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

Milan Cathedral Digitized: A Stone-by-Stone Segmentation Approach Using SAM2

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

[1] 3D Survey Group, ABC Department, Politecnico di Milano, Via Ponzio 31, 20133 Milano, Italy – (kai.zhang, chiara.mea, ahmad.elalailyi, fausta.fiorillo, francesco.fassi)@polimi.it
[2] 3D Optical Metrology (3DOM) Unit, Bruno Kessler Foundation (FBK), Trento, Italy – aelalailyi@fbk.eu
[3] Department of Civil, Architectural, Environmental Engineering and Mathematics (DICATAM), Università degli Studi di Brescia, 25123 Brescia, Italy – luca.perfetti@unibs.it

Keywords: Segmentation, Artificial Intelligence, SAM2, Stone Blocks, Cultural Heritage, Maintenance.

Abstract

High-resolution architectural documentation goes beyond geometry—it requires a deep understanding of the building's structure, materials, and historical layers. This often means interpreting hidden construction logic and identifying even the smallest components, such as individual stones or bricks, to produce meaningful data for conservation, analysis, and interpretation. Identifying and describing all the individual components that constitute the building, such as the type, arrangement, and state of preservation of stones, bricks, mortars, or decorative materials embedded in the walls, is a real challenge due to the large quantity and the complex spatial distribution of each element. Recent advances in AI, particularly foundational models and zero-shot models, offer potential solutions to speed up the documentation process. Taking the gothic complex of Milan Cathedral as the monument object of study, the research hereby presented implements a SAM2 (Segment Anything Model) based stone-by-stone segmentation, leveraging object detector for semantic interpretation. The proposed framework integrates 2D stone block segmentation with photogrammetric 3D reconstruction, enabling accurate projection of semantic labels and geometric data from images to 3D point cloud, allowing a detailed 3D segmentation in all the components of the structure.

1. Introduction

Stone block segmentation has been a keen focus in recent years in the cultural heritage (CH) documentation field. The documentation of architectural assets is among the most enduring and complex activities in field of CH conservation. The manual recognition, mapping, and monitoring of the architecture and its individual components over time, regularly at certain intervals, is an invaluable activity and continues to be widely practiced. However, particularly when it comes to large-scale complex CH assets, the documentation activities and recurrent updates become costly, time-consuming, and burdensome processes.

Milan Cathedral stands as a landmark of Milan city and one of the most important monumental heritages in Italy as well as Europe. As one of the most significant Gothic cathedrals, it features marble cladding. The Candoglia marble, vulnerable to both natural and anthropogenic deterioration agents, requires continuous preservation. The focus of this research starts from this premise, conveying efforts to apply and validate Artificial Intelligence-assisted segmentation of stone blocks on this large case study. In CH assets, mainly in Europe, stone is commonly chosen for its durability and strength. However, constant exposure to weather and unpredictable external factors cause erosion, cracks, biological colonization, black crust, and other decay pathologies (Verges-Belmin, 2008). Consequently, the periodic maintenance and conservation activities become crucial. Knowledge of the restoration history, physical characteristics, and the up-to-date conditions of the building elements benefits context-specific conservation practices (Coletti et al., 2025). Stone block mapping and documentation have become necessary in preservation practices. The recurrent conservation activity of the Milan Cathedral has been lasting over the years, the stone block maintenance is crucial to preserve the Cathedral's aesthetic and structural value, and its replacement or modification is the key feature in the monument conservation activity. However, documentation at the block level - considering the quantity and characteristics of the instances - is time-consuming and a serious logistical and archival challenge.

The recent Artificial Intelligence (AI) boom has provided promising tools and solutions to streamline the documentation process. The research in the field of computer vision has been productive in processing images, accomplishing tasks like image classification (categorizing an image based on its content), object detection (identifying and localizing objects in the scene), and instance segmentation (recognizing every individual object and its borders). AI has already been applied to structural integrity analysis, precise damage assessment, and restoration of CH assets (Llamas et al., 2017). Not only have the visual models been improved over the years, but the emerging AI models also allow open-set detection and are ready-to-use, reducing dependency on labelled data and model training processes. Among all, Segment Anything Model (SAM) (Kirillov et al., 2023), a pre-trained foundational model, has stimulated the development of segmentation-involving pipelines. These AI models enable precise, cost-effective architectural documentation making 2D and 3D data and related documentation available and readable.

As a continuous study of the survey and digitalization activities of Milan Cathedral have been ongoing for many years (Achille et al., 2020, 2012; Spettu et al., 2021), this paper seeks to explore the potential of AI tools, by applying the SAM2 to Milan Cathedral stone block segmentation, integrated with object detection models (comparing the applicability of open-set detector Grounding DINO (Liu et al., 2024) with the close-set detector YOLOv11 (Jocher and Qiu, 2024)) and photogrammetry. The goal is to enable automatic stone segmentation in 2D imagery and implement 3D mapping, contributing to the documentation practices. The concluding phase will involve the visualization of the segmentation outcomes on the 3D point cloud model, showing the ultimate

stone-by-stone segmentation in three dimensions. The research team's aim is towards having the possibility to exploit 3D technologies for various applications, making these advanced tool outputs usable also to the end-user (e.g., Achille et al. 2020, Spettu et al. 2021).

1.1 Related Works

The initial research on the segmentation task employed rule-based approaches, like thresholding, edge detection (Canny, 1983; Sithole, 2008), and region-growing methods (Beucher and Meyer, 1992): no learning from a proper training and dataset, but mainly image manipulations. Later, machine learning enabled segmentation based on feature extraction and pattern recognition (Mohammed and Mihoub, 2024).

In recent decades, deep learning methods have used convolutional networks for end-to-end segmentation. Deep learning models such as FCN (Fully Convolutional Networks) (Long et al., 2015) and U-Net (Ronneberger et al., 2015) initiated the Convolutional Neural Network (CNN)-based segmentation approach. Later Mask-R-CNN (He et al., 2018) added a segmentation head to an object-detection model, allowing instance segmentation. Recently, the transformer-based models entered into computer vision field, e.g., DETR (Carion et al., 2020). A breakthrough solution in segmentation tasks is the Segment Anything Model (SAM) (Kirillov et al., 2023). It has addressed the general image segmentation problem by simplifying the task and becoming a popular off-the-shelf utility. It is, in fact, able to cross various domains without the need for suited training. Developed by Meta AI towards a general-purpose foundational model, SAM was trained on SA-1B (Segment Anything 1-Billion mask dataset), an extremely large segmentation dataset which includes 1 billion masks on over 11 million images. It's a "promptable" (capable of being controlled or customized via prompts, i.e., inputs) model that supports inputs like points, boxes, and masks to guide accurate mask generation for any object in a sequence of images, without extensive training, i.e., zero-shot segmentation. Its capacity to be applied to new and unseen data without prior training is its primary advantage. To be noticed is that SAM does not deal with semantics. Thus, it doesn't give back any label with the segments, requiring additional extensive approaches (classifier or vision-language model) to link each segment to the corresponding semantic meaning.

Many architectural surfaces feature repeating patterns makes them ideal for AI segmentation and single component identification. In fact, they can exploit this feature of built assets to learn consistently, improving performance over different computer vision tasks. An early study of block segmentation compared the performances of multiple segmentation methods upon orthophoto of stone façades (Idjaton et al., 2021), showing the limitations of rule-based techniques. Other research proposed the use of SAM, coupled with a traditional classifier such as Support Vector Machine (SVM) and control of morphological opening to achieve better segmentation results (Lucho et al., 2024). Their results suggest that marrying the segmentation capabilities of SAM with some post-processing and label assigning can improve alignment with architectural features.
Later applications integrate a cross-modality model Grounding DINO (Liu et al., 2024) for primal object detection, guiding the segmentation tasks of SAM, and uses CLIP, a Contrastive Language-Image Pre-training model (Radford et al., 2021) to associate labels to segments (Réby et al., 2023). Their work highlights the potential of combining the large Vision Transformer (ViT) models with semantics. Inspired by previous research, this paper employs SAM2 on the scale of architectural facade, not mapping major architectural elements, but individual small-scale architectural elements (i.e., marble blocks) on 2D imagery. As an additional step, it explores the feasibility of 2D to 3D projection.

Building on this recent line of research, the aim is to conduct segmentation of individual marble blocks for the practical use of Milan Cathedral documentation. As a direct consequence of their use, the masks generated by SAM should be correctly identified, interpreted, and categorized. So, this research applies object detection models - YOLOv11 (Jocher and Qiu, 2024; Redmon et al., 2016) and Grounding DINO - for acquiring semantics information from the 2D images and classifying segments. The object detection models localize the object regions of interest on images, and indicate categories for the instances detected, enabling filtering segmentation results by semantics. Grounding DINO is a cross-modality object detector model that combines language and visual information. It can be promptable with text inputs (natural language), enabling it to locate and label the expected object. Grounding DINO combines a Transformer-based detector (DINO) for modality fusion with grounded pre-training for concept generalization. The tight fusion includes a feature enhancer, a language-guided query selection, and a cross-modality decoder for language-vision modality fusion. The Grounding DINO performs well in open-set object detection on benchmarks like COCO. YOLOv11, on the other hand, is a close-set deep-learning application for object detection, known for its state-of-the-art velocity and efficiency.

The intention is to achieve segmentation at a level of complete architecture of every element composing it, regardless of the complexity of the single case. However, this goal can present several challenges that need to be provided a solution to. While foundational models like SAM offer generalization capabilities, domain-specific fine-tuning may be used in the future to handle the unique shapes and characteristics of Milan Cathedral (spires, very ornated areas, big gothic windows with coloured glass, statues). These elements are very difficult to recognize in other architectural contexts and are therefore challenging to label correctly by a zero-shot model.

2. Stone block detection in the Milan Cathedral

2.1 Methodology

This research presents a practical application exploiting novel AI processing, including object detection and segmentation, aimed at streamlining the segmentation of marble elements within the comprehensive management framework of the cathedral. The investigation is based upon 3D data gathered within the Milan Cathedral survey project that, in the span of over 15 years, achieved a full high-resolution point cloud representation of the full building. Photogrammetry can generate 3D models and orthophotos at a very high resolution (up to the detailed visualization of the mortar gaps), from which semantic and geometric information can be extracted quickly and consistently by AI methods. The high quality required for the photographic acquisition, typical of the architectonic photogrammetric survey, perfectly suits the AI automatic recognition.

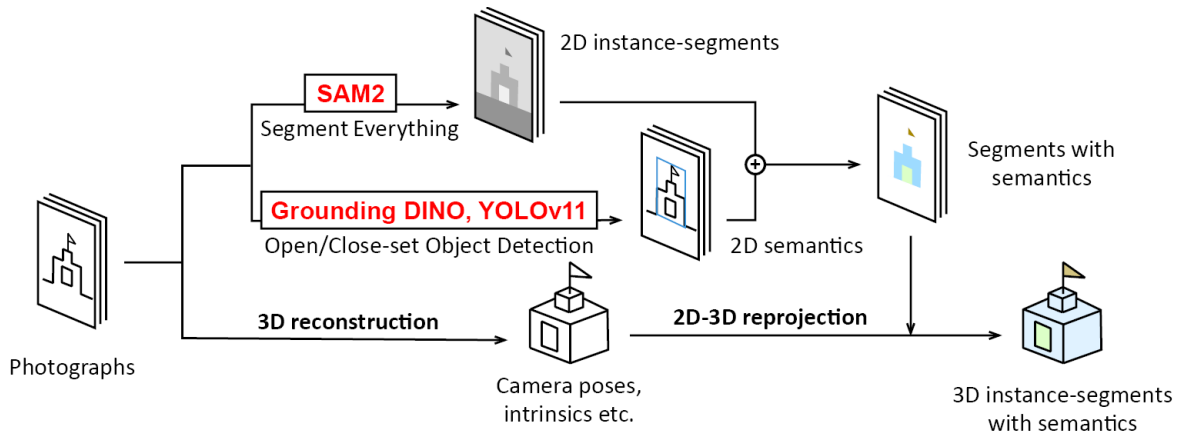

Figure 1. The proposed pipeline of this research: iteration of the segment and classification process until the practical details achieved.

The survey data available offers some key advantages for the implementation of the aforementioned AI framework: (1) the availability of large volumes of data, with the repetitive feature of contents and the images used in the photogrammetric process that are ideal for stone-by-stone segmentation and the detection of marble architectural elements; (2) the known exterior orientation of the images (position and rotation) of each frame in the 3D space, apart from the intrinsic calibration of the cameras derived from the photogrammetric processing. All these factors make it easier to trace back the position of recognized elements from the 2D images to the 3D point cloud geometry.

Aware of this, the objective of this research is to carry out both the segmentation and labelling tasks based on the ready-available and already-oriented image dataset acquired during the photogrammetric survey. The concluding phase will involve the visualization of the segmentation outcomes on the 3D model, showing the ultimate stone-by-stone segmentation in three dimensions.

The presented pipeline involves mainly the following procedures (**Figure 1**): initial segmentation using SAM2, semantic interpretation through object detection, and 3D projection. In detail, the initial step involves using SAM2 to create segmentation masks for all the images used in the 3D reconstruction. This results in a binary mask, which comprises all the segments without semantic labelling. In the second stage, object detection models (Grounding DINO, YOLOv11) are applied for semantic interpretation, comparing their applicability to the specific need of object recognition. The results of the object detection were used as box prompts to guide SAM2 for generating precise masks of certain interference objects that were not marble blocks. These masks can be used to crop off specific chosen element categories from the initial SAM2 segmentation result. Finally, the individualized masks segments of marble block units are projected from 2D images to a 3D model within the tested area of the case study, maintaining the spatial correspondence of the blocks from 2D to 3D space utilizing photogrammetric parameters. The output will be saved to each point of the point cloud as a scalar field, which indicates different block instances, facilitating the later data fruition.

2.2 The Milan Cathedral

Milan Cathedral stands out as a late Gothic masterpiece. Its construction began in 1386 and finished in 1805. It is the largest church in Italy: the external dimension of the Cathedral is 158*93 meters, with a total height of 108.50 meters, covers an area of approximately 12,000 m^2, and has a gross volume of 440,000 m^3. The church features highly rich decorations, counting thousands of marble sculpted statues and decorations.

The Cathedral is characterized by the Candoglia marble coverings. Over time, marble blocks naturally deteriorate as a result of their mineralogical properties and continuous exposure to outdoor environmental factors. These blocks, especially the external ones, are replaced periodically under the recurrent investigation and continuous maintenance. The institution Veneranda Fabbrica del Duomo di Milano has been in charge of this maintenance work for more than 630 years. It's believed that the block-by-block digitalization job is beneficiary to the continuous and accumulating preservation practices and documentation (Fassi et al., 2015).

The study hereby presented was conducted in continuity with the photogrammetric survey, aimed at extracting 1:50 2D drawings, as requested by the Veneranda Fabbrica for the documentation of preservation works. The external facades were surveyed using photogrammetry, because of their extreme flexibility. The survey used the aid of a lifting platform and scaffolding facilities to achieve acquisition positions at altitude, ensuring uniform Ground Sample Distance (GSD) and complete geometry.

The image dataset involved in this research was extracted from the south façade survey, choosing an area equal to the full transept façade on the Southern side of the Cathedral (see **Error! Reference source not found.**). This area is characterized by varied volumes, featuring large, elaborately decorated windows and slit windows that illuminate the staircase volume. In addition to ornamental patterns, the façade includes statues of saints, canopies, "falconature" (the crowning decorative elements), and spires crowning the structure. This specific area, approximately 49 m wide and 64 m high (height calculated from ground level to the tallest spire), was identified to test and illustrate the proposed pipeline. The testing data involves 842 photos, taken with the intention of achieving a GSD of 2mm/pixel. The images were aligned and processed through a photogrammetric workflow in Agisoft Metashape. The generated dense point cloud is composed of 160,282,474 points, representing the surface geometry of the south façade in detail. The data was used to prepare the orthophoto of the whole façade (Perfetti et al., 2019).

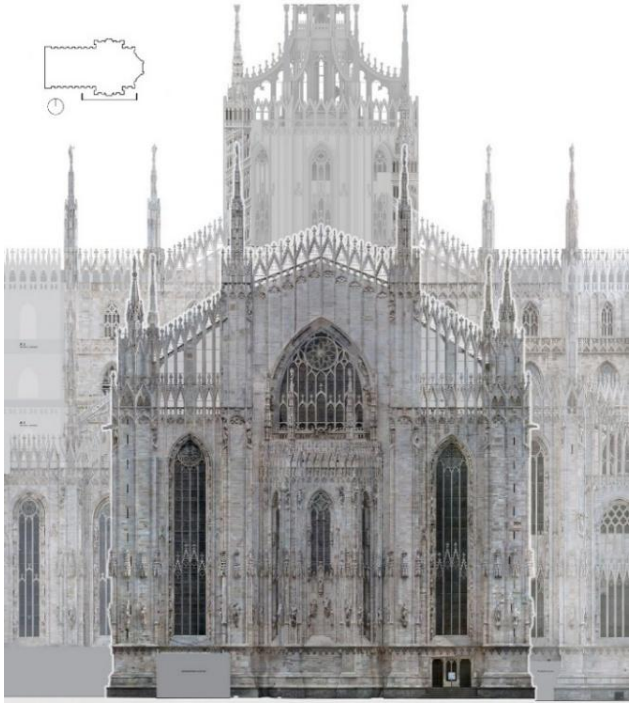

Figure 2. Orthophoto of the South elevation of the transept .

2.3 “Segment Everything” using SAM2

The experiment began with the implementation of SAM2, developed by META, selected for its competitive performance in zero-shot segmentation tasks across diverse domains. Unlike traditional Convolutional Neural Networks (CNNs), SAM2 does not require prior training for specific datasets, making it suitable for applications demanding generalization. Milan Cathedral’s surface, predominantly composed of marble blocks arranged in repetitive yet subtly varied patterns, was considered an appropriate case for automated segmentation and element identification.

However, challenges emerged when applying this pre-trained foundational model to such a complex architectural surface. The visual uniformity of the marble reduces contrast and affects mask precision, considering the marble elements characterized by minimal colour variation, limited depth differences, and mortar gaps closely matching the stone's hue. These characteristics hinder the model’s ability to clearly delineate individual elements. The Cathedral’s stone skin is intentionally designed to appear as a smooth, continuous surface, which offers few visual cues to support the segmentation process. As an expectation, SAM2’s performance will be constrained when dealing with context-specific visual nuances that require more than generalized pattern recognition.

To address the issue of recognising stone blocks when they’re uniform in colour and with shallow gaps between them, some brightness and contrast adjustments in the original photos were necessary. These changes must be applied to keep the image well readable and are particularly useful when images have different exposures or poor distinction between similar materials, such as marble and mortar. So, instead of applying fixed values (that could be applicable in some pictures and not in others, as it’s well-known the long-term survey of such monuments goes on in time and so with varying exposures)– Contrast Limited Adaptive Histogram Equalization (CLAHE) (Zuiderveld, 1994) was used to enhance contrast and aid the segmentation model in reading the borders of singular objects while preserving details.

Prior to being processed with SAM2, the dataset undergoes a pre-processing phase aimed at improving the performance of the subsequent analysis. Firstly, the images were compressed with maximum side dimension to 2048 pixels, while maintaining the block gaps recognizable. In addition, the images were converted to grayscale, reducing the possible confusion caused by the hue, and then went through CLAHE.

The function SAM2AutomaticMaskGenerator (AMG) is embedded with the capability to sample a single-point input prompt in a grid over the image, which guides the SAM2 model to generate multiple masks on the whole image. In this function, no other prompt input was used to infer the masks (no point coordinates or bounding boxes). Multiple parameters were set to balance computational efficiency and segmentation capabilities of SAM2. Most importantly, “points_per_side” parameter is set to 128. It indicates the sampling of “prompt points” on the image at a grid of 128x128, determining 16384 points across the image used to prompt the model. This parameter largely affects segmentation granularity, i.e., the scale of the detecting elements. At each time of the inference, 128 prompt points were fed to the model (points_per_batch). Additionally, a minimum mask area (25 pixels) was set to filter out mispredictions, ensuring that detected elements align with the typical size of stone blocks. Further filtering methods were applied: an Intersection-over-Union (IoU) threshold of 0.7 to retain only masks with high spatial overlap; a Non-Maximum Suppression (NMS) threshold of 0.5 to eliminate redundant overlapping masks; and Mask-to-Mask (M2M) refinement to enhance mask boundary precision.
Eventually, after the pre-processing, SAM2 assigns values to pixels, grouping them based on visual features and prompt input. The model output (**Figure 3**) consists of multiple binary masks, covering all the instances detected. For visualization purposes, the masks inferred from the images were colorized using a set (200) of randomly generated colours. This step supports downstream processing by enabling the clear identification and separation of individual elements within the images.

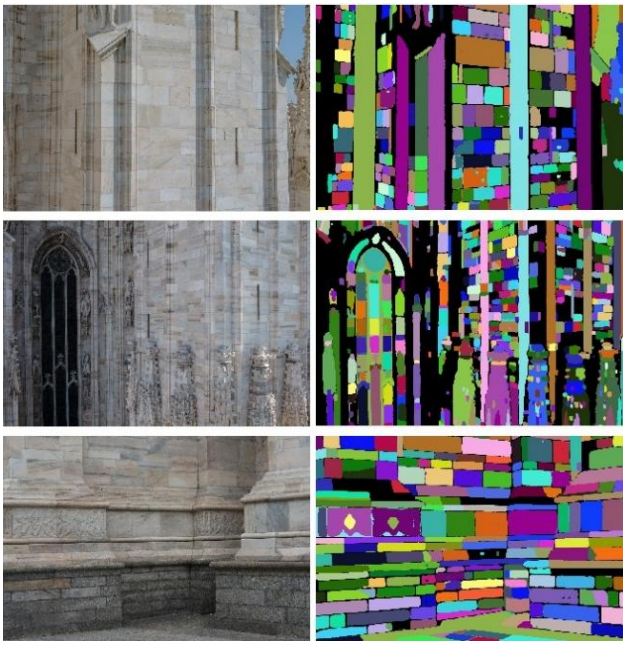

Figure 3. Examples of raw photographs (left) and the initial SAM2 outcomes (right) on the Transept area on the South façade of the Milan Cathedral.

2.4 Detection and Segmentation

After mask creation, the segments generated by SAM2 must be differentiated based on categories, classifying all the instances captured in the imagery for practical reasons. For this purpose, detection model is used to detect objects which cannot be categorized as marble block, such as sculptures, decoration, windows, slit windows, spires, scaffolding and other interferences. In this premise, an object detector can be applied to interpret the semantics. This study tested Grounding DINO and YOLOv11 models, comparing their capability on detecting CH related elements.

2.4.1 Grounding DINO: The pre-trained model Grounding DINO offers satisfying open-set object detection capabilities. It produces bounding boxes of the targeted objects based on textual prompts without the need for extra training.

In this study, Grounding DINO was first applied to interpret the semantics, for its off-the-shelf convenience. Specifically, a checkpoint GroundingDINO_SwinT_OGC was used, with Swin-Tiny as backbone, pretrained on compressed dataset OGC comprises of Objects365, GoldG, and Cap4M datasets. It was applied to detect architectural elements, facilities, and humans, other than marble blocks. Resulting bounding boxes were used to guide the creation of masks in SAM2 (**Figure 4**). This process is parallel to the masks creation discussed above; the output is being used to subtract unwanted regions from the primal segmentation. For the settings of parameters, the box threshold is set to 0.15, and the text threshold to 0.12, to filter the bounding boxes based on the predicted grounding score (confidence that the box matches the text) and the text relevance score (confidence that the text query matches the region). NMS threshold is set to 0.4. The detections are later filtered by the size of bounding boxes (neither too big nor small), removing overlapping boxes. The main filtering criterion was related to the dimension of the box in respect to the full size of the image. Given that targeted elements cover an area of less than 20% of the whole image, any bounding box exceeding that value was filtered out right before the inference on the image.

Text prompts are tested for detecting sculptures, spires, windows, slit windows, falconature, decoration, scaffolding and humans. However, 'falconature' and 'decoration' cannot lead to satisfying detection, even though tested multiple other prompts like: 'gothic hood moulding', 'arch moulding', 'carved detail', and 'ornament' etc.

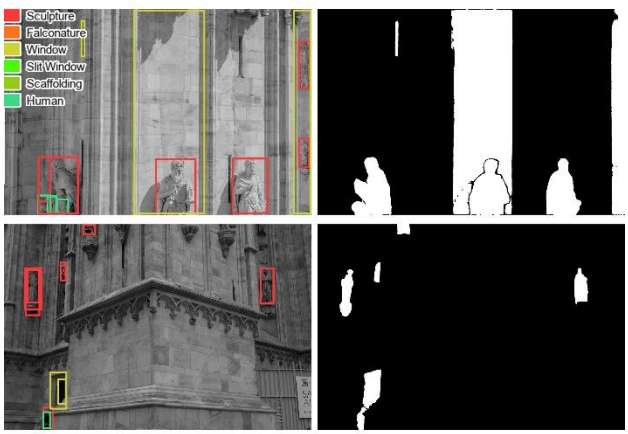

Figure 4. Examples of the detection from Grounding DINO (left) and the SAM masks generated with corresponding box prompts (right).

As the examples in **Figure 4** show, Grounding DINO has achieved promising detection results. However, the results are not prominent when dealing with various ornaments and decorations, which happen to be the most dominant feature of architectural cultural heritage. The ready-to-use open-world detection models are not trained particularly for the heritage context; hence each category will require articulated and concise text-prompts to guide the prediction. The difficulties in the process and low precision output can be expected.

2.4.2 YOLOv11: The outcome from Grounding DINO is promising but has left heavy manual work to improve the results for practical downstream applications. To reduce the dependency on time-consuming manual operation, a YOLOv11 (You Only Look Once) model was tested to identify pre-defined categories, which require manual annotation and model training.
YOLOv11 is a late iteration in the YOLO series, a big march towards real-time detection performance. The model is equipped with sophisticated feature extraction techniques, apart from refined training methodologies. Noticeably, the Cross Stage Partial with Spatial Attention (C2PSA) module was introduced, enhancing the spatial attention in the feature map.

In this application, a set of 143 samples was subsampled from the whole dataset to form a representative training set. The samples were annotated, with the purpose to define 8 categories: sculptures, spires, windows, slit windows, falconature, decoration, scaffolding and humans. In the end, 1895 instances were annotated, with the annotated area taking up to 44.59% of the whole image. The category distribution was 752, 418, 355, 15, 191, 81, 59, 24 instances respectively for used categories mentioned before. The model was trained based on YOLOv11 m variant, upon nine-tenths of the annotated samples, and reached convergence at around 80 epochs. Upon the unseen one-tenth of the data, the model achieved 81.01% precision, 0.69 for the mean Average Precision (mAP) at the Intersection over Union of 0.5 (mAP_0.5), and 0.40 mAP at different IoU thresholds from 0.5 to 0.95 in steps of 0.05 (mAP_0.5:0.95).

Ultimately, this approach resulted in much better performance as expected, since the training samples are collected from the overall dataset. The confusion mainly addresses falconature and scaffolding. In the end, the YOLOv11 detection results were used to guide the creation of SAM2 segments that were used to subtract unwanted areas from the primal segmentation mask previously created. This step contributes to mapping out the unwanted objects for precise marble block segmentation.

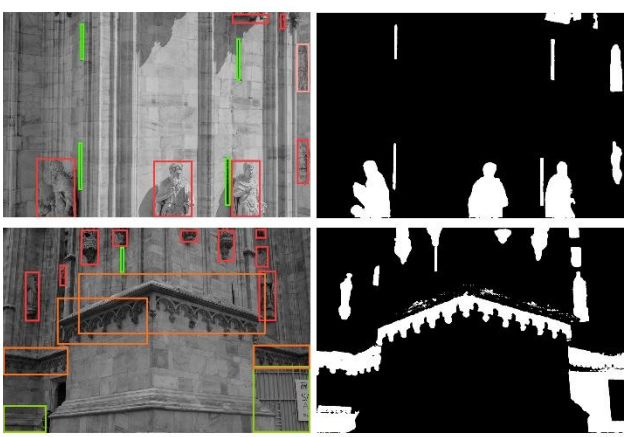

Figure 5. Examples of the detection from YOLOv11 (left) and the SAM masks generated with corresponding box prompts (right).

2.5 Projection to 3D

After mapping out non-targeted (ornated architectural elements) segments, the remaining 2D segments of marble blocks are intersected into three-dimensional space utilizing the known camera poses and intrinsic parameters derived by the photogrammetric project.

The used method is adapted from a previous work (Alami and Remondino, 2024; El-Ayaili et al., 2025), which combines voxel-based ray casting and camera model projections. The method takes segmentation output, dense point cloud, oriented camera parameters as input. During the process, the point cloud is voxelized at a user-defined resolution (in this application, the side length of each voxel cube is set to 0.2 meter, balancing spatial granularity and computational efficiency), and a ray casting scene is generated using Open3D (Zhou et al., 2018).

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 casted rays.

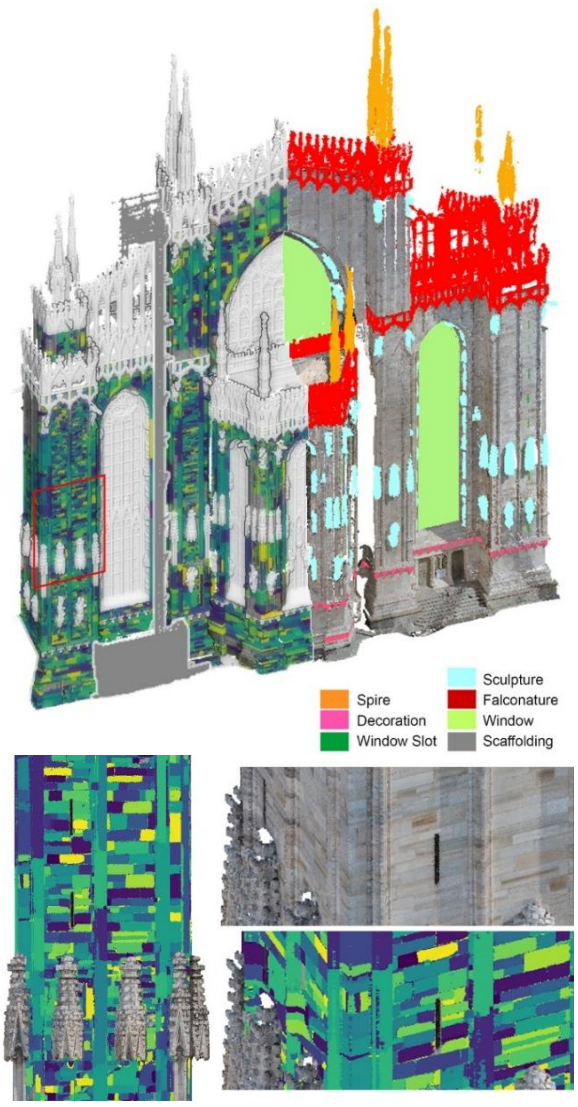

Figure 6. The reconstructed 3D model of the South façade of the transept, with each marble block rendered in different colours based on its corresponding segment label.

As can be seen in **Figure 6**, the segments of the block are interpolated back to 3D space. However, in some area the ray casting strategy has caused minor drifting of the segment projection. For such cases, manual improvement is required to repeat the process only on the problematic area, with eliminating input images that lead to projection from far away and with a wide angle.

3. Discussion

The test intended to involve AI model to accelerate and automatize the semantic segmentation upon 2D data. The 2D outcome has been re-integrated into a 3D reconstructed model of the Milan Cathedral Transept, south facade. The practical test has shown the feasibility but also reveals pros and cons.

In the case study of the Milan Cathedral, SAM2 demonstrates satisfactory performance of AMG function on individualize each marble component presented in the images. The images acquisition is aimed for photogrammetric workflows, that means that the images used are not subject to strong distortion and are captured almost nadiral to the object surface. In addition, the lighting condition is stabilized, harsh shadows and over exposure are avoided. As a result, SAM2 masked accurately the well-defined borders of marble stone elements. The model presents its satisfying performances on detecting small objects, including decayed areas like cracks, material loss. A critical aspect of SAM2 is the over-segmentation, resulted by the grid-sampling prompts and the lack of semantic awareness. It requires extra constraining of model behaviour and result filtering for practical use. Further fine-tuning with elaborate data and post-processing might address this defect. Another critical aspect is that the AMG function is extremely time-consuming. In the presented case study, an image of size 2048*1367 pixels takes 29 minutes to generate over 200+ instances masks, using processer Inter® Core™ i9-9980XE CPU @ 3.00GHz. The occupation of computational resources is noticeable. It this research, multiple devices are used to accelerate the process, about 4 days are used to acquire the initial segments.

One of the primary limitations of the "segment anything" function on SAM2 model is the requirement for mask post-processing, refining and semantics interpretation. To interpret the semantics of the segments, object detection models were implied. In this application, Grounding DINO effectively distinguishes facade elements, even at distances. It encountered similar difficulty related to the object scale in the image as SAM2 did: in non-ortho photos the smaller objects that are far away from the camera may result into a lower confidence score. The utility of Grounding DINO is under criticism for the prompt mechanism that encodes text into dense embeddings that are aligned with visual features. While Grounding DINO exhibits high accuracy if provided simple and direct prompts, like colour description (e.g. "grey"), shape (e.g. "square"), common object categories (e.g. "window", "person"), when it comes to cultural heritage scenarios the specialized terminology will cause crucial problems. In this research, for example, it fails with more domain-specific terms like "falconature", which are rare or absent in the training data. In addition, inaccuracies could rise while using long descriptions for the interested object, which can produce conflicts between levels of text feature (sentence and word) that might be misleading to the detection. In the heritage domain, for instance, the prompt "white stone element" could refer equally to cornice, archivolt, or statue fragment. A single dataset may contain multiple distinct classes, where shape and colour do not always correlate to a single semantic category, and objects within the same class may exhibit entirely diverse shapes and intricate

geometries. Defining precise categories is significantly more complex than general-purpose recognition works. There requires a time-consuming effort of testing to reach accurate grounding prompt, considering the vision-language contextual information. In this case study, Grounding DINO were evaluated on the manual annotated data that were fed to train YOLOv11 model (see **Table 1**). The model achieved higher evaluation score on more general concepts like sculpture, scaffolding, human. However, related to more specific categories like falconature and decoration, the behaviour is neither stable nor acceptable. Grounding DINO can generate multiple detection boxes indicating the same instance (for example . Therefore, using the detections to guide the SAM2 segmentation, the resulted segments will overlap on the same instances though on different scales, slightly supplementing the detection results. By comparing to the SAM2 results guided by manual annotation, the mean IoU reaches 0.309. In the end, consider the Grounding DINO detections require manual inspections before later process, the research moved on applying close-set detector YOLOv11. Yolov11, as one of the latest object detection models, exhibits its advantages in speed and accuracy. In this test, it has shown better performance, even with a limited dataset. Compared to Grounding DINO, close-set detection model has limitations on preparing tailored training dataset, taking into consideration of the time and effort on annotation and model training. These models have satisfying performances but only on fixed pre-defined class vocabulary with limited generalization ability.

	Precision	Recall	F1-score	Support
Sculpture	0.64	0.29	0.40	752
Spire	0.37	0.03	0.06	418
Falconature	0	0	0	355
Decoration	0	0	0	15
Window	0.26	0.41	0.32	191
Slit window	0.21	0.05	0.08	81
Scaffolding	0.44	0.31	0.36	59
Human	0.30	0.33	0.31	24
Macro Avg.	0.28	0.18	0.19	1895
Weighted Avg.	0.39	0.18	0.22	1895

Table 1. classification report of Grounding DINO on annotated ground truth at IoU threshold 0.5.

The accuracy of the projection from 2D-based detection to 3D model is affected by multiple factors: voxel dimension, mask quality, and precision of camera parameters. As steps in the long processing pipeline, photograph quantity, camera calibration, AMG outcome, semantic interpretation and voxel size can introduce mislabeling or outliers into the final 3D output. The last procedure yielded a satisfactory result. Fifty percent of stone surfaces exhibit effective segmentation, with marble blocks properly aligned and the gaps between the blocks clearly delineated. The mis projected component predominantly occurs in regions exhibiting significant geometric alterations, such as angles or decorative elements, which provide challenges for both AI and the photogrammetric method. The reprojection of intricate 3D regions necessitates supplementary geometric constraints to accurately identify the appropriate 3D points for labeling and to resolve ambiguities arising from many intersections of the projection ray with the object.

The large number of instances in such a complex and extensive dataset can pose a challenge for data utilization. Scaling the project to encompass the entire Milan Cathedral requires a robust and scalable strategy. A well-design database with well-defined categories embedded with spatial hierarchy will be essential to arrive to the complete segmentation of the 3D data asset.

4. Conclusion

This paper presented applications of SAM2 in cultural heritage practices of the Milan Cathedral, focusing on segmenting the marble block on the south façade of the transept. The pipeline involves SAM2AutomaticMaskGenerator function to generate masks from the 2D image data, and project them to 3D utilizing photogrammetry techniques. The application further tested the applicability of the open-set detection model Grounding DINO comparing with trained YOLOv11 on finding architectural elements. The application eventually achieved categorization of the marble blocks for the façade of the south transept.

The application has addressed typical time-consuming problems with SAM2 on un-prompted segmentation. It reacts much faster when provided with spatial guidance (points or bounding boxes). The zero-shot model Grounding DINO has a promising performance, but when is applied to specific architectural cultural heritage categories it exhibits limitations. Preparing dataset and training the object detection model (YOLOv11) with pre-defined categories in this test is the most effective solutions. In the large preservation practices with continuous inspection and monitoring needs, a tailored dataset is worth the effort and could support the feasibility of the AI-aided pipeline. Until now, further manual operation and down-stream data fusion research remain. In a broader view of cultural heritage preservation activities, the foundation models like SAM2 can be expected to be integrated for practical utility. Consider a common standard will be established for cultural heritage dataset, open-set detection model can be used for more than general and basic data processing.

Future work could focus on improving the computational efficiency and stabilize the performance of the segment and detection model, verifying the applicability upon the whole Milan Cathedral. The recognition can be targeted to other articulated and complex architectural categories, such as windows and statues, to create a more comprehensive segmented model, storing detailed information about each of its elements on different scales. This pipeline shall be transferable to other case studies, having inherently no overfitting to a specific domain or case study. It can also be accelerated integrating with automatic classification methods of 3D data.

Acknowledgements

Financial support from the program of the China Scholarships Council (grant number: 202208520007) is acknowledged.

References

Achille, C., Fassi, F., Fregonese, L., 2012. 4 Years history: From 2D to BIM for CH: The main spire on Milan Cathedral, in: 2012 18th International Conference on Virtual Systems and Multimedia. Presented at the 2012 18th International Conference on Virtual Systems and Multimedia, 377–382. doi.org/10.1109/VSMM.2012.6365948

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, in: Daniotti, B., Gianinetto, M., Della Torre, S. (Eds.), Digital Transformation of the Design, Construction and Management Processes of the Built Environment. Springer International Publishing, Cham, 331–341. doi.org/10.1007/978-3-030-33570-0_30

Alami, A., Remondino, F., 2024. Querying 3D point clouds exploiting open-vocabulary semantic segmentation of images.

Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLVIII-2-W8-2024, 1–7. doi.org/10.5194/isprs-archives-XLVIII-2-W8-2024-1-2024

Beucher, S., Meyer, F., 1992. The Morphological Approach to Segmentation: The Watershed Transformation, in: Mathematical Morphology in Image Processing. CRC Press.

Canny, J., 1983. A Variational Approach to Edge Detection. AAAI-83 Proceedings, 54-58.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-End Object Detection with Transformers. doi.org/10.48550/arXiv.2005.12872

Coletti, C., Antonelli, F., Germinario, L., Maritan, L., Piovesan, R., Tesser, E., Mazzoli, C., 2025. Investigating stone materials from some European cultural heritage sites for predicting future decay. Rendiconti Lincei Sci. Fis. E Nat. doi.org/10.1007/s12210-024-01298-x

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

Fassi, F., Achille, C., Mandelli, A., Rechichi, F., Parri, S., 2015. A New Idea of Bim System for Visualization, Web Sharing and Using Huge Complex 3d Models for Facility Management. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* XL-5-W4, 359–366. doi.org/10.5194/isprsarchives-XL-5-W4-359-2015

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2018. Mask R-CNN. doi.org/10.48550/arXiv.1703.06870

Idjaton, K., Desquesnes, X., Treuillet, S., Brunetaud, X., 2021. Stone-by-Stone Segmentation for Monitoring Large Historical Monuments Using Deep Neural Networks, in: Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R. (Eds.), Pattern Recognition. ICPR International Workshops and Challenges. Springer International Publishing, Cham, 235–248. doi.org/10.1007/978-3-030-68787-8_17

Jocher, G., Qiu, J., 2024. Ultralytics YOLO11.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment Anything. doi.org/10.48550/arXiv.2304.02643

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, Jie, Jiang, Q., Li, C., Yang, Jianwei, Su, H., Zhu, J., Zhang, L., 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. doi.org/10.48550/arXiv.2303.05499

Llamas, J., M. Lerones, P., Medina, R., Zalama, E., Gómez-García-Bermejo, J., 2017. Classification of Architectural Heritage Images Using Deep Learning Techniques. Appl. Sci. 7, 992. doi.org/10.3390/app7100992

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. doi.org/10.48550/arXiv.1411.4038

Lucho, S., Treuillet, S., Desquesnes, X., Leconge, R., Brunetaud, X., 2024. Weakly Supervised SVM-Enhanced SAM Pipeline for Stone-by-Stone Segmentation of the Masonry of the Loire Valley Castles. J. Imaging 10, 148. doi.org/10.3390/jimaging10060148

Mohammed, A.Sh., Mihoub, Z., 2024. A Review of Image Segmentation Strategies from Classical Methods to Deep Learning, in: 2024 Conference of Young Researchers in Electrical and Electronic Engineering (ElCon). Presented at the 2024 Conference of Young Researchers in Electrical and Electronic Engineering (ElCon), 712–718. doi.org/10.1109/ElCon61730.2024.10468368

Perfetti, L., Fassi, F., Gulsan, H., 2019. Generation of Gigapixel Orthophoto for the Maintenance of Complex Buildings. Challenges and Lesson Learnt. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* XLII-2-W9, 605–614. doi.org/10.5194/isprs-archives-XLII-2-W9-605-2019

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. doi.org/10.48550/arXiv.2103.00020

Réby, K., Guilhelm, A., De Luca, L., 2023. Semantic Segmentation using Foundation Models for Cultural Heritage: an Experimental Study on Notre-Dame de Paris, in: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Presented at the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW),1681–1689. doi.org/10.1109/ICCVW60793.2023.00184

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. doi.org/10.48550/arXiv.1506.02640

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. doi.org/10.48550/arXiv.1505.04597

Sithole, G., 2008. Detection of Bricks in a Masonry Wall, in: The International Archives of the Photogrammetry, 567–72.

Spettu, F., Teruggi, S., Canali, F., Achille, C., Fassi, F., 2021. A HYBRID MODEL FOR THE REVERSE ENGINEERING OF THE MILAN CATHEDRAL. CHALLENGES AND LESSON LEARNT, in: ARQUEOLÓGICA 2.0 - 9th International Congress & 3rd GEORES - GEOmatics and pREServation. doi.org/10.4995/arqueologica9.2021.12138

Verges-Belmin, V., 2008. Illustrated glossary on stone deterioration patterns. ICOMOS.

Zhou, Q.-Y., Park, J., Koltun, V., 2018. Open3D: A Modern Library for 3D Data Processing. doi.org/10.48550/arXiv.1801.09847

Zuiderveld, K., 1994. VIII.5. - Contrast Limited Adaptive Histogram Equalization, in: Heckbert, P.S. (Ed.), Graphics Gems. Academic Press, pp. 474–485. doi.org/10.1016/B978-0-12-336156-1.50061-6