts to automate the segmentation process by pre-defining points across a grid. However, this approach may misidentify small decay areas or building structures as decay regions, leading to inefficiencies and inaccuracies. To address the problem, we propose a two-step approach that integrates the object detection method. First, we train a YOLOv8 model to automatically detect stone decay areas. The bounding boxes generated by YOLOv8 are utilized as specific prompts for the fine-tuned SAM model using box prompts, enhancing its focus and precision.

This combined approach improves automation and accuracy over manual prompting. Figure 2 illustrates the workflow of this method: the YOLO model detects potential decay regions in an image and outlines them with multiple bounding boxes, as highlighted in red color. Subsequently, the fine-tuned SAM model uses these bounding boxes as prompts to generate more precise segmentation masks. Hence, this strategy not only automates the segmentation process but also improves accuracy by concentrating SAM's focus on YOLO-identified areas of interest.

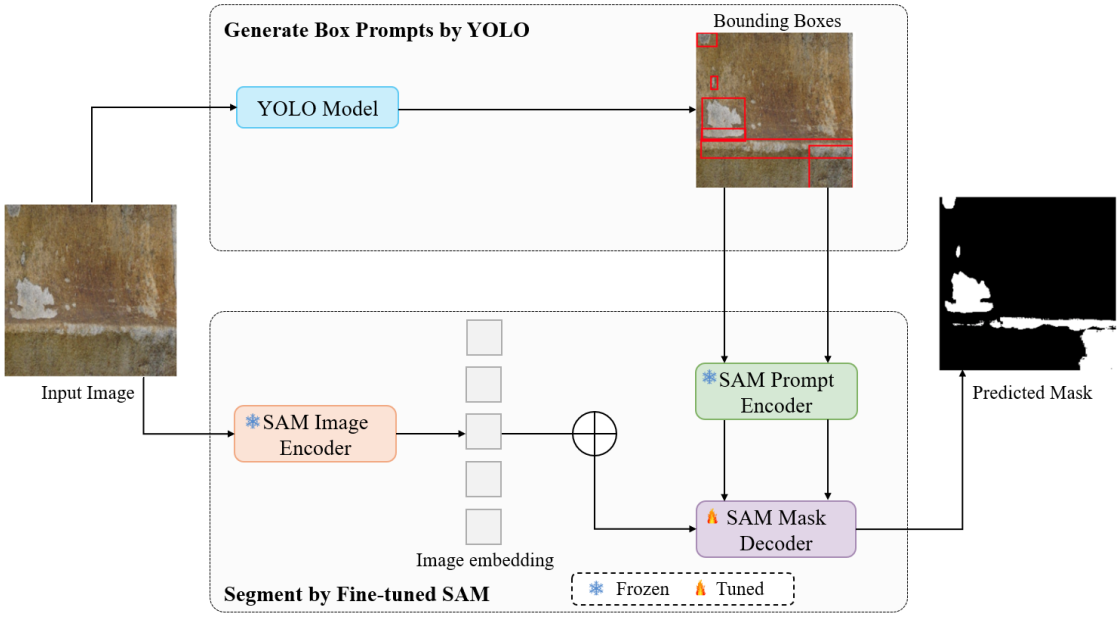


Figure 2. Overview of the proposed segmentation model integrated with YOLO.

3.2 3D Reconstruction Using MVS

For 3D model construction, we employ traditional MVS techniques, which infer depth and surface details from disparities across multiple viewpoints, making them effective for detailed reconstructions from images. Furthermore, they are more efficient in terms of computational speed and simplicity compared with deep learning approaches such as NeRF and 3D GS since they do not require a training process. Specifically, we use COLMAP (Schonberger and Frahm, 2016), a state-of-the-art photogrammetry software that automates the essential processes including camera calibration, image alignment, and dense point cloud generation.

Our pipeline for decay visualization, as shown in Figure 3, integrates the outputs of our segmentation model with COLMAP's 3D reconstruction steps. The process begins with applying the segmentation model to identify and segment decay regions across the visualization dataset. These segmented images are then processed in SfM, starting with feature extraction in the decay-marked regions. SfM subsequently performs feature matching across the dataset, establishing correspondences between images, which are crucial for constructing the 3D structure through triangulation. While an ideal approach would involve reconstructing the site using unmasked images and then projecting the segmented decay regions onto the 3D model based on camera parameters, we simplify this by transferring the segmentation masks directly to the 3D model. We replace the RGB values of the relevant pixels in the original image with the mask colors streamlining the process but potentially reducing the precision, which could be addressed in future work by leveraging camera parameters.

After feature matching, SfM generates a sparse reconstruction to establish camera positions and an initial scene geometry. This sparse model captures the basic structure and serves as the foundation for detailed reconstruction. Subsequently, MVS performs an image undistortion step to ensure accurate geometry. It then computes depth and normal maps for each image based on the sparse model. Depth maps represent the distance of each pixel from the camera, while normal maps describe the orientation of surfaces within the image. Both maps are essential for crafting a more detailed geometric portrayal of each scene. Finally, using the undistorted images, along with the depth and normal maps, MVS synthesizes a dense 3D point cloud with high accuracy and detailed surface textures.

Therefore, this integration not only bridges 2D image analysis with 3D spatial visualization but also provides a robust framework for monitoring and preserving rock and stone-based cultural heritage sites. By merging the precision of segmentation techniques with the depth and detail offered by 3D reconstruction, this framework allows for a more comprehensive understanding of decay patterns and their spatial relationships in historical structures.

3.3 Evaluation Method

To assess the performance of our segmentation approach, we compare the predicted masks against the ground-truth masks using the following three metrics:

Intersection over Union (IoU) quantifies the overlap between the predicted masks and the ground-truth masks.

$$IoU = \frac{Intersect\ \ Area}{Union\ \ Area} \tag{1}$$

Dice Similarity Coefficient (DSC) complements the IoU by also measuring the similarity between the predicted and actual masks but more sensitive to the size of the objects being segmented.

$$DSC = \frac{2 \times Overlap\ \ Area}{Total\ \ Elements} = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{2}$$

where X = the predicted set of pixels
Y = the ground truth set of pixels

The average inference time per image provides a measure of the model's efficiency, thus reflecting its suitability for real-time applications.

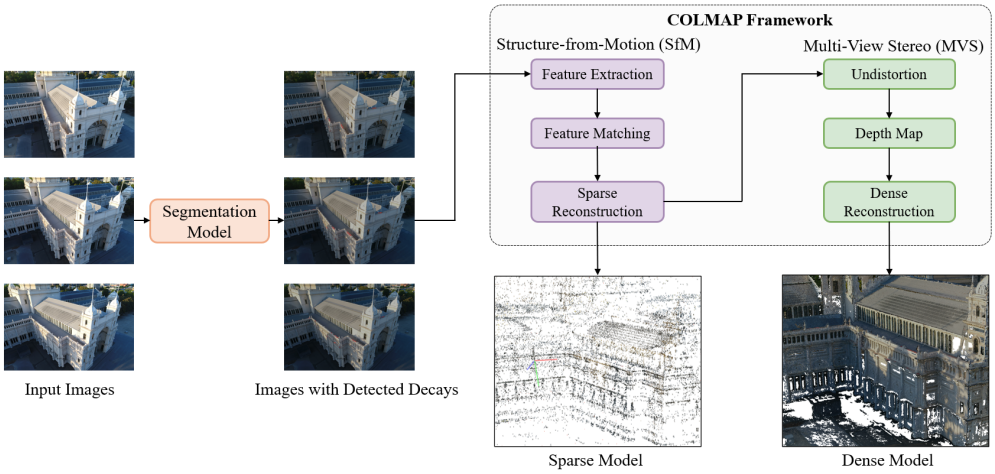


Figure 3. Overview of the proposed decay visualization pipeline.

For 3D model assessment, since no ground-truth is available, we use alternative metrics from COLMAP:

$$Mean\ Error_{\text{reprojection}} = \frac{1}{N}\sum_{i=1}^{N} ||x_i - \hat{x}_i|| \quad (3)$$

where N = total number of 2D points
x_i = the observed 2D point in the original image
$\hat{x}_i$ = corresponding 2D point from the 3D model
$||x_i - \hat{x}_i||$ = the Euclidean distance between points

Additionally, we utilize qualitative method such as visual inspection to further evaluate the result for both segmented masks and the reconstructed model.

4. Experiments and Results

To evaluate the efficiency of our proposed framework, we first fine-tuned the SAM model and trained the YOLOv8 model on our captured dataset. We then applied our segmentation method to a heritage site dataset to evaluate the 3D visualization of decay. The training for SAM and YOLO models was conducted on Google Colab, with NVIDIA L4 GPU. On the other hand, the MVS reconstruction was executed on a personal computer equipped with an NVIDIA RTX 3060Ti GPU, an AMD Ryzen 5 5600X CPU (4.60 GHz), and 64 GB of RAM.

4.1 Datasets

4.1.1 Stone Decay Dataset The Stone Decay Dataset is a 2D image dataset containing high-resolution images of stone from various historical buildings in Melbourne. It includes 348 raw images, representing four common types of stone decay: flaking, black crust, cracking, and contour scaling. Flaking refers to the peeling of the stone's surface scaling in a thin flat; the black crust is frequently dark and composed mainly of particles from the atmosphere due to environmental pollution; cracking represents visible fissures or splits of one part from another in the material; and contour scaling is the separation and spalling of the stone surface along its contours, which often leads to the loss of surface material (Cartwright et al., 2008).

Each image is annotated for different purposes. As demonstrated in Figure 4, for segmentation, ground-truth masks are provided to ensure accurate model training. For object detection, bounding boxes label each decay type distinctly.

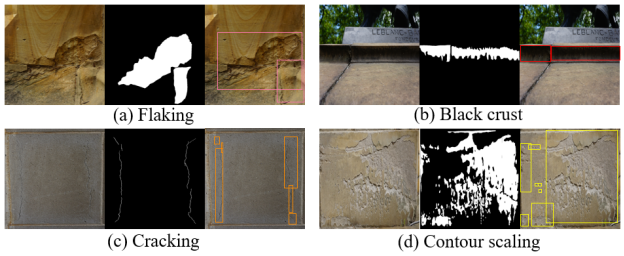


Figure 4. Example images with ground truth annotations. Each set includes the original image, segmentation annotations, and object detection annotations.

4.1.2 Rock Decay Dataset The Stone Decay Dataset is limited for 3D reconstruction tasks due to insufficient overlap between images, which would affect the accuracy of depth estimation when constructing the model. To overcome this limitation and facilitate the visualization of decay types segmented by our fine-tuned SAM model, we created a new dataset, the Rock Decay Dataset, specifically designed for 3D reconstruction.

This new dataset consists of 216 high-resolution images which are manually captured at a rock heritage site with various types of rock decay on the surface in Western Victoria. Unlike the previous dataset, which primarily focused on decay regions, this dataset ensures significant overlap between consecutive images, supporting accurate depth and spatial estimation essential for 3D reconstruction (see Figure 5 (a)). It allows the reconstruction algorithm to better estimate the depth and spatial relationships between different regions of the rock, resulting in a more precise 3D model of the heritage site.

4.1.3 REB-3D Dataset The REB-3D Dataset (Khoshelham, 2018) includes 220 overlapping images of the Royal Exhibition Building in Melbourne, captured by a UAV, as shown in Figure 5 (b). Although the building is primarily made of brick and concrete, it contains several areas of cracking, making it suitable for evaluating the reconstruction effectiveness of our framework. It complements the previously Rock Decay Dataset by offering another scenario, which is a larger site. Hence, this allows for a broader evaluation of our decay visualization framework in diverse heritage conservation contexts.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-1019-2025 | © Author(s) 2025. CC BY 4.0 License.

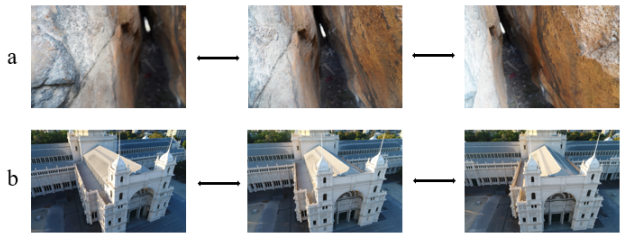


Figure 5. Overlapped images from (a) Rock Decay and (b) REB.

4.2 Segmentation Module

We adopted the SAM architecture proposed by (Kirillov et al., 2023) for decay segmentation and fine-tuned it on our Stone Decay Dataset. The training process employed the ADAM optimizer with a learning rate of 1×10^{-4}, running for 150 epochs. To enhance generalization and prevent overfitting, we applied an early stopping mechanism that breaks the training if no performance improvement was observed over 10 consecutive epochs. Furthermore, we monitored model performance using two metrics: loss and IoU, with Dice-Focal loss selected as the loss function to combine the benefits of dice loss and focal loss.

The overall learning curves for training and validation datasets are illustrated in Figure 6. The model exhibited rapid learning in the early epochs, indicated by a dramatic decrease in both training and validation loss and a corresponding rise in IoU. As training continued, the metrics showed incremental improvement, with fluctuations between epochs 20 and 100, and convergence around the 120th epoch. At this stage, the model achieved near-optimal performance, with an IoU near 0.75, demonstrating effective learning.

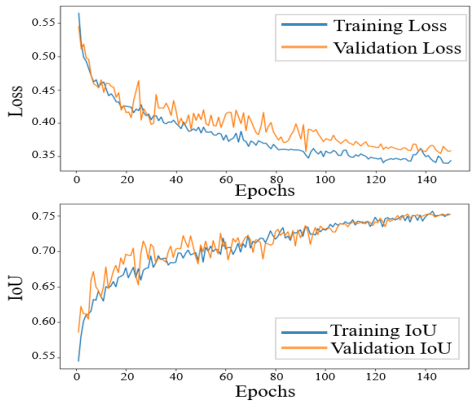


Figure 6. Learning curves of the fine-tuned SAM model.

Moreover, a comparative analysis evaluates the segmentation effectiveness of our proposed fine-tuned SAM model against several SOTA models, including U-Net, U-Net++, YOLOv8-seg, and SAM variants (FastSAM, MobileSAM, SAM2), as shown in Table 1. The models are categorized into interactive methods requiring manual prompts and fully automated approaches. Based on the results, it can be seen that while FastSAM offers the highest speed, its precision is significantly lower. In contrast, our two-step method, though slower than FastSAM, outperforms all other models in both IoU and DSC metrics, achieving the highest accuracy. Notably, the integration in our method allows for a fully automated process, demonstrating its effectiveness and practicality for segmenting decay on heritage buildings and rock heritage sites.

In addition, we assesses the generalization ability of our fine-tuned SAM model on unseen decay types by comparing it against traditional models like U-Net and U-Net++. We collected images with either entirely new decay types or variations in texture, sourced from prior studies (Hou et al., 2024, Bruno et al., 2023, Liu et al., 2020, Cartwright et al., 2008). Figure 7 shows how each model performed on these images. Traditional CNN-based models like U-Net struggled significantly, capturing only small portions of decay regions or failing to segment them altogether in some cases. In contrast, our fine-tuned SAM model, once given the appropriate prompts, successfully segmented these new decay regions, demonstrating robust generalization.

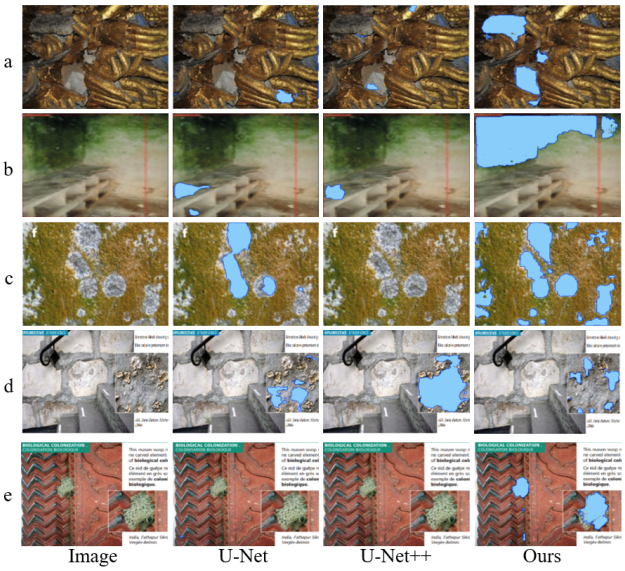


Figure 7. Performance of methods on unseen decay types. (a) golden foil shedding, (b) biological colonization, (c) epilithic biofilms of algae, (d) salt efflorescence, (e) biological colonization. Blue color represents segmentation masks.

4.3 3D Reconstruction Module

In this section, we evaluate the effectiveness of our visualization pipeline on the Rock Decay and REB datasets. Our framework first applies the fine-tuned SAM with YOLOv8 to precisely identify the decay regions and generate corresponding segmentation masks using distinct colors for each decay type. These segmented masks are then overlaid onto the original images, with a 40% transparency, which highlights the decay regions while preserving key image details necessary for the 3D reconstruction. Finally, the annotated images are processed in MVS to generate a detailed 3D model.

Figure 8 illustrates the result of this process on the rock site. The image on the left side represents the original 3D model of the site, while the image on the right demonstrates the segmented 3D model generated by our framework. In the segmented model, various decay types are highlighted with different colors, for example, red for the regions exhibiting black crusts (carbonization/sulfation) and pink for the regions exhibiting contour scaling. Visual inspection of the model reveals that (1) the segmented 3D model maintains a similar visual quality to the original, demonstrating that the overlay does not affect reconstruction fidelity, and (2) the decay regions are effectively highlighted, allowing for comprehensive spatial analysis of the

	Method	Prompt	IoU(%)↑	DSC(%)↑	Time(ms)↓
Interactive	SAM(baseline)	box	45.49	54.47	388.5
	FastSAM	box	35.30	43.30	**13.0**
	MobileSAM	box	44.92	53.95	109.5
	SAM2	box	46.72	55.61	368.3
Automatic	U-Net	/	33.73	50.45	126.6
	U-Net++	/	37.06	54.08	255.6
	YOLOv8-seg	/	31.23	40.75	56.2
	Fine-tuned SAM(ours)	YOLOv8	**53.44**	**63.37**	383.4

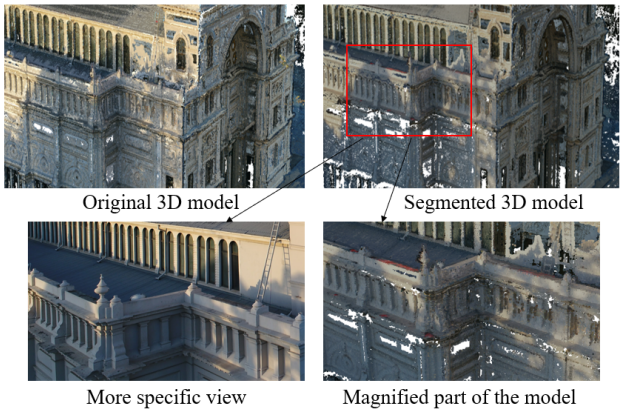
Table 1. Comparative performance of different segmentation models.

extent and distribution of decay. This visualization provides valuable information for conservation planning, enabling targeted preservation efforts based on decay distribution.

Original 3D model Segmented 3D model

Figure 8. Result of decay visualization (red and pink regions) for Rock Decay Dataset.

Figure 9 demonstrates the resulting 3D models of the REB building, with the original model on the left and the segmented model on the right. In the segmented model, the red areas highlight regions affected by what appear to be black crusts. Visual inspection reveals both strengths and limitations in the reconstruction. A key advantage is that the visual quality of the model remains consistent with the original, with decay regions highlighted without compromising structural detail. However, the decay regions are less apparent than in the rock site case and less certain as to their origin as they could be caused by biological matter. This difference is likely due to the UAV capturing images from a greater distance on a larger structure, resulting in decay areas occupying a smaller portion of each image. For example, decay may cover nearly 50% of an image in the Rock Decay dataset, while in this case, it might only represent about 5%, which impacts the segmentation model's effectiveness in identifying small decay areas and confirming their origin. Additionally, the REB building exhibits fewer decayed areas overall, so fewer decay regions appear in the final visualization.



Original 3D model Segmented 3D model

More specific view Magnified part of the model

Figure 9. Result of decay visualization for REB-3D Dataset.

Furthermore, the mean re-projection errors for the original and decay-annotated models in the Rock Decay and REB cases are 1.3512 and 1.3076, and 0.7869 and 0.7358, respectively. These consistently low values suggest accurate alignment of 3D points with the original images. The slightly lower error in the decay-annotated model also indicates that incorporating decay segmentation and overlay does not negatively impact the overall quality of the 3D model.

Moreover, to enhance the clarity of decay regions, especially when they resemble the original stone surface, we utilize a heat map-style overlay for more intuitive visualization.

In Figure 10, the same section of the rock site is presented with a new heat map-style overlay covering the entire site. The new colors now indicate decay probability (as shown by the color bar), ranging from red for high probability of decay (close to 1) to dark blue for non-decayed or background areas (probability = 0). This approach enhances visual visual contrast between decayed and non-decayed areas by highlighting decay with vibrant colors and leaving the background shaded in dark blue. In addition, the heat map provides a probabilistic estimate of decay distribution, which allows for a more specific understanding and could be refined in future work to incorporate decay severity that offers a more detailed and actionable assessment for conservation planning.

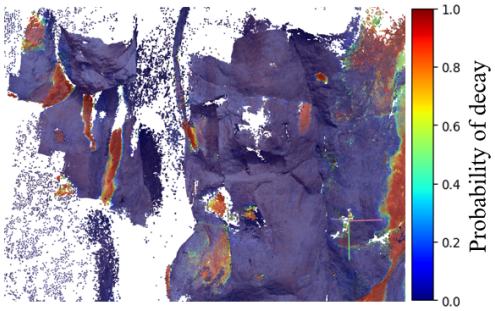


Figure 10. Heat map visualization for rock site.

5. Discussion

The experimental results demonstrate that our proposed framework effectively segments and visualizes decay on stone and rock heritage sites, marking a notable advancement over conventional deep learning methods in this field. Unlike previous segmentation models, which often require extensive re-training or manual input, our framework automatically identifies decay areas with capability for incorporating new decay types. Also, the generated 3D model provides an efficient solution for monitoring the condition of heritage buildings and rock heritage sites compared to traditional in-person inspections.

Our results show that the combination of YOLOv8 for prompt generation with fine-tuned SAM boosts segmentation accuracy and enhances automation, demonstrating superior performance over other SAM variants and providing robust adaptability for identifying unseen decay types. Thus, our technique fills the critical gap in cultural heritage applications, establishing SAM as an effective and practical method.

In addition, the visualization pipeline that incorporates segmented decay regions into a 3D reconstruction model, provides a scalable and efficient solution for assessing the condition of heritage sites. Our framework also simplifies the 3D visualization process by using 2D datasets for segmentation, which are then integrated into a 3D reconstruction. This approach eliminates the need for constructing complex 3D datasets when directly segmenting decay within 3D models, enabling broader implementation across various heritage sites without the requirement for specialized equipment.

Future work could focus on improving segmentation accuracy. For instance, the current fine-tuned model is based on the original SAM structure, but its performance can be enhanced by adopting modifications used in SAM variants such as SAM2, which has demonstrated improvements in segmentation accuracy and efficiency. Moreover, currently, a precise ground-truth based 3D model of a selected heritage site needs to be obtained through close calibration and corroboration by rock and stone conservation experts. Without this process (which needs periodic checking) it is challenging to directly compare the accuracy of our reconstructed 3D model against the real heritage site. Hence, obtaining a ground-truth model in the future would not only allow a more detailed evaluation of our method that utilities SfM and MVS, but also facilitate a comparison of various advanced 3D reconstruction techniques, such as MVSNet, NeRF, and 3D GS.

In addition, our current 3D models indicate decay presence but lack other information such as decay severity, which suggests another direction on extending the model to assess the severity of decay by categorizing decay into levels. For example, mild decay could be represented by values below 0.3, moderate by 0.3-0.6, and severe by values above 0.6. A heat map-style visualization based on these severity degrees would help distinguish more urgent decay areas, with brighter colors indicating more critical decay. This approach is not only useful to quantify the decay and evaluate the risk posed to the heritage site being surveyed but also supports long-term monitoring and assessment.

6. Conclusion

In this paper, we have presented a novel approach for real-time segmentation of decay regions on stone and rock heritage sites by fine-tuning the SAM model and integrating it with YOLOv8. We evaluated our model against several SOTA methods, and the results indicate that it significantly outperforms these approaches in precision while maintaining acceptable processing speeds for real-time applications for cultural heritage. Our model also demonstrates superior generalization ability in identifying various unseen decay types after receiving the prompts compared with traditional data-driven models, such as U-Net and U-Net++. Additionally, leveraging the strengths of our segmentation model, we proposed a pipeline to transform a series of images into a 3D model with visually highlighted decay regions. We applied this pipeline to a case study, and the results demonstrate its effectiveness in successfully detecting and visualizing decay regions, while also showing potential for assessing decay severity and identifying decay in other substrates such as brick and concrete.

7. Acknowledgements

With thanks to staff at the Barengi Gadjin Land Council: Farren Branson, Chrystle Carr, Michael Douglas, Jarra Secombe, and Jake Goodes, Rock Art Officer at Parks Victoria, for data collection and logistical support.

References

Agdas, D., Rice, J. A., Martinez, J. R., Lasa, I. R., 2016. Comparison of visual inspection and structural-health monitoring as bridge condition assessment methods. *Journal of Performance of Constructed Facilities*, 30(3), 04015049.

Basso, A., Condorelli, F., Giordano, A., Morena, S., Perticarini, M., 2024. Evolution of Rendering Based on Radiance Fields. The Palermo Case Study for a Comparison Between Nerf and Gaussian Splatting. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 57–64.

Bruno, S., Galantucci, R. A., Musicco, A., 2023. Decay detection in historic buildings through image-based deep learning. *VITRUVIO-International Journal of Architectural Technology and Sustainability*, 8, 6–17.

Cartwright, T. A., Bourguignon, E., Bromblet, P., Cassar, J., Charola, A. E., De Witte, E., Delgado-Rodrigues, J., Fassina, V., Fitzner, B., Fortier, L. et al., 2008. *ICOMOS-ISCS: illustrated glossary on stone deterioration patterns*. International Council of Monuments and Sites.

Chen, K., Reichard, G., Xu, X., Akanmu, A., 2021. Automated crack segmentation in close-range building façade inspection images using deep learning techniques. *Journal of Building Engineering*, 43, 102913.

De la Fuente, D., Vega, J. M., Viejo, F., Díaz, I., Morcillo, M., 2013. Mapping air pollution effects on atmospheric degradation of cultural heritage. *Journal of cultural heritage*, 14(2), 138–145.

Furukawa, Y., Hernández, C. et al., 2015. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2), 1–148.

Fusco Girard, L., Vecco, M., 2021. The “intrinsic value” of cultural heritage as driver for circular human-centered adaptive reuse. *Sustainability*, 13(6), 3231.

Gopinath, V. K., Ramadoss, R., 2021. Review on structural health monitoring for restoration of heritage buildings. *Materials Today: Proceedings*, 43, 1534–1538.

Grosso, R., Mecca, U., Moglia, G., Prizzon, F., Rebaudengo, M., 2020. Collecting built environment information using UAVs: Time and applicability in building inspection activities. *Sustainability*, 12(11), 4731.

Hatır, M. E., İnce, İ., Korkanç, M., 2021. Intelligent detection of deterioration in cultural stone heritage. *Journal of Building Engineering*, 44, 102690.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hou, M., Huo, D., Yang, Y., Yang, S., Chen, H., 2024. Using mask R-CNN to rapidly detect the gold foil shedding of stone cultural heritage in images. *Heritage Science*, 12(1), 46.

Jocher, G., Chaurasia, A., Qiu, J., 2023. Ultralytics YOLO.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 1–14.

Khoshelham, K., 2018. Smart heritage: challenges in digitisation and spatial information modelling of historical buildings. *2nd Workshop on Computing Techniques for Spatio-Temporal Data in Archaeology and Cultural Heritage*, University of Melbourne Melbourne, Australia, 7–12.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. et al., 2023. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Larochelle, H., Erhan, D., Bengio, Y., 2008. Zero-data learning of new tasks. *AAAI*, 1number 2, 3.

Liu, X., Koestler, R. J., Warscheid, T., Katayama, Y., Gu, J.-D., 2020. Microbial deterioration and sustainable conservation of stone monuments and buildings. *Nature Sustainability*, 3(12), 991–1004.

Liu, Y., Lin, Y., Yeoh, J. K., Chua, D. K., Wong, L. W., Ang, M. H., Lee, W., Chew, M. Y., 2021. Framework for automated UAV-based inspection of external building façades. *Automating cities: Design, construction, operation and future impact*, 173–194.

Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B., 2024. Segment anything in medical images. *Nature Communications*, 15(1), 654.

Ma, Z., Liu, S., 2018. A review of 3D reconstruction techniques in civil engineering and their applications. *Advanced Engineering Informatics*, 37, 163–174.

Mazzacca, G., Karami, A., Rigon, S., Farella, E., Trybala, P., Remondino, F., 2023. NERF FOR HERITAGE 3D RECONSTRUCTION. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 1051–1058.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2021. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7), 3523–3542.

Murtiyoso, A., Karwel, A., Grussenmeyer, P., 2024. Comparison of state-of-the-art Multi-view stereo solutions for close range heritage documentation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 317–323.

Nooralishahi, P., Ramos, G., Pozzer, S., Ibarra-Castanedo, C., Lopez, F., Maldague, X. P., 2022. Texture analysis to enhance drone-based multi-modal inspection of structures. *Drones*, 6(12), 407.

Osco, L. P., Wu, Q., de Lemos, E. L., Gonçalves, W. N., Ramos, A. P. M., Li, J., Junior, J. M., 2023. The segment anything model (sam) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124, 103540.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., Wang, X.-Z., Wu, Q. J., 2022. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4), 4051–4070.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L. et al., 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 234–241.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Spennemann, D. H., Graham, K., 2007. The importance of heritage preservation in natural disaster situations. *International Journal of Risk Assessment and Management*, 7(6-7), 993–1001.

Tan, Y., Li, S., Liu, H., Chen, P., Zhou, Z., 2021. Automatic inspection data collection of building surface based on BIM and UAV. *Automation in Construction*, 131, 103881.

Timothy, D. J., 2014. Contemporary cultural heritage and tourism: Development issues and emerging trends. *Public Archaeology*, 13(1-3), 30–47.

Verykokou, S., Ioannidis, C., 2023. An overview on image-based and scanner-based 3D modeling technologies. *Sensors*, 23(2), 596.

Wang, B., Aboah, A., Zhang, Z., Pan, H., Bagci, U., 2023. Gazesam: Interactive image segmentation with eye gaze and segment anything model. *NeuRIPS 2023 Workshop on Gaze Meets ML*.

Welponer, M., 2019. Open-source image-based 3D reconstruction pipelines: Review, comparison and evaluation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-2/W17*, 331–338.

Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S., Hong, C. S., 2023. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*.

Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J., 2023. Fast segment anything. *arXiv preprint arXiv:2306.12156*.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, Cham, 3–11.