DIGITAL HERITAGE (2025)
S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, and F. Remondino (Editors)

2D and 3D Semantic Segmentation for Interpreting and Understanding 3D Heritage Spaces

Ahmad El-Alailyi[1,2], Gabriele Mazzacca[1,3], Ashkan Alami[1,4], Nazanin Padkan[1,3], Narges Takhtkeshha[1,5], Francesco Fassi[2] and Fabio Remondino[1]

[1] 3D Optical Metrology (3DOM) Unit, Bruno Kessler Foundation (FBK), Trento, Italy
Email: < aelalailyi><gmazzacca><npadkan>< ntakhtkeshha><remondino>@fbk.eu
[2] 3D Survey Group, ABC Department, Politecnico di Milano, Milano, Italy – Email: Francesco.fassi@polimi.it
[3] Dept. Mathematics, Computer Science and Physics, University of Udine, Italy
[4] Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
[5] Department of Geodesy and Geoinformation, TU Wien, Vienna, Austria

Abstract

The 3D digitization of Cultural Heritage (CH) sites has become increasingly requested for documentation, preservation, and analysis applications. Beyond capturing 3D spatial geometry, the semantic interpretation and understanding of digital models are critical for enabling meaningful CH studies and facilitating informed conservation strategies. However, manual annotation and classification of architectural elements and surface pathologies remain labor-intensive and time-consuming, underscoring the need for automated approaches. This study presents a comparative analysis between two distinct semantic segmentation frameworks: (1) a 2D-to-3D pipeline that projects 2D image-based detections onto 3D point clouds produced with V-SLAM data and (2) direct segmentation methods of 3D point clouds acquired with portable LiDAR sensors. These frameworks are evaluated on data acquired using two distinct mobile mapping systems (MMS): (1) a fisheye multi-camera Visual SLAM-based portable system (ATOM-ANT3D) for the 2D-to-3D pipeline; (2) a LiDAR-based MMS (Heron MS Twin Color) for the 3D segmentation methods. Achieved results demonstrate the ability of the proposed frameworks to generate semantically enriched 3D heritage data, with the 2D-to-3D method slightly outperforming the 3D segmentation techniques.

1. Introduction

Accurate and detailed documentation of Cultural Heritage (CH) sites has historically been fundamental for preservation, conservation, and restoration. Traditional documentation techniques started with manual drawings and photographic archives utilizing hand surveys with simple tape measurements, point laser distance measurers, plumb lines, and theodolites. In recent decades, advancements in 3D digital technologies have significantly enhanced CH documentation processes. Specifically, Terrestrial Laser Scanning (TLS) and photogrammetry have nowadays become the standard techniques, enabling precise 3D geometric capture of heritage structures and environments [EGV*07; EBR*08; RR10; FAF11; Rem11; KBKB17; VRP*23; VRR*23]. However, while these methods produce highly accurate geometric data, the resulting documentation often lacks essential semantic information required for comprehensive interpretation and analysis of architectural elements and surface pathologies. Several studies have emphasized the importance of integrating geometric 3D point clouds and models with semantic understanding to enable more comprehensive and meaningful CH documentation [GR19; TGR*20; YHL23; ZF25]. Building on this, recent research has begun exploring 2D semantic segmentation as a computationally efficient alternative to 3D point cloud-based semantic segmentation. These image-based methods offer a promising direction, reflecting the growing interest in efficient interpretation and classification techniques within the CH documentation workflows.

1.1 Paper's aim

Motivated by the need for comprehensive and semantically enriched CH sites documentation, this study proposes and compares two distinct semantic segmentation methodologies: (1) a 2D-to-3D semantic segmentation approach combining multimodal large language models (MLLMs) and conventional object detection methods; (2) 3D point cloud semantic segmentation approaches based on machine or deep learning methods. The 2D-to-3D approach is applied to data collected with the ATOM-ANT3D multi-camera Visual SLAM-based portable system [EPFR24], whereas 3D point clouds are acquired with the HERON MS TWIN Color MMS [Gex25] (Figure 1). The innovative aspects of the work include:

- insights into 2D/3D segmentation strategies to enhance the interpretability and usability of 3D heritage documentation;
- use of MLLM in CH scenarios.

The work intends to shed light on semantic segmentation approaches applied to V-SLAM- or LiDAR-based data to support the heritage community in need of enriched 3D surveying data with meanings and interpretations.

2. Related works

2.1 Mobile mapping systems (MMS)

Mobile mapping systems (MMS) equipped with LiDAR (Light detection and ranging) sensors, cameras, or a fusion of both, have emerged as a promising alternative to static classical survey techniques, offering flexibility and significantly reducing data acquisition times [EAQ22; XZY*22] proving effective in diverse and complex applications [NMR*17; DTF*21; ALL*23; TABF24]. Vision-based MMS have gained popularity for rapid image data acquisition and efficient 3D reconstruction workflows [OS19; TMBR21; EPFR24; ETM*24; PFV24] producing texture-rich data suited for photorealistic modeling and benefiting from recent advancements in image-based semantic segmentation. However, their reliance on adequate lighting, surface texture, and the challenges posed by motion blur, occlusions, and geometric distortions, especially when using wide-angle/fisheye lenses,

DOI: 10.2312/dh.20253047

requires careful handling during data acquisition and processing. On the other hand, LiDAR-based MMS have been predominant in 3D mapping, operating in environments with limited texture and varying lighting conditions. Often vehicle-mounted, backpack-carried, or handheld, LiDAR systems efficiently document complex geometries in diverse indoor and outdoor contexts [NMR*17; Będ24] and in CH documentation applications [DTF*21; ALL*23; TABF24]. They are typically complemented by onboard cameras for color information; however, these cameras generally provide lower-resolution textures compared to specialized photogrammetric setups, limiting their capability to capture fine surface details.

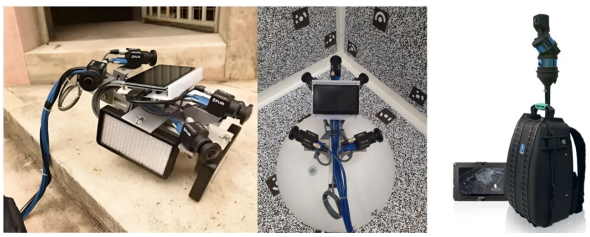

Figure 1. *The two mobile mapping systems used in this study are: ATOM-ANT3D (left) and HERON MS TWIN Color (right).*

2.2 Semantic segmentation and feature detection

2.2.1 Vision- and Multimodal large language models (VLM/MLLM). Classical machine learning methods laid the foundation for modern image segmentation. Support Vector Machines (SVMs) classify pixels by identifying optimal hyperplanes in high-dimensional spaces [CGRL20]. Decision trees [MFP*14] distinctly classify objects by dividing them based on pixel intensity values. Conditional Random Fields (CRFs) represent images as probabilistic graphs, refining segmentation by modeling contextual relationships between pixels. In contrast, deep learning methods like Fully Convolutional Networks (FCNs) [LSD15] perform pixel-level classification without manual feature extraction. Architectures such as DeepLab [CPK*17] use dilated convolutions and Atrous Spatial Pyramid Pooling (ASPP) for multi-scale context, while Pyramid Scene Parsing Network (PSPNet) [ZSQ*17] captures global context via pooling strategies.

Advances in multi-modal Artificial Intelligence (AI) have produced diverse models optimized for tasks like object detection, segmentation, and visual grounding featuring varied architectures (e.g., transformer-based, multimodal, etc.) and approaches (e.g., prompt-driven segmentation, zero-shot detection). Segment Anything Model (SAM) [KMR*23] is a foundational segmentation model utilizing vision transformers trained on extensive segmentation datasets. It uses prompt encoders (e.g., a point on an object or a bounding box) to condition the mask decoder, predicting object masks from embeddings. Large Language and Vision Assistant (LLaVA) [LLWL23] integrates visual encoders with large language models (LLMs) for vision-language reasoning tasks, though without direct segmentation outputs. Sa2VA [YLZ*25] is a unified architecture for dense grounded understanding of images and videos, marrying SAM2 with LLaVA. It uses SAM2's visual encoding and the LLM's self-attention over combined visual and textual tokens, allowing natural language interactions that yield segmentation masks for referenced objects. Grounding DINO [LZR*23] provides open-set detection using transformer-based fusion of vision and language inputs, whereas Grounded-SAM [RLZ*24] integrates Grounding DINO and SAM for prompt-driven segmentation within bounding boxes.

2.2.2 3D point clouds segmentation. Supervised methods learn a mapping from 3D inputs to point labels using annotated data. Classical methods, such as Random Forests [Bre01; GDPR18], classify points based on handcrafted features, combining individual tree predictions through majority voting. PointNet [QSMG16], a pioneering point-based network, processes points independently using multilayer perceptrons (MLPs) and aggregates global context via max-pooling. PointNet++ [QYSG17] adds hierarchical grouping for capturing local context and multi-scale features. PointCNN [LBS*18] generalizes convolutional neural networks (CNNs) to point clouds by learning an X-transformation that canonicalizes local neighborhoods, facilitating structured, convolution-like operations. Graph-based approaches like Superpoint Graph (SPG) [LS17] segment clouds into geometric primitives (i.e., superpoints) and apply graph convolutional networks for classification. Point Convolutional Networks utilize continuous convolution kernels that operate on point coordinates, such as PointConv [WQF18] and KPConv [TQD*19]. Inspired by Transformers in NLP and image analysis, self-attention networks have been adapted to point clouds. Point Transformer [ZJJ*20] is a fully attention-based network that treats the point cloud as a set and learns interactions among points through self-attention. Superpoint Transformer [RRL23] partitions point clouds into a hierarchical superpoint structure, leveraging the self-attention mechanism to group superpoints in meaningful regions, improving efficiency and speed in large-scale point cloud segmentation. Given the high annotation cost, unsupervised methods have emerged to segment or pre-train features from unlabeled point clouds. Traditional geometry-based techniques, including region growing, density clustering, and RANSAC-based primitive fitting [ZZCL19], are effective for simple geometries but struggle with detailed, irregular shapes. Recent deep unsupervised methods, such as GrowSP [ZYWL23], leverage neural feature extraction and iterative merging into meaningful segments without human labels.

2.3 Applications in Cultural Heritage

The combination of 2D object detection and 3D spatial enrichment has been studied in CH. For instance, Faster R-CNN was used by Pathak et al. [PSW*21] to identify damages in 3D model rendered views. Similarly, UV-mapped textures or orthoimages from 3D reconstructions are used by texture-based classification methods [GR19]. Recent methods shift toward working directly on 2D images, rather than renders from 3D models. Foundational vision-language models made this process more accessible without training requirements. Réby et al. [RGD23] used the Segment Anything Model (SAM) with Grounding DINO to achieve object localization and semantic labelling at the image level. Similarly, Galanakis et al. [GLM*24] explored SAM's capabilities for detailed structural analysis and segmentation at the individual stone level. In 3D-based point cloud semantic segmentation, machine learning methods have been applied in the CH field [FKP*20]. Random Forests (RF) were adopted due to their adaptability to limited training data and their ability to handle variable heritage geometries [ZF25]. For example, Multi-Level Multi-Resolution (MLMR) classification, hierarchically segments point clouds from coarse to fine detail, balancing computational cost with accuracy [TGR*20]. As deep learning gained traction, researchers have been exploring different methods in the field of CH. A deep learning framework based on an enhanced DGCNN (Dynamic Graph Convolutional Neural Network) architecture was proposed for semantic segmentation of 3D point clouds applied to the diverse ArCH dataset of historical buildings [PPM*20]. Recent deep learning-based fusion approach combining image segmentation and Point Transformer networks was proposed to

improve semantic segmentation in cultural heritage [BMB*24]. However, a direct comparative analysis spanning traditional machine learning, deep learning, and recent multimodal large vision-language models (MLLMs) for 2D-to-3D semantic segmentation remains underexplored in the literature.

3. Case study: Palazzo Ducale in Venice, Italy

This paper presents the architectural documentation and 3D semantic interpretation of the ground-floor arcade corridor in the inner courtyard of the Palazzo Ducale (Doge's Palace) in Venice, Italy. A masterpiece of Venetian Gothic and Renaissance architecture, the palace features a continuous system of arcades and loggias surrounding the courtyard. The study focuses on the section running along the Piazzetta and Renaissance wings (Figure 2), which historically served both ceremonial and practical functions as a sheltered passageway.

Figure 2. *The surveyed area of Palazzo Ducale in Venice: inner courtyard (left) and arcade corridor (right).*

The inner courtyard arcade corridor of the Palazzo Ducale is surveyed using the ATOM-ANT3D fisheye multi-camera MMS [EPFR24; PFV24] and the HERON MS TWIN Color LiDAR MMS[1]. Using ATOM-ANT3D, a total of 19,755 images (3,951 images per camera) is acquired and processed using a V-SLAM and 3D reconstruction pipeline [EMT*25] (Figure 3a). Then, a dense 3D cloud is generated (Figure 3b) to be used in the semantic segmentation process. Considering the use of wide-angle fisheye lenses for open-space data acquisition, capturing extensive scene information but potentially introducing noise at greater depths and near lens boundaries due to distortion, and a segmentation workflow based on 2D-to-3D semantic projection (Section 4.1.2), a down-sampled point cloud of approximately 5 million points was used. This choice balances scene coverage, projection accuracy, and computational efficiency. The resulting 3D labels can later be interpolated onto the full-resolution dense cloud. On the other hand, the Heron MS Twin Color LiDAR MMS data acquisition provided a high-resolution point cloud composed of ca 101 mil points (Figure 3c). To reduce computational time of the semantic segmentation process, the LiDAR point cloud size is subsampled at a 1-centimeter resolution, resulting in a ca 51 mil points dataset. The semantic classes considered for spatial interpretation and conservation purposes of the heritage area include arches, mouldings, openings (e.g., windows and doors), pavement/floor, columns, walls, and cracks.

4. Methodologies

The following sections present two approaches, based on different sensorial data (V-SLAM and LiDAR), for the semantic segmentation of heritage scenes to support interpretation, understanding, and conservation.

4.1 2D-to-3D semantic segmentation with V-SLAM data

4.1.1 2D semantic segmentation. The extraction of architectural features corresponding to predefined classes is performed using the MLLM Sa2VA [YLZ*25], and the conventional object detection model YOLOv8 [JCQ23] is specifically used for crack detection. Sa2VA uniquely integrates a vision-language model (i.e., LLaVA) with the Segment Anything Model (SAM), allowing segmentation guided by detailed textual descriptions rather than simple class names. This capability is crucial, as common open-vocabulary models struggle with accurately segmenting fine-grained architectural components. Hence, it is chosen for its capabilities and state-of-the-art performance.

Initial attempts to apply semantic prompt queries directly to the raw fisheye images demonstrated inconsistent mask accuracy. Therefore, fisheye images are converted into rectilinear ones using OpenCV's [Bra00] fisheye calibration module with the simplified Kannala-Brandt model [KB06], generating pixel-wise mappings for remapping images to simulate a pinhole camera view with zero distortion [ZEPF24]. Furthermore, generic or high-level prompts initially yielded ambiguous or incomplete results. To address this, we developed a curated set of task-specific semantic queries, incorporating contextual and architectural terminology aligned with the model's vision-language representations, significantly enhancing retrieval precision for the defined classes.

For object segmentation targeting cracks located on columns, YOLOv8 [JCQ23] is employed. Specifically, a pre-trained version of YOLOv8 from Padkan et al. [PBMR23], trained on approximately 400 annotated images encompassing a variety of crack types, is used. Despite the availability of numerous object detection and semantic segmentation models, we chose the YOLO family due to its effectiveness when trained on relatively small datasets and its performance in detecting small objects such as cracks. To improve detection resolution, each input image is divided into four smaller sub-images, allowing for a more focused crack analysis. YOLOv8 is applied independently to each sub-image, and binary masks are merged to reconstruct a full-size mask aligned with the original image dimensions. However, the domain gap between training data and our test environment occasionally caused false positives, such as misclassifying structural borders, wall features, and spacing between the floor stones as cracks. To address this issue, Sa2VA-generated column masks are used to isolate cracks within column regions, effectively reducing false detections from irrelevant background areas.

Figure 3. *Top-view of the ATOM-ANT3D trajectory and sparse point cloud generated with the developed V-SLAM approach (a); close-view of the photogrammetric dense point cloud (b); HERON point cloud of the courtyard area (c).*

[1] https://shorturl.at/BNrAO

4.1.2 2D-to-3D semantic projection. To project semantic segmentation masks onto the 3D point cloud, a method adapted from [AR24] is used. The pipeline combines voxel-based ray casting and camera model projections. The input dense point cloud is voxelized at a user-defined resolution, and a ray casting scene is generated using Open3D [ZPK18]. Voxelization plays a critical role by preventing rays from passing through sparse regions of the point cloud and mistakenly labeling geometry behind the actual surface acting as an occupancy grid, allowing rays to terminate on the first encountered surface even when no exact point is directly hit. For each oriented camera, rays are cast into the scene based on the camera's intrinsic and extrinsic parameters, identifying visible surface points and transferring the corresponding semantic labels from the 2D masks onto the 3D voxels encapsulating the 3D points intercepting the casted rays, coupled with a neighborhood expansion search (Figure 4).

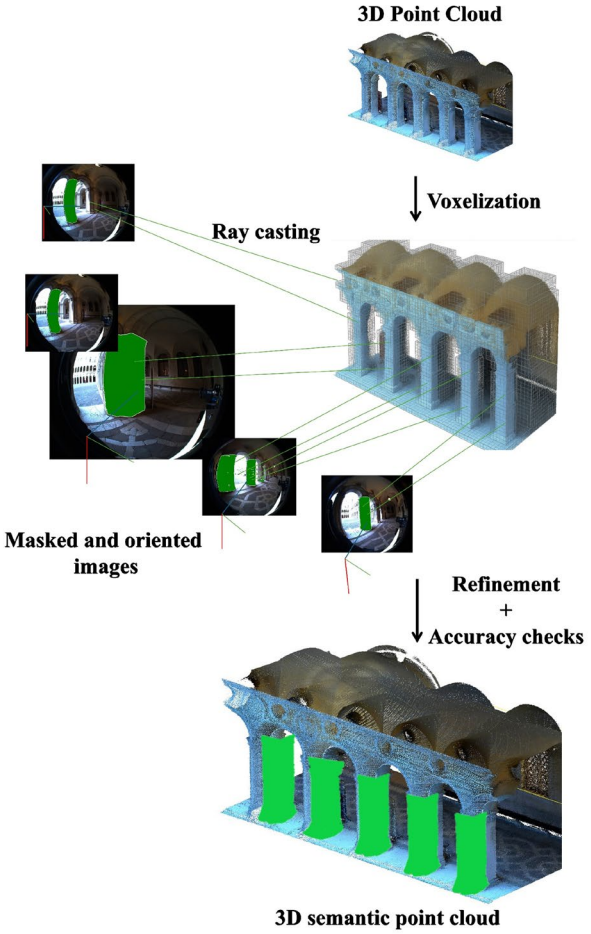

Figure 4. *The proposed 2D-to-3D semantic segmentation projection scheme for the V-SLAM data (images and point clouds).*

The method's accuracy relies on multiple factors, including voxel resolution, mask quality, and precision of camera parameters. Coarse voxelization, segmentation errors, and calibration drift can introduce mislabeling or outliers into the final 3D semantic output. Therefore, to enhance efficiency and robustness, a multi-view voting strategy is implemented, where each point accumulates label votes from all valid projections, and the most voted label is assigned. While this approach improved labeling quality in parts of the scene, it simultaneously introduced outliers in areas affected by variations in mask segmentation accuracy. To further address missed labeling on the 3D point cloud caused by the 2D-to-3D projection limitations, we developed a dedicated point cloud post-processing pipeline. Given the color/label-to-class mapping, normals are estimated to capture local surface geometry, and a region growing algorithm is applied exclusively to unlabeled regions where 2D masks failed to project correctly. Labels from initially labeled points are propagated to neighboring unlabeled points based on spatial proximity (i.e., defined search radius and label majority fraction) and normal similarity, ensuring that only points with a sufficient fraction of similar labeled neighbors are incorporated.

While effective for larger architectural elements, the voxel-based approach with fine, small-scale features such as cracks, requires significantly finer voxel resolutions. To address this, a complementary technique was adopted, leveraging the inherent 2D-to-3D correspondences provided by Metashape [Agi25] between the original images and the dense point cloud. Specifically, we used the generated 2D masks to directly select their corresponding 3D points on the dense cloud, ensuring the annotation of cracks. This method is well-suited for isolating individual, small-scale classes like cracks rather than large-scale multi-class segmentation.

4.2 3D Semantic segmentation on LiDAR data

4.2.1 3D point cloud semantic. To assess the effectiveness of 3D point cloud semantic segmentation on the LiDAR-based 3D point cloud (Figure 3c), three well-established methods based on supervised machine/deep learning are employed for comparative analysis [GR20; BMB*24]: Random Forest, Point Transformer, and Superpoint Transformer. In all cases, a combination of covariance features [WJHM15] and sensor-based features is provided to the architectures to facilitate convergence and yield prediction models with high levels of accuracy.

4.2.2 Data preparation: annotations, training, test, and validation. A subset of the dense point cloud, which contained most of the object classes of interest, generated by the Heron MS Twin Color MMS is selected and partitioned into training and validation sets, divided into ca 70% for training and ca 30% for validation. Both subsets are manually annotated and used during the training phase of the 3D pipelines, combined with covariance and sensor-based features. The covariance features Anisotropy, Linearity, Normal change rate, Omnivariance, Planarity, Roughness, Sphericity, and Verticality are calculated at different radii (0.2m, 0.4m, 0.6m, 0.8m, and 1m) using CloudCompare [Clo24]. The model demonstrating the best performance on the validation set is selected for the final evaluation on unseen data. Both subsets are meticulously chosen to ensure they represent the entire dataset accurately. For a consistent evaluation across the methods, the Random Forest, Point Transformer, and Superpoint Transformer models are trained to predict two additional classes, Vault and other objects present in the corridor, beyond the classes segmented by the 2D-to-3D projection pipeline. Since the 2D-to-3D dataset did not include Vault (i.e., due to weak 3D reconstruction on textureless surfaces) or other objects not included in the 2D image segmentation search, either in the labeling stage or on the point cloud itself, points corresponding to these classes are removed from the test point cloud before evaluation. All segmentation results are subsequently reprojected onto this reduced test cloud. To ensure comparability, points predicted as Vault or other objects by the 3D-based methods are reassigned to the "unlabeled" category (class Other), consistent with the 2D-to-3D approach where unclassified points are similarly labeled. This strategy guarantees a fair evaluation framework across all methods, with the class *Other* representing either unlabeled background, missed classifications, and points

originally predicted as Vault and other objects, now reassigned. Hence, class *Other* is not a meaningful architectural element, but a catch-all noise class.

5. Results, evaluations, and discussion

Sa2VA 2D segmentation successfully segmented features in the fisheye images; however, its performance did not generalize across all classes. The distortion introduced by the 190° field-of-view lenses adversely affected certain classes' segmentation, leading to extended masks beyond object boundaries or erroneously covering unwanted regions in the images (Figure 5).

Figure 5. *The impact of image un-distortion on MLLM detection and segmentation performance for the wall (left) and openings (right) classes, with fisheye (top) and equirectangular images (bottom).*

Figure 6. *Detected and segmented architectural elements from the undistorted image dataset.*

This can be attributed to multiple factors. Features near the image edges, where distortion is highest, often appear warped and inconsistent. Moreover, detecting complex structures depends on clearly differentiating features from the broader global context. As MLLMs are primarily trained on rectilinear images, their performance can be limited by the lack of robust learned representations for fisheye images. Therefore, applying calibration as a pre-processing step helped correct distortions, preserve spatial accuracy, and improve segmentation reliability (Figures 5 & 6). The proposed 2D and 3D methodologies for extracting semantic classes from the surveying data are evaluated using qualitative and quantitative analyses against a manually annotated ground truth point cloud serving as the reference. In addition to visual assessments, by inspecting the segmented point clouds against the annotations, standard segmentation metrics, including precision, recall, and F1-score, are computed to objectively quantify performance (Tables 1 and 2).

(%)	2D-to-3D Semantic Segmentation	3D Semantic Segmentation		
		R.F.	P.T.	S.T.
Weighted Precision	93.48	78.50	86.19	89.81
Weighted Recall	89.80	78.21	85.31	90.28
Weighted F1 Score	91.03	76.50	85.22	89.92

Table 1. *Table reporting the metric evaluations including the Weighted Precision, Weighted Recall, and Weighted F1 Score as percentages for the presented approaches. R.F. = Random Forest; P.T. = Point Transformer; S.T. = Superpoint Transformer*

%	F1 Score per class			
	2D-to-3D Semantic Segmentation	3D Semantic Segmentation		
		R.F.	P.T.	S.T.
Arches	93.64	66.18	77.36	98.81
Mouldings	79.97	28.68	69.79	65.32
Openings	84.02	45.01	73.55	63.77
Pavement/ Floor	84.70	99.51	99.06	98.74
Columns	94.41	98.02	90.64	98.56
Wall	90.27	63.44	76.79	86.33
Other	33.61	95.86	93.70	91.56
Avg. (exc. Other)	87.83	66.81	81.20	85.25

Table 2. *Per-class F1 Score comparison between the presented approaches. The last row reports the average F1-score computed across the main architectural classes excluding the "Other" class. Results are reported as percentages across different architectural elements. R.F. = Random Forest; P.T. = Point Transformer; S.T. = Superpoint Transformer*

Particular attention is given to class-wise evaluations to highlight differences in capturing complex spatial relationships and delineating class boundaries. This combined analysis provided robust validation of each method's effectiveness in both visual interpretation and quantitative performance for CH 3D data enrichment. Figure 7 presents a qualitative comparison of each technique's results. Our proposed 2D-to-3D segmentation method produced well-defined classes, though some noise appeared (i.e.,

mislabeled points), especially around floor-level element connections, due to inaccuracies in 2D mask boundaries, slight intrinsic/extrinsic calibration errors, voxel resolution, and noisy 3D points intercepting casted rays. The Random Forest approach results show well-defined columns, arches with noticeable noise, and pavement/floor. However, the wall, mouldings, and openings are poorly interpreted. Results from the Point Transformer-based method indicate generally well-defined columns, except for outliers on flat areas that are misinterpreted as a wall class. Arches are clearly defined but include similar noise patterns to those seen in columns. Wall and floor are accurately segmented. Openings showed mixed results, with accurate segmentation for windows but incorrect labeling for doors (i.e., misinterpreted and classified as wall). Mouldings are mostly defined but show partial misclassification as walls in some areas. The Superpoint Transformer method demonstrated ca similar results to Point Transformer for mouldings and floor, but with higher segmentation accuracy on columns, arches, and wall. The visual analysis of the results is consistent with the quantitative evaluations reported in Tables 1 and 2. The proposed 2D-to-3D semantic segmentation method, along with the Point Transformer and Superpoint Transformer predictions, achieved ca comparable accuracy across all weighted metrics in Table 1 (i.e., Precision, recall, and F1 score), with Point Transformer showing slight underperformance. In contrast, the Random Forest method demonstrated lower overall accuracy. The 2D-to-3D method achieved ca consistent high accuracy in the per-class F1 scores for arches (93.64%), mouldings (79.97%), openings (84.02%), floor (84.70%), columns (94.41%), and wall (90.27%). Nonetheless, it demonstrated the lowest accuracy for the pavement/floor class compared to other methods. Random Forest exhibited strong performance for columns (98.02%) and pavement/floor surfaces (99.51%), but struggled in capturing mouldings (28.68%), openings (45.01%), and wall (63.44%). Similarly, the Point Transformer achieved strong performance on columns (90.64%) and pavement (99.06%), yet introduced some outliers on the wall area, leading to lower segmentation accuracy (76.79%). Superpoint Transformer demonstrated robust and consistent segmentation across most classes, notably achieving near-perfect scores for arches (98.81%), pavement/floor (98.74%), and columns (98.56%) while underperforming on mouldings (65.32%) and openings (63.77%). The low F1 score in the *Other* class for the 2D-to-3D approach reflects how objects in images that are not included in any class definition and query prompting are left unlabeled. On the other hand, the 3D semantic segmentation approaches always try to perform a prediction; therefore, the Vault and other objects that are predicted now match the defined class Other. To further highlight the overall performance and comparability between methods, an average F1 score excluding the 'Other' class was computed. The achieved results were 87.83% for the 2D-to-3D method, 66.81% for Random Forest, 81.20% for Point Transformer, and 85.25% for Superpoint Transformer confirming the overall comparable performance between the 2D-to-3D method and 3D deep learning methods. A key distinction lies in the requirement for training and manual annotations. The 3D-based methods required a manually annotated 3D training set. This process could be time-consuming and resource-intensive when applied across multiple CH sites with varying styles and scalability of architectural elements. In contrast, the 2D-to-3D pipeline using Sa2VA, which is pre-trained on large-scale diverse image-text datasets, is applied without additional dataset-specific training. However, in the context of cultural heritage, where features can be highly detailed, non-standardized, and visually subtle, prompt engineering becomes a critical step to guide the model using descriptive textual prompts. Thus, while Sa2VA eliminates the need for dataset-specific training, it still demands user expertise and interaction to tailor the prompts appropriately.

Figure 8 presents the results of the 2D-to-3D semantic segmentation comparison with the Superpoint Transformer, selected as the most accurate among the 3D semantic segmentation methods, on two sections of the arcade corridor.

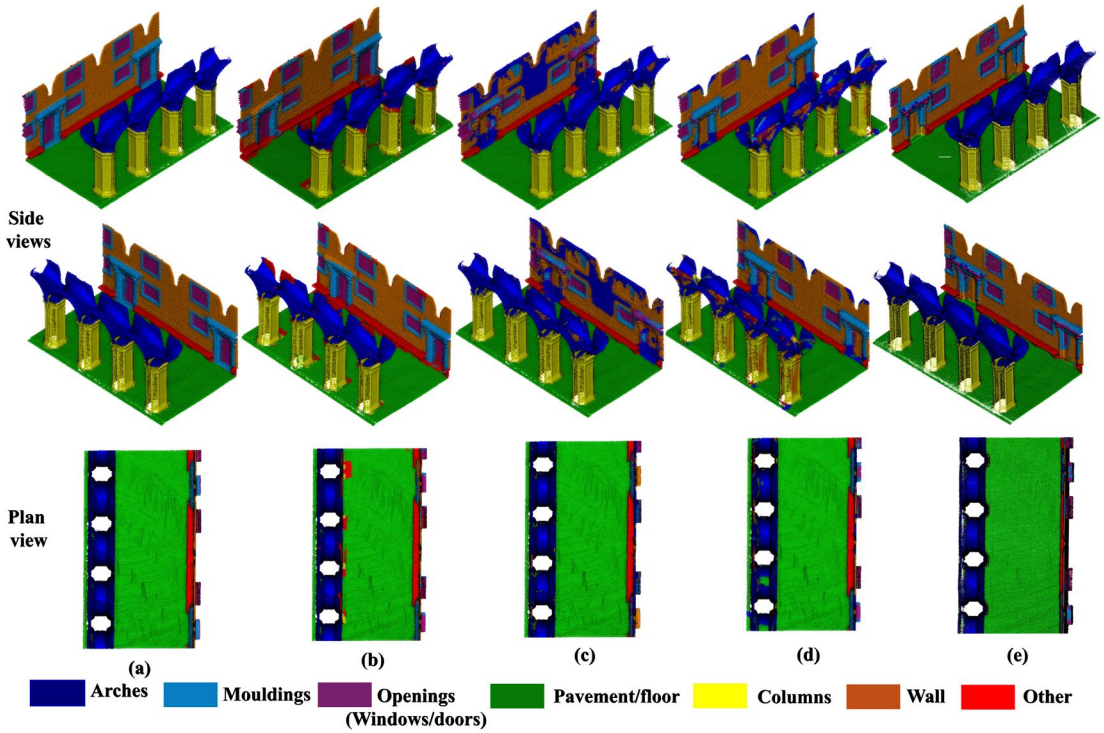

Figure 7. *Semantic 3D point clouds and object classes: (a) ground truth (manually annotated) dataset; (b) 2D-to-3D semantic segmentation results; (c) 3D prediction results using Random Forest; (d) Point Transformer and (e) Superpoint Transformer.*

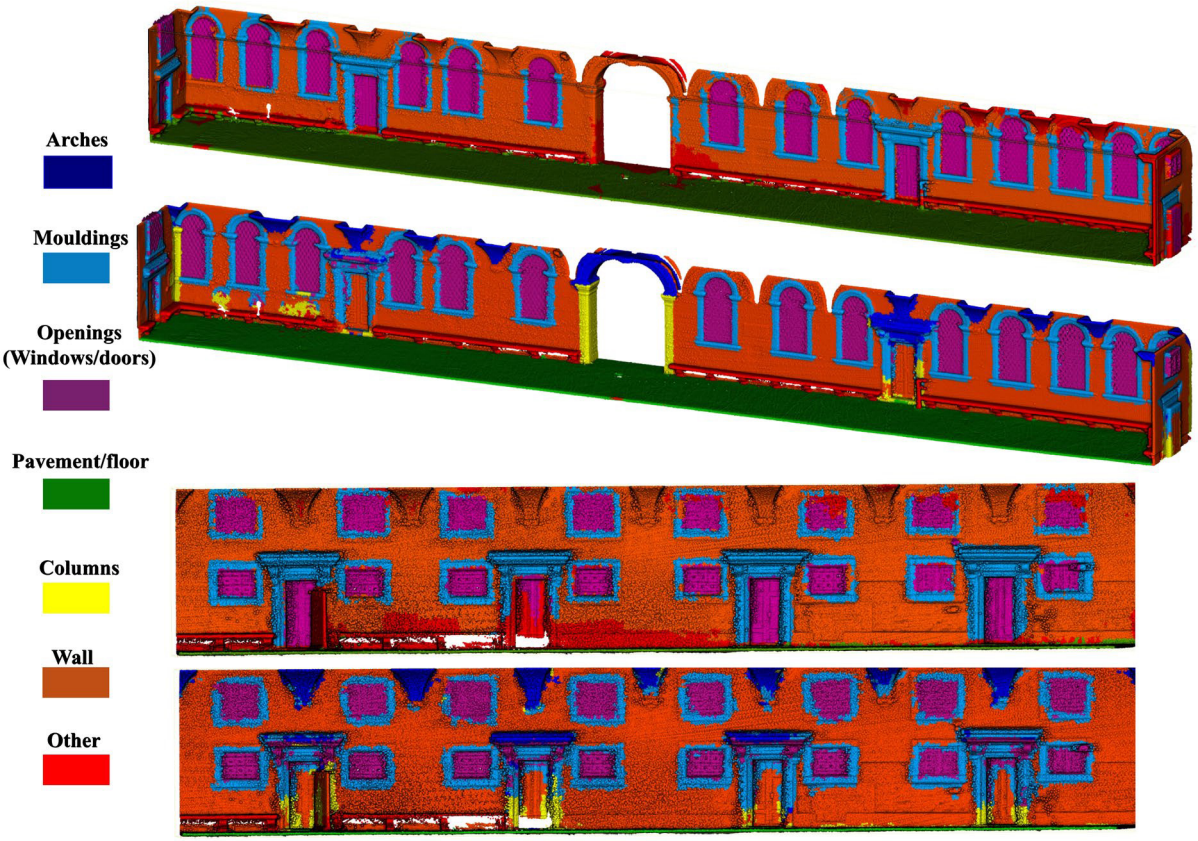

Figure 8. *Results of the 2D-to-3D semantic segmentation approach (top) and Superpoint Transformer (bottom) on two sections of the surveyed area.*

The 2D-to-3D approach struggled slightly to define the full shape of some of the mouldings. Additionally, the Superpoint Transformer predicted some elements as columns and arches, even though these classes are not present in the evaluated sections. The results demonstrate the generalizability of the methods at larger spatial scales. Direct crack detection on 3D point clouds remains a challenging task. While modern 3D point cloud semantic segmentation approaches perform well for distinct and large-scale object classes, they often struggle with small/tiny scale detections. Cracks exhibit extremely small spatial dimensions compared to the overall scene, and the resolution of typical 3D point clouds is often insufficient to capture the fine details required for reliable detection. Consequently, point cloud segmentation algorithms face difficulties in distinguishing cracks from normal surface variations. In contrast, as described in Section 4.1.1, crack detection based on image analysis benefits from richer texture information and color gradients, enabling a higher degree of detection efficiency. By merging the column masks obtained through Sa2VA with the YOLOv8-based crack detection (Figure 9), cracks could be isolated on the segmented columns, removing outlier detections from surrounding regions. Subsequently, given the 3D results, efficient 3D crack localization and visualization are performed. Figure 10 presents the results of our 2D-to-3D columns semantic segmentation. Additionally, the proposed crack detection and projection pipeline results are presented in Figure 11. Nevertheless, some outliers persist at the column level, which we attribute to the absence of a specialized training dataset specifically targeting this typology of cracks, as our model is trained on a more generalized crack detection dataset. Additionally, factors such as camera-to-object distance, blur, and over-exposure affect the efficiency of the 2D crack detection and segmentation.

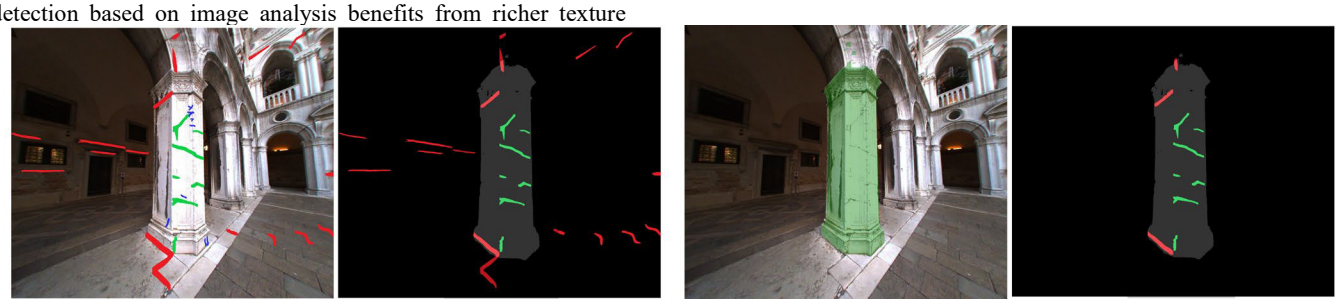

Figure 9. *Crack detection performed with YOLOv8 (left) and column-only crack detection achieved by integrating Sa2VA segmentation masks with YOLOv8 (right). Correct detections are presented in green, incorrect detections in red, and missed detections in blue*

Figure 10. *Semantic segmentation results for the class "column" based on the 2D-3D approach.*

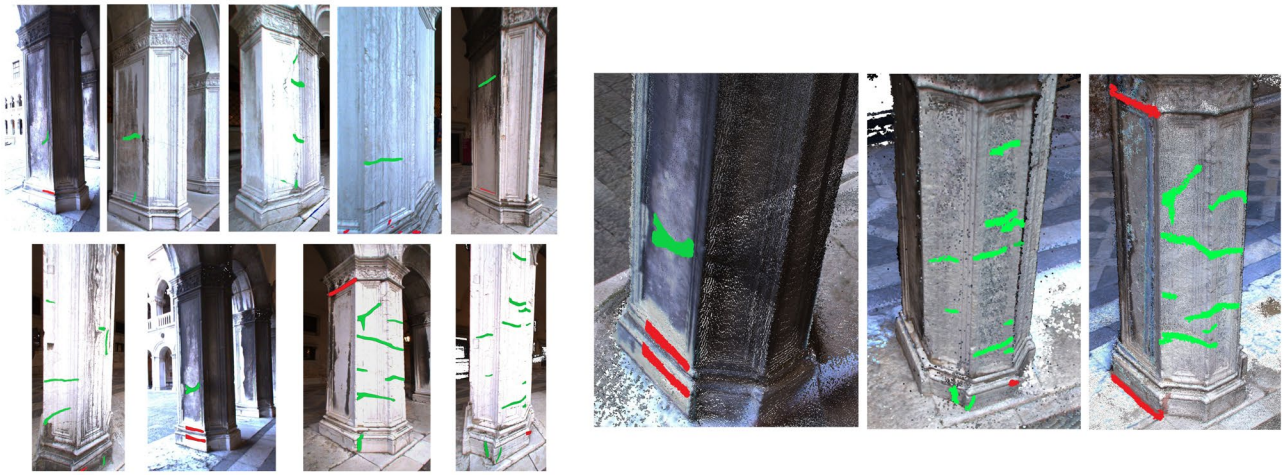

Figure 11. *Cracks detected in the images (left) and a close-view of their subsequent projections onto the 3D columns (right).*

6. Conclusions

This study aimed to investigate how actual AI-based semantic segmentation methods could be applied to V-SLAM data (i.e., images and point clouds) or directly to LiDAR data (i.e., point clouds). Semantic segmentation results are nowadays fundamental to allow better interpretation and understanding of Cultural Heritage 3D models. We presented a 2D-to-3D semantic segmentation framework that projects elements detected in the images through a MLLM onto dense 3D reconstruction, as well as three established 3D point cloud segmentation approaches: Random Forest, Point Transformer, and Superpoint Transformer. While all approaches are capable of segmenting major architectural elements, varying levels of noise are observed in each method. Qualitative and quantitative evaluations confirmed that the 2D-to-3D segmentation pipeline, Point Transformer, and Superpoint Transformer-based predictions produced comparable results, with the 2D-to-3D approach achieving slightly higher overall accuracy. The Random Forest method exhibited a lower overall accuracy compared to the aforementioned approaches. Furthermore, the proposed crack detection and projection pipeline proved effective in isolating fine-scale crack features on segmented columns, overcoming limitations inherent to direct 3D point cloud-based crack detection. Subsequently, when combined with the 3D results, we are able to project, localize, and interpret cracks on the 3D point cloud. Overall, the integration of semantic information into 3D reconstructions provides a robust and scalable solution for enhancing CH site documentation and preservation. As a general guideline, the authors highlight that MLLMs are primarily pre-trained on common imagery compared to the manual annotation and dataset-specific training required by 3D point cloud–based semantic segmentation methods; therefore, detecting unique and highly specialized architectural features requires careful prompt engineering and, in some cases, custom fine-tuning or the provision of reference examples. Moreover, as the scale of the dataset increases, greater user supervision becomes necessary to identify and filter out outlier mask detections produced by the MLLM (subsequently projected to the 3D data). Future developments will focus on advancing towards robust and fully automated segmentation pipelines and accurate 2D-to-3D projection techniques, particularly generalized for complex heritage datasets.

Acknowledgments

This work is partially funded by "Boostech Valorization Program 2022" funded by the Italian "Piano Nazionale di Ripresa e Resillienza—NextGenerationEU with the goal of industrializing the Ant 3D prototype, which is already the subject of the patent proposal n° 102021000000812. (24 January 2023).

References

[Agi25] Agisoft Metashape (Version 2.2). Agisoft: St. Petersburg, Russia, www.agisoft.com/, 2025.

[ALL*23] ARICÒ, M., LA GUARDIA, M., LO BRUTTO, M., et al. "Mobile Mapping for Cultural Heritage: the Survey of the Complex of St. John of the Hermits in Palermo (Italy)". Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLVIII-1/W1-2023, 2023, 25–32.

[AR24] ALAMI, A. and REMONDINO, F. “Querying 3D Point Clouds Exploiting Open-Vocabulary Semantic Segmentation of Images”. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLVIII-2/W8-2024, 2024, 1–7.

[Będ24] BĘDKOWSKI, J. “Open Source, Open Hardware Hand-Held Mobile Mapping System for Large Scale Surveys”. SoftwareX, 2024.

[BMB*24] BASSIER, M., MAZZACCA, G., BATTISTI, R., et al. “Combining Image and Point Cloud Segmentation to improve Heritage Understanding”. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-2/W4-2024, 2024, 49–56.

[Bra00] BRADSKI, G. “The OpenCV Library”. Dr. Dobb's Journal of Software Tools, 2000.

[Bre01] BREIMAN, L.: “Random Forests”. Machine Learning. Vol. 45(1). 2001, 5-32.

[CGRL20] CERVANTES, J., GARCIA-LAMONT, F., RODRÍGUEZ-MAZAHUA, L., LOPEZ, A. “A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends”. Neurocomputing, 2020, pp. 189-215.

[Clo24] CloudCompare (version 2.13.1.) [GPL software], retrieved from www.cloudcompare.org/, 2024.

[CPK*17] CHEN, L.C., PAPANDREOU, G., KOKKINOS, I., et al. “Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”. IEEE TPAMI, 2017, 40, 834–848.

[DTF*21] DI STEFANO, F., TORRESANI, A., FARELLA, E.M., et al. “3D Surveying of Underground Built Heritage: Opportunities and Challenges of Mobile Technologies”. Sustainability, 2021.

[EAQ22] ELHASHASH, M., ALBANWAN, H. and QIN, R. “A Review of Mobile Mapping Systems: From Sensors to Applications”. Sensors, 2022.

[EBR*08] EL-HAKIM, S., BERALDIN, J., REMONDINO, F., et al. “Using Terrestrial Laser Scanning and Digital Images for 3D Modelling of the Erechtheion, Acropolis of Athens”. DMACH. 2008, pp. 3–16.

[EGV*07] EL-HAKIM, S., GONZO, L., VOLTOLINI, F., et al. “Detailed 3D Modelling of Castles. International Journal of Architectural Computing”. 2007, 199–220.

[EMT*25] EL-ALAILYI, A., MORELLI L., TRYBAŁA P., et al. “Optimizing Multi-Camera Mobile Mapping Systems with Pose Graph and Feature-Based Approaches”. Remote sensing (In review), 2025.

[EPFR24] ELALAILYI, A., PERFETTI, L., FASSI, F. and REMONDINO, F. “V-SLAM-Aided Photogrammetry to Process Fisheye Multi-Camera Systems Sequences”. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-2/W4-2024, 2024, 189–195.

[ETM*24] ELALAILYI, A., TRYBAŁA, P., MORELLI, L., et al. “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, 2024, 147–154.

[FAF11] FASSI, F., ACHILLE, C. and FREGONESE, L. “Surveying and Modelling the Main Spire of Milan Cathedral Using Multiple Data Sources”. The Photogrammetric Record, 2011, 462–487.

[FKP*20] FIORUCCI, M., KHOROSHILTSEVA, M., PONTIL, M., et al. “Machine Learning for Cultural Heritage: A Survey”. Pattern Recognition Letters, 2020, 133, pp. 102–108.

[GDPR18] GRILLI, E., DININNO, D., PETRUCCI, G., and REMONDINO, F. “From 2D to 3D Supervised Segmentation and Classification for Cultural Heritage Applications”. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLII-2, 2018, 399–406.

[Gex25] Gexcel Srl. HERON MS TWIN Color – Portable 3D Mapping System, retrieved from heron.gexcel.it. 2025.

[GLM*24] GALANAKIS, D., LUCHO, S., MARAVELAKIS, E., et al. “Segment Anything Model for Scan-to-Structural Analysis in Cultural Heritage”. IEEE, 2024, pp. 1–7.

[GR19] GRILLI, E. and REMONDINO, F. “Classification of 3D Digital Heritage”. Remote Sensing, 2019.

[GR20] GRILLI, E. and REMONDINO, F. “Machine Learning Generalisation across Different 3D Architectural Heritage”. ISPRS. International Journal of Geo-Information , 2020, 9, no. 6: 379.

[JCQ23] JOCHER, G., CHAURASIA, A. and QIU, J. “YOLO by Ultralytics” (Version 8.0. 0). 2023, pp. 09-24.

[KB06] KANNALA, J. and BRANDT, S.S. “A Generic Camera Model and Calibration Method for Conventional, Wide-Angle, and Fish-Eye Lenses”. IEEE Trans. Pattern Anal. Mach. Intell, 2006, 1335–1340.

[KBKB17] KAN, T., BUYUKSALIH, G., KAYA, Y. and BASKARACA, A.P. “The Importance of Digital Methods in Preservation of Cultural Heritage: the Example of Zirnikli Mansion”. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-5/W1, 2017, 175–186.

[KMR*23] KIRILLOV, A., MINTUN, E., RAVI, N., et al. “Segment Anything”. arXiv:2304.02643, 2023.

[LBS*18] LI, Y., BU, R., SUN, M., et al. “PointCNN: Convolution On X-Transformed Points”. arXiv:1801.07791, 2018.

[LLWL23] LIU, H., LI, C., WU, Q. and LEE, Y.J. “Visual Instruction Tuning”. arXiv:2304.08485, 2023.

[LSD15] LONG, J., SHELHAMER, E., and DARRELL, T. “Fully Convolutional Networks for Semantic Segmentation”. In Proceedings CVPR, 2015.

[LS17] LANDRIEU, L. and SIMONOVSKY, M. “Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs”. arXiv:1711.09869, 2017.

[LZR*23] LIU, S., ZENG, Z., REN, T., et al. “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection”. arXiv:2303.05499, 2023.

[MFP*14] MARKUŠ, N., FRLJAK, M., PANDŽIĆ, I., et al. “Object Detection with Pixel Intensity Comparisons Organized in Decision Trees”. arXiv:1305.4537, 2014.

[NMR*17] NOCERINO, E., MENNA, F., REMONDINO, F., et al. “Investigation of Indoor and Outdoor Performance of Two Portable Mobile Mapping Systems”. SPIE Optical Metrology, 2017, p. 103320I.

[OS19] ORTIZ-CODER, P. and SÁNCHEZ-RÍOS, A. “A Self-Assembly Portable Mobile Mapping System for Archeological

Reconstruction Based on VSLAM-Photogrammetric Algorithm". Sensors, 2019.

[PBMR23] PADKAN, N., BATTISTI, R., MENNA, F. and REMONDINO, F. "Deep Learning to Support 3d Mapping Capabilities of a Portable VSLAM-Based System". Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-1/W1-2023, 363–370.

[PFV24] PERFETTI, L., FASSI, F. and VASSENA, G. "Ant3D—A Fisheye Multi-Camera System to Survey Narrow Spaces". Sensors, 2024.

[PPM*20] PIERDICCA, R., PAOLANTI, M., MATRONE, F., et al. "Point Cloud Semantic Segmentation Using a Deep Learning Framework for Cultural Heritage". Remote Sensing, 2020, 12(6), pp. 1005.

[PSW*21] PATHAK, R., SAINI, A., WADHWA, A., et al. "An Object Detection Approach for Detecting Damages in Heritage Sites Using 3-D Point Clouds and 2-D Visual Data". J. Cult. Herit., 2021,48, pp. 74–82.

[QSMG16] QI, C.R., SU, H., MO, K. and GUIBAS, L.J. "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation". arXiv:1612.00593, 2016.

[QYSG17] QI, C.R., YI, L., SU, H. and GUIBAS, L.J. "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space". arXiv:1706.02413, 2017.

[Rem11] REMONDINO, F. "Heritage Recording and 3D Modeling with Photogrammetry and 3D Scanning". Remote Sensing, 2011, 1104–1138.

[RGD23] RÉBY, K., GUILHELM, A. and DE LUCA, L. "Semantic Segmentation Using Foundation Models for Cultural Heritage: an Experimental Study on Notre-Dame de Paris". IEEE/CVF (ICCVW), 2023, pp. 1681–1689.

[RLZ*24] REN, T., LIU, S., ZENG, A., et al. "Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks". arXiv:2401.14159, 2024.

[RR10] REMONDINO, F. and RIZZI, A. "Reality-Based 3D Documentation of Natural and Cultural Heritage Sites—Techniques, Problems, and Examples". Appl Geomat, 2010, 85–100.

[RRL23] ROBERT, D., RAGUET, H., and LANDRIEU, L. "Efficient 3D Semantic Segmentation with Superpoint Transformer". Proc. IEEE/CVF ICCV, 2023, 1785–1795.

[TABF24] TRECCANI, D., ADAMI, A., BRUNELLI, V. and FREGONESE, L. "Mobile Mapping System for Historic Built Heritage and GIS Integration: A Challenging Case Study". Appl Geomat, 2024, 293–312.

[TC18] TEICHMANN, M., CIPOLLA, R. "Convolutional CRFs for Semantic Segmentation". arXiv:1805.04777, 2018.

[TGR*20] TERUGGI, S., GRILLI, E., RUSSO, M., et al. "A Hierarchical Machine Learning Approach for Multi-Level and Multi-Resolution 3D Point Cloud Classification". Remote Sensing, 2020.

[TMBR21] TORRESANI, A., MENNA, F., BATTISTI, R. and REMONDINO, F. "A V-SLAM Guided and Portable System for Photogrammetric Applications". Remote Sensing, 2021.

[TQD*19] THOMAS, H., QI, C.R., DESCHAUD, J.-E., et al. "KPConv: Flexible and Deformable Convolution for Point Clouds". arXiv:1904.08889, 2019.

[VRP*23] VIEIRA, M.M., RIBEIRO, G., PAULO, R., et al. "Strategy for HBIM Implementation Using High-Resolution 3D Architectural Documentation Based on Laser Scanning and Photogrammetry of the José De Alencar Theatre". Digital Applications in Archaeology and Cultural Heritage, 2023.

[VRR*23] VILEIKIS, O., RIGAUTS, T., ROUHANI, B., et al. "Dive into Heritage: A Digital Documentation Platform of World Heritage Properties in the Arab States Region". Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-M-2-2023, 1613–1620.

[WJHM15] WEINMANN, M., JUTZI, B., HINZ, S., MALLET, C. "Semantic Point Cloud Interpretation Based on Optimal Neighborhoods, Relevant Features and Efficient ClassifiersISPRS Journal of Photogrammetry and Remote Sensing, 2015, 286-304.

[WQF18] WU, W., QI, Z. and FUXIN, L. "PointConv: Deep Convolutional Networks on 3D Point Clouds". arXiv:1811.07246, 2018.

[XZY*22] XU, X., ZHANG, L., YANG, J., et al. "A Review of Multi-Sensor Fusion SLAM Systems Based on 3D LIDAR". Remote Sensing, 2022.

[YHL23] YANG, S., HOU, M. and LI, S. "Three-Dimensional Point Cloud Semantic Segmentation for Cultural Heritage: A Comprehensive Review". Remote Sensing, 2023.

[YLZ*25] YUAN, H., LI, X., ZHANG, T., et al. "Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos". arXiv:2501.04001, 2025.

[ZEPF24] ZHANG, K., ELALAILYI, A., PERFETTI, L., and FASSI, F. "Cost-Effective Annotation of Fisheye Images for Object Detection". Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-2/W8-2024, 2024, 491–498.

[ZF25] ZHANG, K. and FASSI, F. "Transforming Architectural Digitisation: Advancements in AI-Driven 3D Reality-Based Modelling". Heritage, 2025.

[ZJJ*20] ZHAO, H., JIANG, L., JIA, J., et al. "Point Transformer". arXiv:2012.09164, 2020.

[ZPK18] ZHOU, Q.-Y., PARK, J. and KOLTUN, V. "Open3D: A Modern Library for 3D Data Processing". arXiv:1801.09847 , 2018.

[ZSQ*17] ZHAO, H. SHI, J., QI, X., et al. "Pyramid Scene Parsing Network". In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6230–6239.

[ZYWL23] ZHANG, Z., YANG, B., WANG, B. and LI, B. "GrowSP: Unsupervised Semantic Segmentation of 3D Point Clouds". arXiv:2305.16404, 2023.

[ZZCL19] ZHANG, J., ZHAO, X., CHEN, Z. and LU, Z . "A Review of Deep Learning-Based Semantic Segmentation for Point Cloud". IEEE Access, 2019, 179118–179133.