The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-M-9-2025
30th CIPA Symposium “Heritage Conservation from Bits:
From Digital Documentation to Data-driven Heritage Conservation”, 25–29 August 2025, Seoul, Republic of Korea

Coupling V-SLAM and Semantic Segmentation for Cultural Heritage Documentation

Ahmad El-Alailyi [1,2], Kai Zhang [1], Chiara Mea [1], Luca Perfetti [3], Fabio Remondino [2], Francesco Fassi [1]

[1] 3D Survey Group, ABC Department, Politecnico di Milano, Milano, Italy – (ahmad.elalailyi, kai.zhang, chiara.mea, francesco.fassi)@polimi.it
[2] 3D Optical Metrology (3DOM) Unit, Bruno Kessler Foundation (FBK), Trento, Italy – (aelalailyi, remondino)@fbk.eu
[3] Department of Civil, Architectural, Environmental Engineering and Mathematics (DICATAM), Università degli Studi di Brescia, Brescia, Italy - luca.perfetti@unibs.it

Keywords: Multi-camera Mobile Mapping System, V-SLAM and 3D Reconstruction, Deep Learning, Heritage Documentation, Architectural and Pathologies Semantics, 2D-to-3D Enrichment.

Abstract

3D digitization has become an essential tool in cultural heritage documentation, offering unprecedented opportunities for preservation, analysis, and dissemination. Beyond only capturing 3D spatial geometry, the semantic enrichment of 3D models is rapidly evolving offering a more efficient interpretation and usage of 3D data. Traditionally, 3D semantic enrichment has relied on point cloud-based segmentation. However, 3D point cloud-based segmentation approaches can struggle with the efficient identification of small-scale, geometric elements, or visually ambiguous classes, limiting their applicability in such contexts. This study leverages the rich contextual and textural information of 2D imagery to detect challenging semantic categories, such as fine architectural elements (e.g., individual stone blocks) and material decay (e.g., material detachment and material loss), using deep learning-based 2D semantic segmentation techniques. These detections are then projected into 3D space through a 2D-to-3D semantic segmentation framework that couples V-SLAM and 3D results with the 2D predictions. The framework is evaluated on data acquired using the fish-eye multi-camera mobile mapping system ATOM-ANT3D in two challenging case study environments. Achieved results demonstrate a reliable level of accuracy given the inherent complexity of targeted classes, enhancing the interpretability of 3D models by providing meaningful and metrically interpreted objects classifications in 3D models. (Demonstration video)

1. Introduction

The demand for accurate and intuitively interpretable 3D spatial data models has significantly increased in the Cultural Heritage (CH) field for documentation and monitoring purposes. Although Terrestrial Laser Scanning (TLS) and conventional photogrammetry (Luhmann et al., 2007; Remondino, 2011; Fassi et al., 2011; Pritchard et al., 2017) can achieve accurate 3D reconstructions, their limited mobility and workflows reduce their effectiveness in complex environments (Berra and Peppa, 2020; Muralikrishnan, 2021). In particular, narrow spaces like staircases and corridors restrict instrument setup and limit unobstructed observation points, while irregular structure geometry often causes occlusions and incomplete data capture. Mobile Mapping Systems (MMS), leveraging Simultaneous Localization and Mapping (SLAM) processing methods, have gained traction due to their flexibility, speed, and real-time mapping capabilities (Elhashash et al., 2022; Elalailyi et al., 2024b). Specifically, Visual SLAM (V-SLAM) uses sequences of images to estimate the position of a sensor within an environment (Davison et al., 2007). As a result, visual-based MMSs have recently become popular for rapid 3D data acquisition and efficient 3D reconstruction (Ortiz-Coder and Sánchez-Ríos, 2019; Kuo et al., 2020; Torresani et al., 2021), and in particular, the use of systems utilizing multi-camera fish-eye setups for narrow and complex surveying applications (Elalailyi et al., 2024a; Perfetti et al., 2024a).

At the same time, Artificial Intelligence (AI) is playing a crucial role in enhancing 3D spatial data understanding. Previous applications of AI in CH have mainly focused on 3D point cloud-based classification (Grilli and Remondino, 2019; Teruggi et al., 2020). Simultaneously, AI-driven 2D image data processing has played an important role in improving visual-based 3D spatial data applications. Convolutional Neural Networks (CNNs) have improved feature detection and matching, with notable methods including SuperPoint (DeTone et al., 2017) and D2-Net (Dusmanu et al., 2019). Conventional frameworks such as YOLO (You Only Look Once) (Redmon et al., 2015) have enabled fast 2D image object detection. AI-driven approaches have also been integrated for 2D semantic segmentation (Chen et al., 2016) or scene understanding (Murez et al., 2020), enabling a more automated and intelligent interpretation of visual data. Recent developments in multi-modal AI have introduced advanced object detection, semantic segmentation and visual grounding techniques (i.e., models that link natural language phrases to specific objects in an image). These models are capable of performing prompt-driven segmentation, and zero-shot detection that enable automated segmentation and identification without the need for training or manual labelling. Segment Anything Model (SAM) (Kirillov et al., 2023), have boosted the analysis and understanding of images utilizing vision transformers and prompt encoders to predict object masks. Grounding DINO (Liu et al., 2023b) is an open-set object detector that integrates vision and language through a transformer-based architecture. It extends the DINO detection framework, a DETR-style transformer model that uses object queries, by incorporating text prompts from a language encoder, such as CLIP's text model. In this context, CLIP (Contrastive Language–Image Pretraining) (Radford et al., 2021) plays a foundational role by aligning visual and textual representations, enabling models like Grounding DINO to interpret visual scenes through natural language. Building on these capabilities, Galanakis et al., (2024) investigated SAM's potential for stone-level structural analysis and segmentation. SAM, in combination with detection models such as Grounding DINO, has enabled effective object detection and semantic labelling (Réby et al., 2023). El-Alailyi et al., (2025a) presented an approach for 2D-to-3D semantic segmentation utilizing Sa2VA (Yuan et al., 2025), a unified architecture for dense grounded knowledge of images combining SAM2 and LLaVa (Large Language and Vision Assistant) (Liu et al., 2023a), enabling architectural features detection, and conventional object detection YOLOv8 (Jocher et al., 2023) to identify cracks.

Despite the growing interest in AI-driven CH 3D analyses, significant practical challenges remain. Complex CH environments, characterized by intricate spatial configurations (e.g., tight corridors and irregular geometries), variable illumination conditions, and challenging surface texture, can

hinder the efficiency of 2D semantic segmentation and object detection, as well as compromise the completeness of 3D reconstructions. These limitations are especially critical when dealing with subtle or non-standard element detections, which are rarely addressed in conventional detection and 3D data enrichment frameworks but are essential for CH documentation.

1.1 Paper’s Aim

To address the need for semantically enriched 3D models that include challenging and underrepresented object detection scenarios, this study proposes a framework that couples V-SLAM and 3D results with 2D image-based semantic segmentation and object detection using zero-shot detectors, foundational models and conventional supervised detectors. The method specifically targets stone blocks as key architectural elements, along with material degradation features such as material detachment and material loss. The pipeline is tested and validated using on-site acquired data with the fish-eye multi-camera system ATOM-ANT3D (Figure 1) (Elalailyi et al., 2024a; Perfetti et al., 2024b) in two challenging and historical CH sites.
This work intends to shed light on semantically enriched 3D representations of surveyed heritage scenes to support enhanced 3D architectural documentation, feature recognition, and improved conservation planning with metric information.

Figure 1. The ATOM-ANT3D fish-eye multi-camera portable mobile mapping system.

2. Methodology

2.1 V-SLAM and 3D Reconstruction

The acquired multi-camera image datasets are processed using a V-SLAM and 3D reconstruction pipeline (El-Alailyi et al., 2025b). Depending on the mapped environment settings, the synchronized multi-camera configuration can generate up-to four independent V-SLAM trajectories, each corresponding to a distinct stereo camera pair. A post-processing multi-camera pose graph optimization fuses available trajectory estimates into a single, consistent solution leveraging redundancy in camera observations and pose estimations. First, it initializes a nonlinear factor graph, with nodes representing full keyframe poses and edges encoding relative pose constraints between consecutive keyframes. To account for pose uncertainty, a noise model coupled with a loss function is adopted to increase robustness against outliers. Then the optimization is performed via a Gauss-Newton optimizer, minimizing nonlinear residuals to achieve a globally consistent trajectory. Subsequently, a multi-view feature-based optimization performed in Metashape (Agisoft Metashape, 2024) refines single stereo V-SLAM estimates or the unified trajectory obtained by the previous optimization, which serves as the initial pose estimate for the five-camera rig. The initial trajectory helps identify spatially proximate cameras for matching and detects possible candidate loop closures. Multi-view triangulation is later performed to generate a sparse 3D tie points cloud, which is further refined through a constrained bundle adjustment (BA) using the pre-calibrated rigid relative orientation of the multi-camera system. Points with high reprojection errors are iteratively filtered, and optimization proceeds until the RMSE converges below a one-pixel threshold, ensuring geometric consistency and accurate camera poses estimation across the datasets (Figure 2).

2.2 AI-based 2D Segmentation

While 3D reconstruction provides a spatial and geometric representation, it lacks semantic information that can be important for architectural documentation, object understanding, and CH conservation efforts. The use of AI-driven tools can help to speed up the interpretation process, efficiently analysing the acquired datasets without requiring additional instruments or long manual operations (Zhang et al., 2024). Many architectural surfaces feature repeated patterns and materials, making them ideal for “zero-shot” AI models (e.g., SAM2). In this study, we test three methods for object detection: (i) SAM2 to generate zero-shot masks, (ii) Grounding DINO, through text-based prompts, for automated detection of architectural elements, and (iii) the Yolo-v5 as a close-set detection model. While zero-shot foundational models offer significant flexibility and automation, they may fall short when dealing with highly specific or less-represented features, such as challenging architectural components and material decay. In such cases, conventional supervised models like YOLO remain necessary, as they allow for targeted detection of well-defined object classes when trained with a carefully curated, annotated dataset.

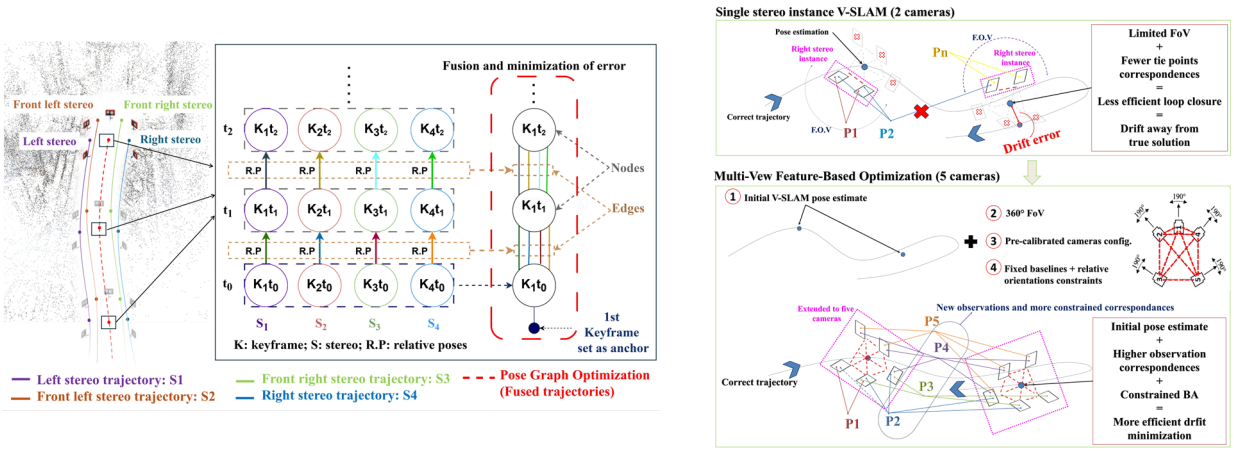

Figure 2. The multi-camera poses graph optimization (left) and the multi-view feature-based optimization (right) schemes.

2.3 Coupling of 2D Semantics and 3D Reconstruction

The coupling of 2D semantic information extracted from the fish-eye images with the 3D reconstruction is applied by adapting the method proposed by Alami and Remondino, (2024) and El-Alailyi et al., (2025a). Dense point clouds are initially derived by densifying the 3D reconstruction and subsampled at 1 cm space between adjacent points to reduce the computational overhead of the projection algorithm, and later, after the projection process is finalized, the results are interpolated back onto the full high-resolution dense cloud. First, the subsampled point clouds are converted to a voxel grid at a user-specified resolution, and a ray tracing environment is constructed. Using intrinsic calibration, 2D labels are corrected to remove lens distortion and projected into 3D using the known optimized V-SLAM camera poses (Section 2.1). The rays intersect visible voxels encapsulating the corresponding 3D points, and the semantic labels from the 2D masks are assigned to the corresponding 3D voxels enhanced through local neighbourhood search to increase spatial coverage and label completeness (Figure 3). In the proposed case study 1, the Minguzzi staircase (Section 3.1), SAM2 generated semantic masks of stone blocks detections are randomly labeled, meaning that the same stone may receive different label IDs across images, or conversely, different stones may be assigned the same label ID. This inconsistency, combined with the challenging geometry of the site, characterized by a narrow, spiraling staircase and the use of wide-angle fish-eye cameras, introduces further complexity. The viewpoint of each stone block varies significantly from one image to another, and in many cases, stones are only partially visible in some views due to occlusions or limited field of view, while being fully visible in others. To address these challenges, a joint 2D-3D integration strategy is adopted. First, all mask IDs across the image set are reassigned to unique labels. During the projection stage, 2D masks are sequentially projected onto the 3D point cloud. When a 3D point is first intersected by a projected mask, the semantic label is stored. As the process continues, if the same 3D point is intersected by a different mask (i.e., representing a candidate for the same stone block), a 3D voting mechanism is triggered which compares the total number of 3D points currently labelled by each competing 2D mask and assigns the 3D points to the label with the greater overall spatial coverage in 3D, favouring the label that corresponds to a more complete view of the object in 2D (e.g., fully visible masked stone block). Through this iterative process, overlapping and redundant 2D masks are merged based on geometric evidence in 3D space, resulting in a more coherent semantic segmentation of stone blocks directly on the 3D model. The accuracy of the proposed method is influenced by several factors, including the voxel grid resolution, the quality of the input segmentation masks, and the accuracy of camera calibration. Inaccurate calibration, segmentation errors, or overly coarse voxelization may lead to mislabelling, incomplete coverage, or the introduction of outliers in the final 3D semantic model.

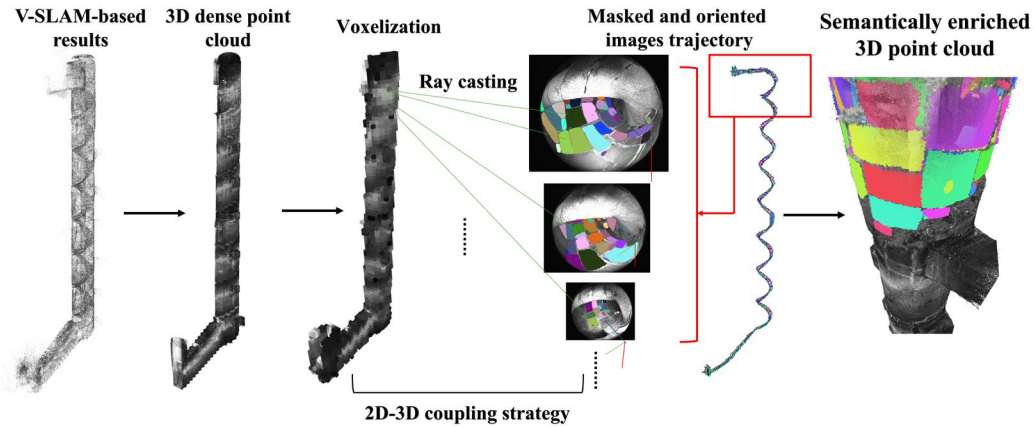

Figure 3. The proposed 2D-to-3D semantic segmentation projection and coupling pipeline of the 2D semantic masks with the 3D reconstruction results (images and point clouds).

3. Case Studies, Data Processing, and Validation

3.1 Case Study One: Minguzzi Spiral Staircase in the Milan Cathedral, Italy

Within the main façade of the Milan Cathedral (Italy), the Minguzzi spiral staircase is located inside the front-right pylon situated at the southwest corner of the cathedral, extending approximately 25 meters in height. The staircase features a central marble column with a diameter of about 40 cm, and a narrow passage that measures, transversally, just 70 cm in width, making movement inside highly constrained (Figure 4). The challenging environmental conditions within the staircase further complicate documentation efforts. The interior is characterized by extremely low luminance, with poor-quality artificial lighting and limited natural illumination from a few exterior window openings. These constraints necessitate specialized 3D mapping approaches to effectively capture and analyse the structure.

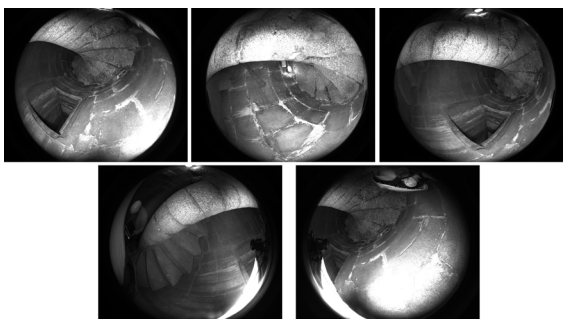

Figure 4. Example fish-eye images from the Minguzzi survey showing stone blocks with atypical geometry and surface conditions due to white material build-up.

Research on Milan Cathedral has been extensive, with several studies focusing on architectural documentation, structural

analysis, and digital surveying (Achille et al., 2020; Perfetti et al., 2024a).

3.2 Case Study Two: The Pozzi Cells Area of the Doge’s Palace in Venice, Italy

The Pozzi cells and passageways (Figure 5) within the Doge’s palace in Venice exhibit centuries-old stone and brick construction, showing signs of deterioration, peeling, and discoloration. These confined spaces, characterized by narrow corridors and small cells, highlight the restrictive nature of the historical detention areas. The limited manoeuvring space and uneven lighting conditions create exposure inconsistencies.

Figure 5. Example images from the Pozzi cells area survey.

3.3 Data Acquisition, Processing and Validation

The Minguzzi dataset utilized in this study is collected in a previous survey (Elalailyi et al., 2024a) through an ascending spiral trajectory field acquisition, which starts from the lower entrance and ends at the upper exit of the staircase. Over a span of approximately 8 minutes, a total of 7,905 images are captured (1,581 images/camera) and processed (El-Alailyi et al., 2025b). In the second case study, the Doge’s Palace acquisition is part of a bigger survey covering three floors of the structure, capturing the architectural details and spatial layout of each level. In this work, only the first floor of the surveyed area is considered, with ca. 25 m long trajectory along the ca. 1.2 m wide passage where 4,700 images (940 images/camera) are captured and processed using the V-SLAM and 3D reconstruction framework (Section 2.1). 3D dense clouds are later generated for both the Minguzzi and Pozzi cells area with ca. 34 and 20 million points, respectively (Figure 6).

The AI-based 2D semantic segmentation for both case studies is detailed as follows. In the Minguzzi case study, and given the large data volume, a fine-tuned SAM2 foundational model is applied to generate zero-shot segmentation masks to extract stone blocks, which constitute the key elements of the staircase, essential for maintenance and documentation. The central camera image dataset of ATOM-ANT3D, sampled every 5 frames per second given the high image overlap, is chosen due to its unobstructed viewing point of the stone blocks. To address the challenge of recognizing stone blocks that are highly uniform in color and separated by only shallow gaps and surface build-up material, it is necessary to adjust the brightness and contrast of the images. These enhancements are needed to maintain overall image readability, so an adaptive approach is required. Instead of applying fixed values, CLAHE (Contrast Limited Adaptive Histogram Equalization) (Zuiderveld, 1994) is used. This technique improves local contrast, helping the segmentation model better identify the edges of individual stones while preserving fine details. To isolate stone blocks from other architectural elements, Grounding DINO was employed to detect and label non-target objects (e.g., central marble column, windows, and stair steps) using text prompts, generating bounding boxes with associated confidence scores. These were then used to guide the SAM2 model in an attempt to produce refined masks, which were subsequently combined with automatically generated masks to subtract the unwanted elements.

In the Pozzi cells area, the used data covers an area including guardian room and prison cells, and a narrow corridor. This area is mostly built with stones and bricks, with some areas covered with mortar. The building material of the prison cells demonstrates different degrees of deterioration. The material surfaces indicate typical material (1) detachment and (2) loss issues, due to human intervention and humidity factors. The detection of the decay phenomenon in cultural heritage remains challenging. In this application, YOLOv5 detection model is utilized for the object detection task using a dataset comprised of 184 images. The model is trained with compressed images (640*640 pixels) reaching convergence at around 40 epochs. The detection is further used to guide the segmentation of SAM2, generating masks of decay by each category. The results of the 3D reconstruction camera's trajectory, dense clouds and the 2D AI-based detection are later coupled using our proposed 2D-to-3D semantic segmentation framework (Section 2.3). To validate the effectiveness of the coupling approach in producing semantically enriched 3D models, both qualitative and quantitative evaluations are conducted on representative sections of each case study. Given the sensitivity and complexity of the targeted classes, particularly small-scale or degraded features prone to visual ambiguity, ground truth is carefully defined by selecting the most visually clear and geometrically distinct 3D instances. This selective strategy ensured a reliable reference for validating the 2D-to-3D projection pipeline, allowing us to confidently assess potential errors introduced by the 2D object detection and semantic segmentation masks and the projection process into the 3D model, minimizing subjectivity. For the Minguzzi staircase, a set of 9 well-preserved stone blocks with clearly visible boundaries are manually traced on the 3D model, excluding those affected by deterioration, occlusion, or surface build-up (Section 4). In the Doge’s Palace dataset, 25 instances of material detachment and 11 instances of material loss are similarly delineated on the dense 3D cloud, selected based on their visual and geometric clarity (Section 4). In both cases, precision, recall, F1-score, and Intersection over Union (IoU) are computed to assess the accuracy and completeness of the 2D-to-3D semantic projection relative to the manually defined ground truth.

4. Results, Evaluations and Discussion

Figure 6 demonstrates the results of the 3D reconstruction trajectories and tie point clouds, and their corresponding 3D dense clouds used as the input 3D models to the 2D-to-3D semantic enrichment. In the context of 2D-to-3D semantic segmentation, the accuracy of object detection and the quality of 2D masks are critical, as the projected semantic information directly influences the reliability and completeness of the 3D labelling and semantic enrichment. The quality of the 2D semantic segmentation is a limiting factor, and the presence of inaccurate 2D masks can lead to mislabeling, ambiguity, or loss of detail in the 3D representation, particularly in complex environments. Given the unique texture and geometric characteristics of the Minguzzi staircase, the use of wide-angle fish-eye imagery, and domain discrepancies between the training data of the model, usually based on commonly found objects, the attempts to automatically distinguish stone blocks using Grounding DINO and the SAM2 generated masks from the

central marble column, windows, and stair steps are ineffective. As a result, a fully automatic classification approach couldn’t reliably differentiate between stone blocks and other elements, consequently requiring manual intervention.

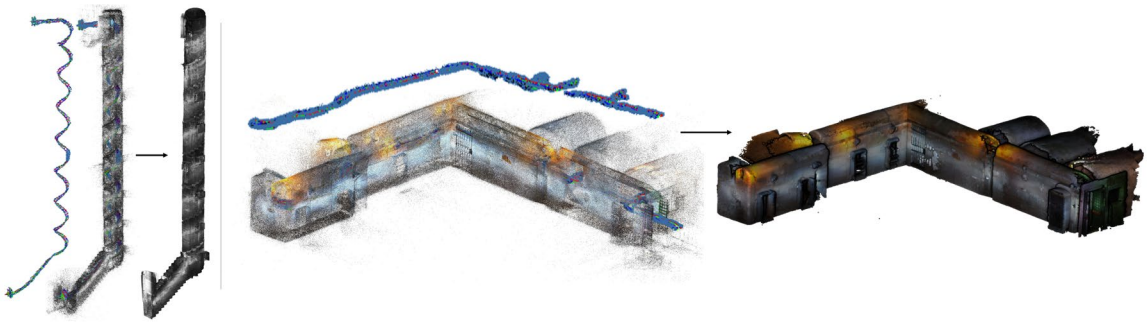

Figure 6. Results (trajectory and 3D tie points) and their corresponding 3D dense point clouds for Minguzzi (left) and Pozzi cells area (right) case studies.

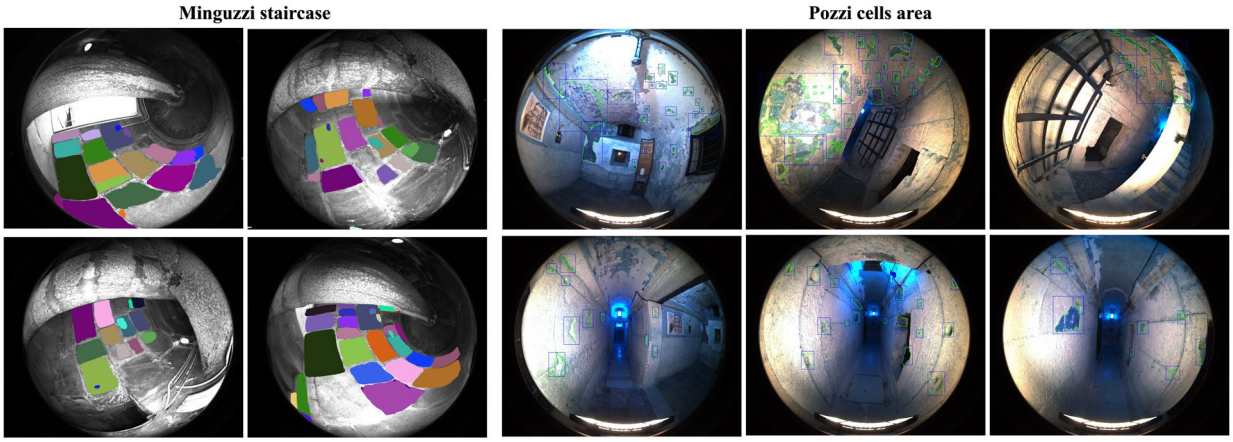

Figure 7. Examples of the detected and segmented classes from the two case studies: stone block segmentation in the Minguzzi staircase (left); material detachment (top) and material loss (bottom) detected in the Pozzi cells area (right).

Figure 7 reports the results from the SAM2 2D stone block segmentation in the Minguzzi case study and the YOLOv5/SAM2 coupling for decay detection at the level of material detachment and material loss for the Pozzi cells area. Figure 8 presents the qualitative visual assessment of selected results from each case study, reporting the overlap (green), false positives (red) and false negatives (blue) of the 3D detection with respect to their corresponding ground truth. The results of the Minguzzi staircase (Figure 8a) demonstrate that the projected masks align well with their respective ground truth locations. However, the segmentation boundaries at the stone blocks' edges exhibit limited accuracy. This is attributed to a combination of factors such as: (i) the presence of the white surface build-up between stone blocks, which visually obscures the true boundaries and misleading SAM2 segmentation, (ii) the use of wide-angle fish-eye images adding a warping effect to the stone blocks and (iii) limitations inherent to the projection process itself, such as slight inaccuracies in camera intrinsic and extrinsic parameters or voxel resolution constraints. Figures 8b & 8c illustrate selected results for detachment and material loss, respectively, in the Pozzi cells area. In both cases, multiple segmentation challenges are evident. Several regions show clear instances of false positive and false negative detections, reflecting the difficulty in accurately delineating degradation boundaries. Additionally, certain well-defined masks exhibit slight spatial offsets from the ground truth, likely due to minor projection inaccuracies, challenging viewpoints, or limitations in the ray projection and voxel intersection mechanisms used during the 2D-to-3D semantic segmentation framework. Notably, some detected masks are located near the edges of the fisheye images (Figure 7), where high optical distortion further complicates both segmentation and projection accuracy. These results underscore the sensitivity of the pipeline to environmental complexity, surface irregularities, and image-based distortions. Table 1 reports on the results of the metric evaluation of the proposed 2D-to-3D semantic segmentation pipeline across three targeted classes. Each class is assessed using four standard segmentation metrics: Precision, Recall, F1 Score, and Intersection over Union (IoU). The stone blocks segmentation in the Minguzzi dataset achieved a precision of 94.81% and recall of 88.41%, indicating that the majority of predicted blocks are correct, and most actual blocks are successfully identified. This balance is reflected in a strong F1 Score of 91.50% and an acceptable IoU of 84.33%, confirming accurate and consistent alignment between the projected 2D semantic labels and the 3D ground truth.

(%)	*2D-to-3D Semantic Segmentation*		
	Minguzzi	Pozzi cells area	
	Stone blocks	Material detachment	Material loss
Precision	94.81	73.53	81.21
Recall	88.41	88.36	85.91
F1 Score	91.50	80.26	83.49
IoU	84.33	67.03	71.66

Table 1. Metrics for the presented 2D-to-3D semantic segmentation.

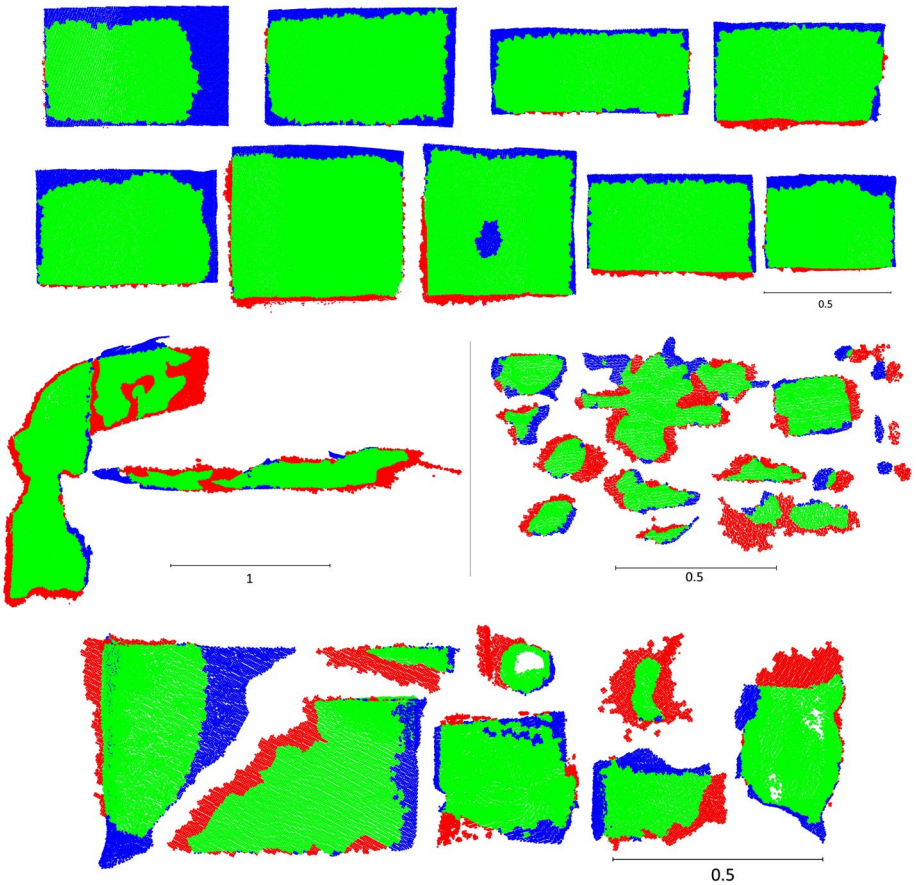

Figure 8. Qualitative evaluation of 2D-to-3D semantic segmentation results across both case studies: stone block segmentation in the Minguzzi staircase (top row); material detachment detection (middle row); and material loss detection (bottom row) in the Pozzi cells area. Green: overlap with ground truth; red: false positives; blue: false negatives. Units in meters.

In contrast, the material detachment class, which represents more irregular and visually ambiguous features, shows lower precision (73.53%) but a high recall (88.36%). This suggests that the method effectively identifies most detachment, instances but also includes a higher rate of false positives, likely due to the subtle and complex textures of degraded surfaces and the over segmented masks. Consequently, the F1 Score drops to 80.26%, with an IoU of 67.03%, reflecting the increased difficulty in accurately delineating these features. The material loss class reports a precision and recall at 81.21% and 85.91%, respectively. This results in an F1 Score of 83.49% and an IoU of 71.66%, indicating relatively consistent performance, in both detecting and projecting these localized areas of material loss. Figure 9 presents the results from the 2D-to-3D semantic segmentation of both case studies, highlighting the ability of our proposed pipeline to create semantically enriched 3D models with metric information.

5. Conclusions

This study presented a framework for enriching 3D reconstructions with 2D semantic segmentation on fish-eye images collected with the multi-camera mobile mapping system ATOM-ANT3D to produce semantically meaningful 3D models for Cultural Heritage (CH) documentation. Two complex case studies are used to evaluate the proposed framework targeting challenging key architectural elements (i.e., stone blocks) and material decay (i.e., material detachment and material loss). Qualitative and quantitative analyses are performed highlighting the strengths and limitations of the approach. Although a fully automatic grounding technique for the isolation of SAM2-generated stone block masks from other architectural elements is difficult due to a combination of factors such as the use of wide-angle fish-eye imagery, complex environment geometry, and non-uniform object surface conditions, requiring expert intervention, the 2D-to-3D projection framework nonetheless achieved robust stone blocks 3D segmentation on the Minguzzi staircase. This confirms that given reliable 2D segmentations, the proposed framework produces 3D labels that align well with the ground truth. In contrast, the coupling of YOLOv5 and SAM2 proved effective in detecting more ambiguous and irregular features, such as material detachment and material loss. While this approach achieved high recall and acceptable overall performance, it faced certain challenges, resulting in slightly lower precision and IoU scores. These challenges include distortions near image borders (caused by fish-eye lenses), surface irregularities, and obstructed object boundaries due to the difficulty of detecting less visually distinct features. Nonetheless, the pipeline demonstrates a viable and efficient method for enriching cultural heritage (CH) 3D models with semantic data. By enhancing the interpretation of 3D models, this study contributes to improved documentation, monitoring, and preservation of CH. Future work will focus on improving the quality of 2D detection masks, increasing projection accuracy, and developing robust automated object detection methods tailored specifically for CH applications.

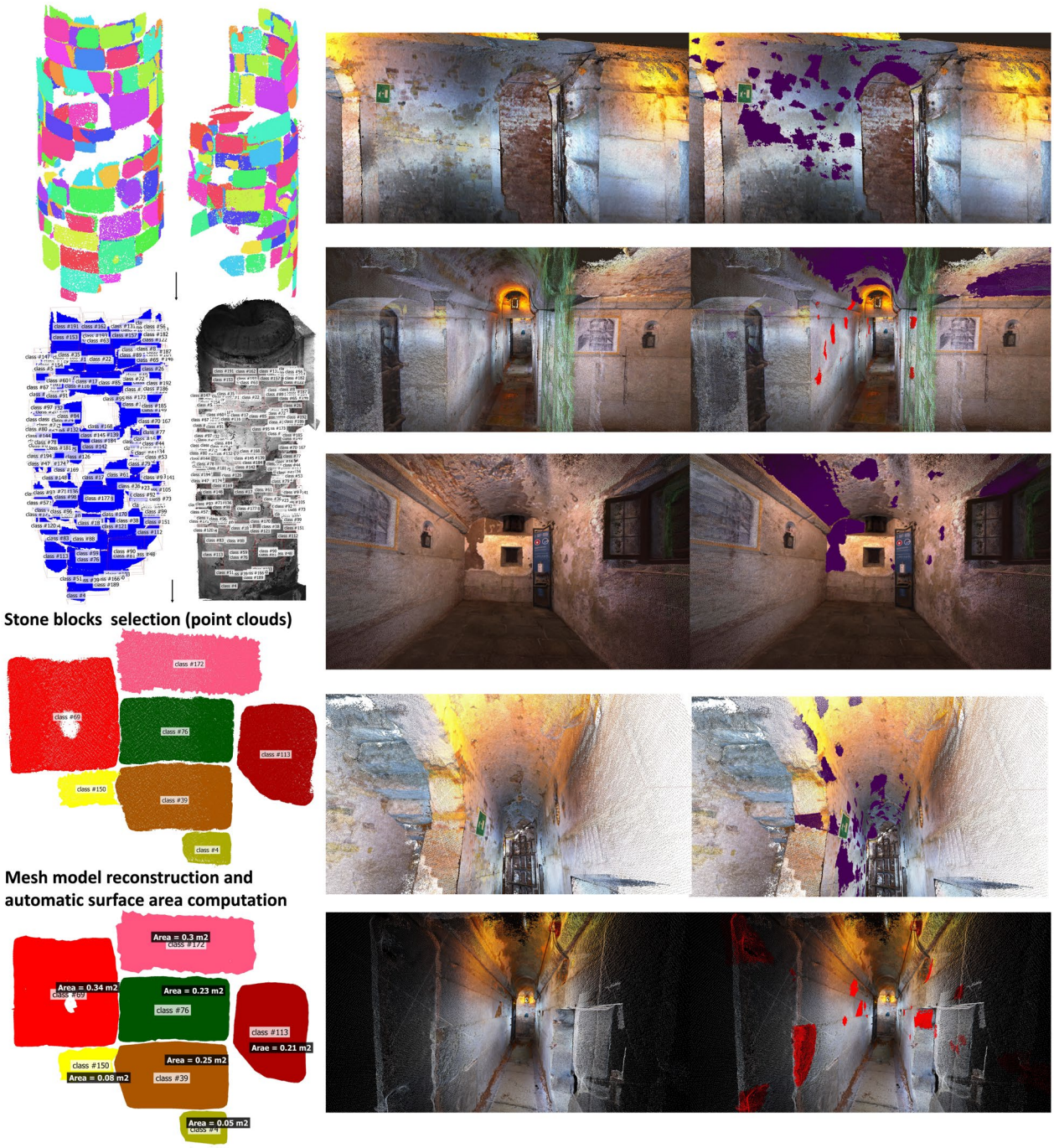

Figure 9. Results of the 2D-to-3D semantic segmentation: stone-block detection and unique label assignment with metric information in the Minguzzi staircase (left); material detachment (purple) and material loss (red) in the Pozzi cells area before and after (right). (point clouds and mesh models)

References

Achille, C., Fassi, F., Mandelli, A., Perfetti, L., Rechichi, F., & Teruggi, S., 2020. From a Traditional to a Digital Site: 2008–2019. The History of Milan Cathedral Surveys. *Digital Transformation of the Design, Construction and Management Processes of the Built Environment,* pp. 331–341.

Alami, A. & Remondino, F., 2024. Querying 3D Point Clouds Exploiting Open-Vocabulary Semantic Segmentation of Images. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLVIII-2/W8-2024,* pp. 1–7.

Berra, E. F., & Peppa, M. V., 2020. Advances and Challenges of UAV SFM MVS Photogrammetry and Remote Sensing: Short Review. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-3/W12-2020, pp. 267–272.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L., 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE TPAMI,* pp. 834–848.

Davison, A. J., Reid, I. D., Molton, N. D., & Stasse, O., 2007. MonoSLAM: Real-time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), pp. 1052–1067.

DeTone, D., Malisiewicz, T., & Rabinovich, A., 2017. SuperPoint: Self-Supervised Interest Point Detection and Description. *Proc. CVPRW,* pp. 337–33712.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., & Sattler, T., 2019. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. *IEEE/CVF CVPR,* pp. 8084-8093.

El-Alailyi, A., Mazzacca, G., Alami, A., Padkan, N., Takhtkeshha, N., Fassi, & F., Remondino, F., 2025a. 2D and 3D Semantic Segmentation for Interpreting and Understanding 3D Heritage Spaces. *Eurographics* (in press).

El-Alailyi, A., Morelli, L., Trybała, P., Fassi, F., & Remondino, F., 2025b. Optimizing Multi-Camera Mobile Mapping Systems with Pose Graph and Feature-Based Approaches. *Remote Sensing* (in review).

Elalailyi, A., Perfetti, L., Fassi, F., & Remondino, F., 2024a. V-SLAM-aided Photogrammetry to Process Fisheye Multi-Camera Systems Sequences. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLVIII-2/W4-2024, pp. 189–195.

Elalailyi, A., Trybała, P., Morelli, L., Fassi, F., Remondino, F., Fregonese, L., 2024b. Pose Graph Data Fusion for Visual- and LiDAR-based Low-Cost Portable Mapping Systems. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLVIII-2/W8-2024, pp. 147–154.

Elhashash, M., Albanwan, H., & Qin, R., 2022. A Review of Mobile Mapping Systems: From Sensors to Applications. *Sensors*, 22(11), 4262.

Fassi, F., Achille, C., Fregonese, L., 2011. Surveying and Modelling the Main Spire of Milan Cathedral Using Multiple Data Sources. *The Photogrammetric Record,* 26, pp. 462–487.

Galanakis, D., Lucho, S., Maravelakis, E., Bolanakis, N., Konstantaras, A., Vidakis, N., Petousis, M., Treuillet, S., Desquesnes, X., & Brunetaud, X., 2024. Segment Anything Model for Scan-to-Structural Analysis in Cultural Heritage. *Proc. EEITE*, pp. 1–7.

Grilli, E. & Remondino, F., 2019. Classification of 3D Digital Heritage. *Remote Sensing*, Vol. 11(7), 847.

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

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., & Girshick, R.B., 2023. Segment Anything. *IEEE International Conference on Computer Vision*, pp. 3992-4003.

Kuo, J., Muglikar, M., Zhang, Z., & Scaramuzza, D., 2020. Redesigning SLAM for Arbitrary Multi-Camera Systems. *Proc. ICRA*, pp. 2116–2122.

Liu, H., Li, C., Wu, Q., Lee, Y.J., 2023a. Visual Instruction Tuning. *NeurIPS.*

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, Jie, Jiang, Q., Li, C., Yang, Jianwei, Su, H., Zhu, J., Zhang, L., 2023b. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv.* arXiv:2303.05499.

Luhmann, T., Robson, S., Kyle, S., & Harley, I., 2007: Close Range Photogrammetry: Principles, Techniques and Applications. *Wiley*.

Agisoft Metashape, 2024. Agisoft: St. Petersburg, Russia, Version 2.2. www.agisoft.com/

Muralikrishnan, B., 2021. Performance Evaluation of Terrestrial Laser Scanners - A Review. *Measurement Science and Technology*, *32*(7), 072001.

Murez, Z., van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V., & Rabinovich, A., 2020. Atlas: End-to-End 3D Scene Reconstruction from Posed Images (Version 3). *arXiv*. arXiv:2003.10432.

Ortiz-Coder, P., & Sánchez-Ríos, A., 2019. A Self-Assembly Portable Mobile Mapping System for Archaeological Reconstruction Based on VSLAM-Photogrammetric Algorithm. *Sensors*, *19*(18), 3952.

Perfetti, L., Bruno, N., & Roncella, R., 2024a. Multi-Camera Rig and Spherical Camera Assessment for Indoor Surveys in Complex Spaces. *Remote Sensing*, 16(23), 4505.

Perfetti, L., Fassi, F., & Vassena, G., 2024b. Ant3D—a Fisheye Multi-Camera System to Survey Narrow Spaces. *Sensors*, 24(13), 4177.

Pritchard, D., Sperner, J., Hoepner, S., & Tenschert, R., 2017. Terrestrial Laser Scanning for Heritage Conservation: the Cologne Cathedral Documentation Project. *ISPRS Annals of the Photogramm. Remote Sens. Spatial Inf. Sci*., IV-2/W2, pp. 213–220.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning Transferable Visual Models From Natural Language Supervision. *Proc. PMLR*.

Réby, K., Guilhelm, A., & De Luca, L., 2023. Semantic Segmentation using Foundation Models for Cultural Heritage: an Experimental Study on Notre-Dame de Paris. *Proc. ICCVW*, pp. 1681–1689.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2015. You Only Look Once: Unified, Real-Time Object Detection. *Proc. CVPR,* pp. 779-788.

Remondino, F., 2011. Heritage Recording and 3D Modeling with Photogrammetry and 3D Scanning. *Remote Sensing* 3, pp. 1104–1138.

Teruggi, S., Grilli, E., Russo, M., Fassi, F., Remondino, F., 2020. A Hierarchical Machine Learning Approach for Multi-Level and Multi-Resolution 3D Point Cloud Classification. *Remote Sensing*, 12(16), 2598.

Torresani, A., Menna, F., Battisti, R., Remondino, F., 2021. A V-SLAM Guided and Portable System for Photogrammetric Applications. *Remote Sensing*, Vol.13(12), 2351.

Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.-H., 2025. Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos. *arXiv*. arXiv:2501.04001

Zhang, K., Mea, C., Fiorillo, F., & Fassi, F., 2024. Classification and Object Detection for Architectural Pathology: Practical Tests with Training Set. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLVIII-2/W4-2024, pp. 477–484.

Zuiderveld, K., 1994. Contrast Limited Adaptive Histogram Equalization., in: Graphics Gems IV. *Academic Press*.