A Roman Carved Tale Modelled in 3D and Interpreted with AI

Gabriele Mazzacca[1,2] , Andrea Sterpin[3,4], Simone Rigon[1], Marco Medici[3], Fabio Remondino[1]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy – (gmazzacca, srigon, remondino)@fbk.eu
[2] Dept. Mathematics, Computer Science and Physics, University of Udine, Italy
[3] INCEPTION, Spinoff of the University of Ferrara, Italy – marco.medici@inceptionspinoff.com
[4] Department of Architecture, University of Ferrara, Italy – andrea.sterpin@unife.it

Keywords: Photogrammetry, 3D Modelling, Semantic Segmentation, Multimodal Large Language Models.

Abstract

This study proposes an innovative methodology for documenting and semantically analysing cultural heritage by integrating artificial intelligence (AI) with a photogrammetric 3D model. The case study is the Trajan’s Column in Rome, a monumental structure adorned with a continuous helical relief depicting Emperor Trajan’s Dacian campaigns. AI-driven semantic segmentation is used to identify key elements (such as human figures, battle scenes and natural motifs) within the digitised sculptural narrative. Starting from a high-resolution photogrammetric 3D model, the column’s texture is divided into multiple segments and a multimodal large language model (MLLM) is applied to produce context-aware segmentation masks via natural language prompts. Results are then projected onto the 3D geometry and visualised through a web-based 3D viewer.

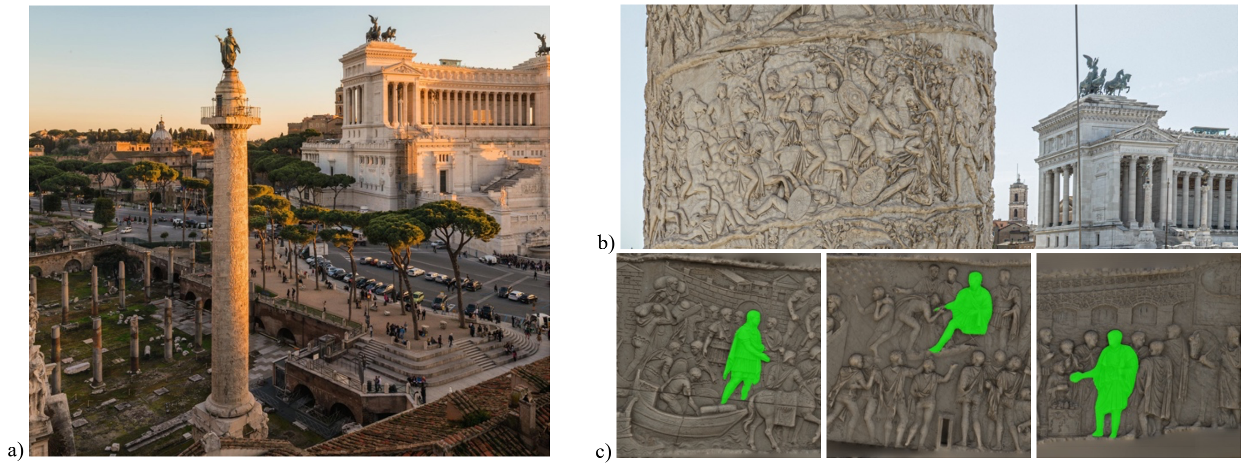

Figure 1. The Trajan’s column in Rome, ca 44 m height, located within the forum (a; source: National Geographic); close-up of the tale which commemorate Traian's win against the Dacian (b); MLLM-based identification of the Emperor Trajan (repeated 58 times) within the 200 m long carved frieze (c).

1. Introduction

In recent years, 3D documentation of Cultural Heritage scenarios has been democratised with reliable, efficient and automated procedures based on image or range data, acquired from ground or UAV platforms (Remondino, 2011; Remondino et al., 2017; Stathopoulou et al., 2019; Farella et al., 2022). However, automatic semantic enrichment and analysis of such 3D models remain an open challenge. Augmenting 3D data with qualitative information is crucial for enhancing its accessibility and usability, yet this task has posed significant difficulties for years (De Luca, 2013; Maietti et al., 2018; Yang et al., 2023). Cultural heritage assets, with their complex geometries, material diversity, and stylistic variability, present significant challenges for recently developed artificial intelligence (AI) solutions for semantic segmentation.

This paper focuses on the emerging opportunities offered by multimodal large language models (MLLMs), which leverage natural language and zero-shot learning strategies. The proposed method uses photogrammetrically derived mesh textures as input for the AI-based semantic segmentation process.

To validate the proposed workflow, the study focuses on the Trajan’s Column, an iconic monument in the Imperial Fora of Rome, renowned for the helical frieze that encircles and adorns its shaft. As highlighted by Hölscher (2004) and Conti and Moffa (2022), the frieze of Trajan’s Column functions as a monumental semantic system: a complex compositional narrative that projects the symbolic power of imperial triumph into stone, combining historical accuracy and evocative strength in a form of immersive 3D storytelling.

This iconic carved narrative serves as a compelling case study for testing the capabilities of the proposed segmentation approach. Nevertheless, the intricate compositional layout, stylised figures and scenes, monochromatic surfaces, and visible erosion make the automated semantic segmentation particularly challenging.

1.1 Paper’s aim

The work presents the 3D documentation of the Trajan's Column in Rome and its interactive analysis through a multimodal large language model (MLLM). The paper will examine the following innovative aspects of this work:

- the complex 3D survey of the Column;
- an automated methodology to unwrap the 3D column and derive high-resolution orthoimages of the carved frieze;
- a zero-shot referring image segmentation process that uses a natural language-driven querying interface, requires no specific retraining, and is capable of interpreting, identifying,

and segmenting scenes, characters, and anthropic or natural elements within the carved narrative;

- an online 3D viewer for interactive access to the results.

Compared to other works that have coupled AI foundation models to perform segmentation of low-level heritage classes (Reby et al., 2023), our approach relies solely on recent MLLM foundation models to tackle complex semantic segmentation tasks that require advanced image and text understanding as well as interpretative capabilities.

2. Related works

The semantic enrichment of 2D and 3D digital heritage data fundamentally improves their interpretative potential and usability, enabling more nuanced documentation, analytical processes and visualisations (Poux et al., 2017; Teruggi et al., 2021). For years, authors have presented manual and interactive frameworks to represent 3D heritage structures based on their morphological description and constituent elements (Attene et al., 2007; Manferdini et al., 2008; Serna et al., 2012; De Luca et al., 2013).

The semantic segmentation of 2D images has been a subject of research for many years, from the early classical methods based on pixel-level features and graph-based algorithms (Boykov and Jolly, 2001) to the first deep learning approaches, such as AlexNet model, which demonstrated unprecedented performance on the largest image dataset of the time, ImageNet (Krizhevsky et al., 2012). Then, the introduction of convolutional neural networks revolutionised the field, enabling end-to-end segmentation without the need for manually designed low-level features. Long et al. (2015) introduced Fully Convolutional Networks for entire images, preparing the arrival of subsequent models such as U-Net (Ronneberger et al., 2015) and Mask R-CNN (He et al., 2017), which are now fundamental for the segmentation of complex images.

The last decade has seen significant progress in 3D semantic segmentation of point clouds. Early efforts employed traditional machine learning methods (Weinmann et al., 2015; Grilli et al., 2019), but researchers have increasingly investigated deep learning approaches (Matrone et al., 2020a,b). PointNet and PointNet++ (Qi et al., 2017a, 2017b) pioneered this transition, and subsequent models, such as PointCNN (Li et al., 2018), KPConv (Thomas et al., 2019), Point Transformer (Zhao et al., 2021) and Super Point Transformer (Robert et al., 2023), have further expanded the capabilities of 3D semantic segmentation techniques.

A third option, involving the integration of 2D and 3D methodologies, has also seen a great deal of research in the last years, with two main types of approaches tested. The first employs deep learning techniques on high-resolution images obtained from photogrammetric surveys, to then accurately reproject inferences onto the corresponding 3D point clouds by leveraging known camera orientations (Bassier et al. 2024). The second focuses on the direct semantic segmentation of textured 3D meshes, generating UV maps from the textured 3D model, training segmentation algorithms to identify semantic classes within the texture space and then mapping back the segmented texture onto the 3D meshes (Grilli et al., 2018; Grilli and Remondino, 2019).

The connection between the 2D and 3D world opens up the possibility to leverage the most recent advancements in the Computer Vision field. Specifically, the employment of recent foundation models (i.e. large deep learning models trained on massive amounts of data) is used to accomplish a wide variety of functions. Nevertheless, the use of foundation models in digital heritage remains limited due to the peculiarity of the subject, which is often very different from the data used to train these foundation models. First attempts were done by Reby et al. (2023) combining Segment Anything (SAM) (Kirillov et al., 2023), Grounding Dino (Liu et al. 2024) and CLIP (Radford et al., 2021) to perform low-level characterisation of building objects. More recently, the rise of Multi-modal Large Language Models (MLLMs) has allowed more complex image scene-understanding tasks using natural language inputs (Fei et al., 2024; Yan et al., 2024; Zhang et al., 2024). One of these tasks, known as referring image segmentation (RIS), allows a user to describe the target object for segmentation through the means of descriptive textual inputs ("prompts"). This is a highly desirable feature in the cultural heritage field because it leverages experts' core knowledge to generate precise prompts that produce meaningful segmentation results.

3. The Trajan's Column in Rome

Erected in 113 AD at the centre of Trajan's forum, the column is widely regarded as one of the most sophisticated expressions of Roman imperial visual culture and rhetoric (Beard, 2007). Commissioned by Emperor Trajan to commemorate the Dacian wars (101-102 and 105-106 AD), it stands as a masterpiece of ancient engineering and sculptural narrative, serving for centuries as a monument of historical memory (Beard, 2007; Coarelli, 2000). With an overall height of approximately 44 meters (including a 6.3 m for the base, 0.8 m for the pedestal, 26.6 m for the shaft, 1.5 m for the platform, 4.7 m for the capital block and a 4.2 m statue on the top), the column is composed of 17 blocks of Luna marble enclosing a helical staircase lit by small slit windows. Its most celebrated feature is the spiralling frieze, stretching over 200 meters in length and depicting more than 2,500 carved figures. The carved narrative sequence reports scenes of marches, battles, military councils, rituals and construction phases, distinguished by naturalism, iconographic richness and refined compositional structure. The visual progression is organised along hierarchical and perspectival lines, guiding the viewer's gaze upward along the shaft. To read the entire story, a viewer would have to move 23 times around the column. Trajan himself appears nearly 60 times in commanding positions, embodying the idealised figure of the Roman *princeps*-strategist, legislator and guarantor of order. Beyond its celebratory function, the column serves as a vital historical source, offering insight into Roman military logistics, Dacian ethnography, construction techniques, and imperial iconography (Lepper and Frere, 1988; Coarelli, 2000; Hölscher, 2004).

Due to its elaborate formal and narrative structure, as well as its complex material characteristics, the Trajan's Column is a representative case study for testing advanced digital approaches for scene understanding in cultural heritage documentation and valorisation.

4. Methodology

4.1 3D surveying and modelling

The geometric survey of the column and its surroundings was first conducted using a multi-sensor approach, combining terrestrial laser scanning (TLS) with photogrammetric techniques. The TLS acquisition was performed with a Leica P40, acquiring a total of 21 scan positions, which are subsequently aligned through iterative closest point (ICP) registration. The pairwise registration yielded a RMSE of 6.7 mm, with a standard deviation of 0.64 mm across the scan pairs, indicating internal consistency and stability within the alignment network. Given the instrumental specifications of the Leica P40 - declaring a typical ranging error of 3 mm at 50 meters - the achieved alignment results are well within the expected precision range of the device, supporting the reliability of the TLS dataset as the primary geometric constraint for the photogrammetric process. In fact, given the ongoing underground works of the metro line below the archaeological area, TLS is adopted as the primary surveying method instead of a traditional network of fixed topographic points. TLS acquisition, indeed, provides a dense and redundant geometric dataset, enabling the selection a posteriori of architectural features for the photogrammetric alignment and control. The final TLS point cloud has approximately 307 mil. points with an average sampling distance of 2 mm (Figure 2).

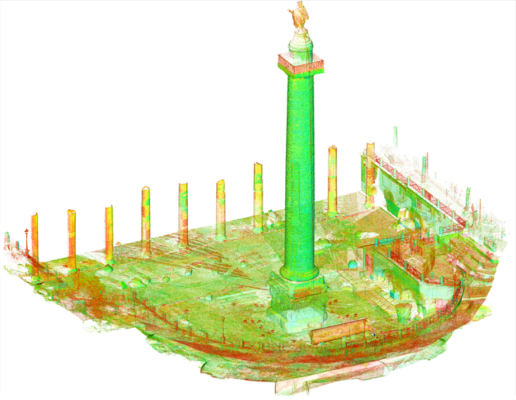

Figure 2. TLS point cloud of the surveyed area (ca 50x50m).

The photogrammetric survey was performed with both terrestrial and drone images (Table 1), acquired ca 2-3 weeks after the TLS.

	Camera	*Resol.*	*Focal*	*# img*	*Min/Max GSD (*)*
ground	Sony Alpha 7 IV	33 Mpx	24-105 mm	801	1-3 mm
drone	DJI Mavic 3M	20 Mpx	24 mm	1485	0.75-2 mm

Table 1. Specifications of the terrestrial and aerial photogrammetric data (* on the Column).

The photogrammetric adjustment integrated both image sets (Figure 3, Table 2), with image orientation and scaling constrained through a set of homologous points identified between the TLS point cloud and the photogrammetric images.

RMSE on Check Points	8.6 mm
Avg redundancy	9.17
Mean reprojection error	1.19 px
Avg. observation per image	6350

Table 2. Bundle adjustment metrics.

The dense image matching process on the Column generated a coloured dense point cloud of approximately 117 million points, providing additional radiometric detail complementary to the TLS geometry. As the used checkpoints (13) revealed a quite high RMSE within the bundle adjustment, the photogrammetric and TLS point clouds are further aligned using an ICP approach. The cloud-to-cloud distance between the two datasets resulted in a mean distance of 5mm (Figure 4c), which is considered a satisfactory metric given the size of the Column.

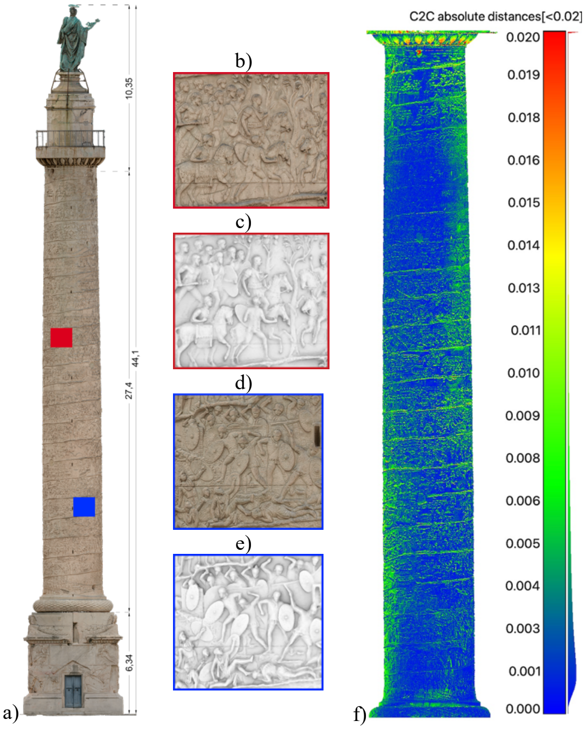

Figure 4. Textured photogrammetric 3D model (a) with close-up views on the carved frieze and tale (b, c, d, e) with and without texture. Cloud-to-cloud comparison of the TLS and photogrammetric surveys (f).

A high-resolution mesh model is then derived from the dense point cloud, containing around 20 million faces. This model is textured with 12 high-resolution texture maps (8192x8192 pixels), featuring a resolution of 1 mm/pix and enabling detailed visual analysis of the monument’s sculptural elements (Figure 4).

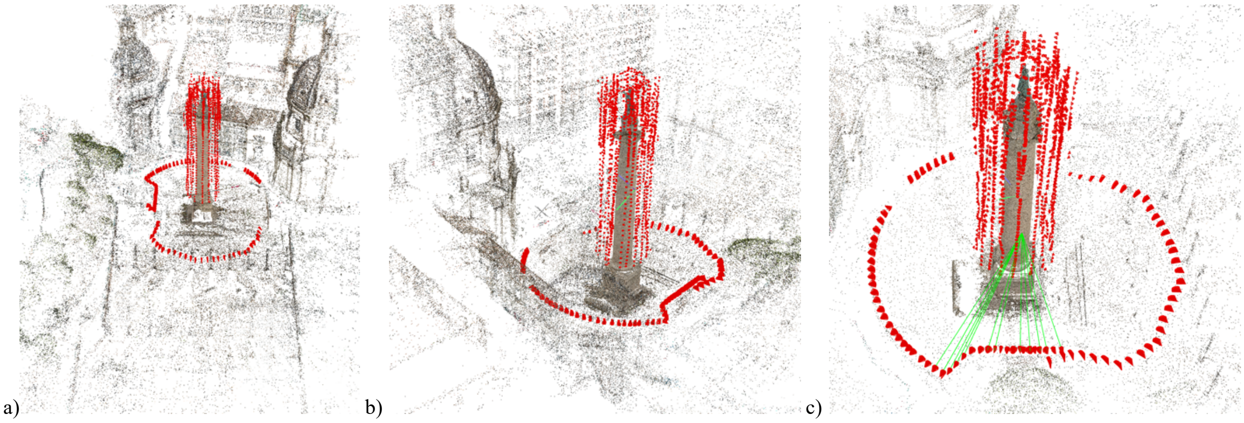

Figure 3. Different views of the recovered camera poses and sparse point cloud of the Column and its surrounding.

https://doi.org/10.5194/isprs-annals-X-M-2-2025-253-2025

4.2 3D model simplification

In addition to the comprehensive 3D documentation of the Column, the development of an AI-assisted tool for the interpretation of the carved frieze required a structured approach to data preparation. Two key processing stages are identified as essential to achieve this objective: (i) the discretisation of the frieze into a continuous, ribbon-like structure, guided by the visual segmentation naturally suggested by the cornice that separates the sculptural registers; (ii) the optimisation of the model’s geometric and textural components to enable efficient AI-based querying, analysis, and visualization.

The high-resolution textured 3D model is processed to generate a simplified mesh representation of the column shaft. This is achieved through the subdivision of the model into smaller, logically organised segments via parametric scripting (*Rhinoceros3D* + *Grasshopper*), producing both a cylindrical (*wrapped*) configuration and a corresponding planar (*unwrapped*) version. Simplified geometries are essential to reduce computational complexity, while smaller mesh segments facilitate the handling of high-resolution textures during both generation and AI-based processing.

A crucial step in this workflow is the extraction of a guiding polyline along the neutral axis of the cornice, which served as a geometric reference for segmentation. This polyline is derived through a surface analysis based on local covariance features - linearity, planarity, surface variation, and omnivariance - ensuring precise alignment with the central trajectory of the sculptural frieze (Figure 5).

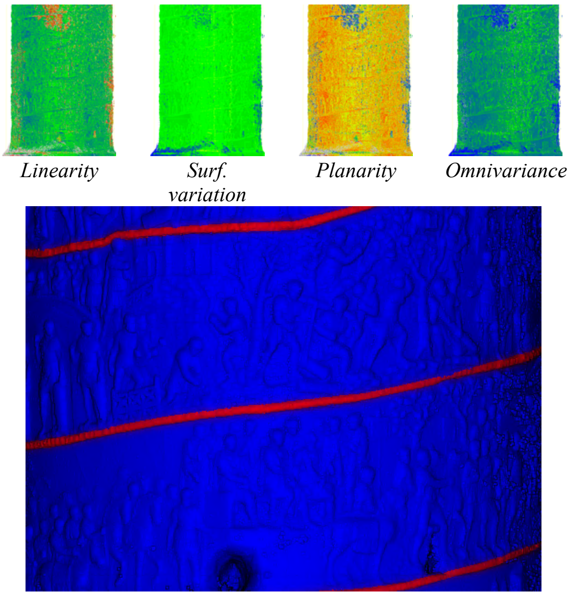

Figure 5. Covariance features at various radii (top) used to extract the polyline (bottom) delimiting the tale along the Column.

As a first step, an approximate cylindrical surface is generated by lofting a series of circular sections with varying diameters, spaced at regular 1-meter intervals along the shaft. This surface provides a geometric reference for the frieze's development. Then, a helical curve with a constant pitch is created following the frieze’s central path. Both the central path curve and the cornice polyline are segmented into 26 polylines using a vertical plane aligned with the column’s longitudinal axis (Figure 6a).

Within each of the 26 segments, a group of three reference polylines is defined to guide the surface generation: the central helical one, representing the main path of the frieze, and two additional polylines, derived from the cornice, marking the lower and upper edges of the sculptural register. This polyline triplet is essential to accurately capture both the unfolding trajectory of the relief and its vertical extent, providing a stable geometric framework for mesh simplification.

A lofting operation is indeed performed between the curves of each group, resulting in the 26 meshes that comprise the simplified wrapped model of the shaft. The use of these three curves allowed the creation of ribbon-like mesh segments that maintain topological consistency across the model, ensuring seamless alignment between adjacent areas as neighbouring meshes share common borders.

This structure also enabled the generation of both a wrapped (cylindrical) and an unwrapped (planar) version of the model, supporting flexible applications in visualisation, texture mapping, and AI-driven analysis.

To preserve dimensional coherence between the two configurations, the central helical polyline served as the primary reference during the unwrapping process. Its segments are reoriented onto a vertical plane, maintaining their original lengths, while the lower and upper polylines are adjusted accordingly to fit the transformed layout. This approach ensured that the simplified mesh retained the proportions and spatial logic of the original column, despite its geometric complexity.

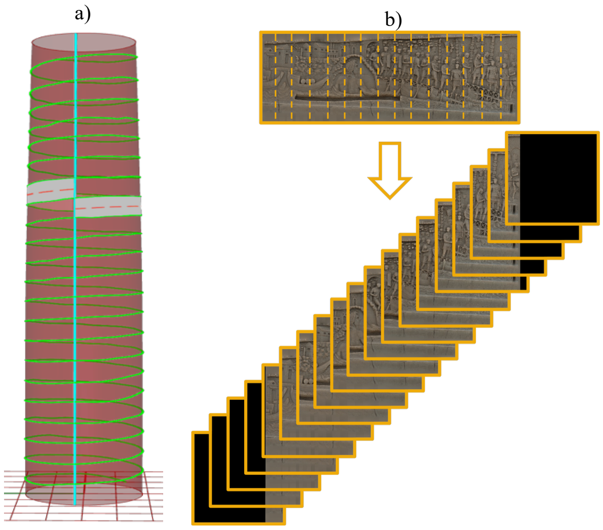

Figure 6. The cylinder split into 26 ribbons (a). Subdivision of textures into multiple overlapping tiles for semantic segmentation (b).

4.3 UV mapping, texturing and surface remeshing

Using the 26 ribbons created as described above, an automated pipeline is developed within *SideFX Houdini* to extract both chromatic and geometric information from the photogrammetric textured mesh and allows for their reprojection onto the simplified (wrapped and unwrapped) mesh counterparts with the goal of reconstructing the monument’s original details.

The process began with the computation of UV coordinates on the simplified wrapped ribbons. Using these coordinates and taking the photogrammetric textured mesh as input, texture maps are generated through baking operations. These included diffuse, ambient occlusion, curvature, and normal maps.

The UV coordinates derived from the wrapped configuration are then transferred to the unwrapped segments, ensuring a consistent spatial correspondence between the two versions of the shaft. This transfer is made possible by the similar topological base structure shared by both configurations, as established through their generation process outlined in Section 4.2. Subsequently, each wrapped segment undergoes a remeshing phase aimed at achieving a denser and more uniform distribution of vertices, thereby enhancing the mesh’s potential for local detail representation. For each newly generated vertex, the distance from the original surface is calculated and stored as a geometric attribute. Then the spatial data is transferred from the wrapped to the unwrapped geometry by leveraging the shared UV layout. The depth of the meshes is then reconstructed, preserving the original geometric features.

In preparation for the AI-based segmentation described in Section 4.4, the images are further processed to aid the model in its task by increasing the visibility of contours and critical details. As the column is made of a single material and presents limited radiometric information, the diffuse map is blended with the curvature map to combine the subtle detail cues highlighted by the shadows in the diffuse map with the geometric features of the frieze, such as edges and creases, emphasised by the curvature map. Since the variation of the curvature is more significant than the absolute depth alone for identifying the contours, the use of the depth map was not considered necessary.

The hybrid images are then subdivided into 1152 tiles of 640×640 pixels (Figure 6b) to feed to the MLLM model. This specific tile size is deemed optimal in our specific case because it offers a good balance between image detail and manageable data size, but working with higher resolution images may produce different results. Each tile has an 80% overlap with the adjacent ones to reach a good contextual understanding of the scene and to avoid excessive straight cuts. Since the original hybrid images have variable dimensions due to the shape of the frieze, particularly in length, the number of tiles generated per image differs accordingly. Some images produce only fifteen tiles, while others, being longer, require fifty or more to ensure full coverage. Because tiles overlap by 80%, each pixel is predicted five times in the final image; a threshold-based vote is then applied to select the final pixel value and generate the final masked images.

4.4 AI semantic segmentation

Among the methods able to perform referring image segmentation (RIS), Sa2VA (Yuan et al., 2025) is employed: Sa2VA is a multi-modal large language model able to segment a specific object in an image or video based on an input linguistic query (“prompt”). Sa2VA combines SAM-2 (Ravi et al., 2024), a foundational video segmentation engine, with LLaVA (Li et al., 2024), a state-of-the-art vision language model. In our pipeline, we handcrafted “instruction tokens” to generate segmentation masks for the interpretation of the Column. According to the carved narrative explored above, five distinct semantic classes are identified on the frieze: *people*, *Emperor Trajan*, *battle scenes*, *Roman military standards*, and *vegetation*. Each of these target classes is chosen because it requires high levels of scene understanding, and each prompt is designed to describe the main characteristics of each class. While the classes *people*, *Roman military standards*, and *vegetation* are mainly discernible by their geometric and physical characteristics, the segmentation of *battle sequences* and *Emperor Trajan* is based on the interpretation of dynamic cues, such as gestures and group interactions, not just physical appearance, and allow us to test the MLLM model capabilities to successfully process highly interpretational queries. Classes requiring very specific domain knowledge, but of potential heritage interest nonetheless, are excluded from the current study because the model lacks specific heritage-based training.

After the prompt designing phase, the generated 1152 image tiles are queried to perform the inference of the target classes. For every tile, a query per class is performed, and the results are validated by a human operator to remove erroneous segmentation. The segmented images are then reassembled into the original 26 ribbons for being visualised on the wrapped and unwrapped meshes of the 3D model.

5. Segmentation and visualisation results

5.1 Quantitative analysis

A subset of the predicted images (6) is used for deriving some metrics to evaluate the performance of the MLLM-based segmentation method. For each target object, the prediction is compared with a corresponding mask annotated by an expert user. To evaluate the outcome of the segmentation pipeline, common Machine Learning metrics are used:

$$Precision = \frac{Tp}{Tp + Fp}$$

$$Recall = \frac{Tp}{Tp + Fn}$$

$$F1\ score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

$$IoU = \frac{Tp}{Tp + Fp + Fn}$$

where, for each considered class, Tp (true positive), Tn (true negative), Fp (false positive), and Fn (false negative) come from the computed Confusion Matrix. All metrics, reported in Table 3, are the averages of the metrics calculated for each class on each image.

%	*People*	*Emperor Trajan*	*Battle scenes*	*Roman military standards*	*Vegetation*
Precision	79.09	89.66	45.26	79.28	92.13
Recall	87.71	87.33	47.71	80.70	58.19
F1 Score	82.07	88.44	46.34	79.84	63.02
IoU	69.78	83.39	43.41	77.07	53.22

Table 3. Per-class metrics for the MLLM-based segmentations (visual in Figure 7).

5.2 Qualitative analysis

The segmentation results across all five classes (*people*, *Emperor Trajan*, *battle scenes*, *Roman military standards*, and *vegetation*), while occasionally incomplete, often provided meaningful semantic insights, especially considering the complexity of the scenes (Figure 7).

The prediction of the class *people* is generally correct, especially in scenes with clearly defined human figures. Despite some inconsistencies, such as the inclusion of shields in some instances and their exclusion in others, Roman soldiers are correctly identified and segmented. This suggests that more precise prompt phrasing, particularly concerning the description of weapons, could yield more consistent results. Inconsistencies due to partial or over-segmentation can also be seen in other instances, probably due to occlusion problems (human figures hidden by other objects in the scene) or lack of crisp contours, determined by the damaged eroded surface of the frieze in that area. Finally, over-segmentation of human figures by including nearby animals.

For the *Emperor Trajan* class, the segmentation proved to be particularly challenging. The Emperor's depiction in the frieze is not consistently clear, often blending with other figures, making it difficult to distinguish even for the expert human eye. The MLLM had to identify Trajan as a subcategory within the broader *people* class. Despite a detailed prompt (“short and cropped hair, no beard, no helmet, no laminar armour, wearing a long, pleated tunic that reaches his knees and a cloak draped over his shoulders and across one of his arms...”), some partial and incorrect classifications or clear misidentifications in more ambiguous contexts are present.

The *battle scenes* class yielded the highest number of segmentation errors. Given the stylised and dynamic nature of combat illustrations, character postures and contextual elements often led to confusion. Missing or false positive detections are present in the results, e.g. in scenes depicting troops in motion but without explicit combat cues.

The Roman military standards category features small dimensions of the emblems, morphological resemblance to other vertical elements (such as limbs or branches) and occasional surface erosion. Nonetheless, the segmentation results are generally satisfactory, despite errors due to partial occlusions, erosions or crowded areas.

The *vegetation* class, though stylised and often depicted in the background, is segmented with generally positive results. Trees and branches are correctly identified in many instances, though partial or incorrect segmentations are also observed.

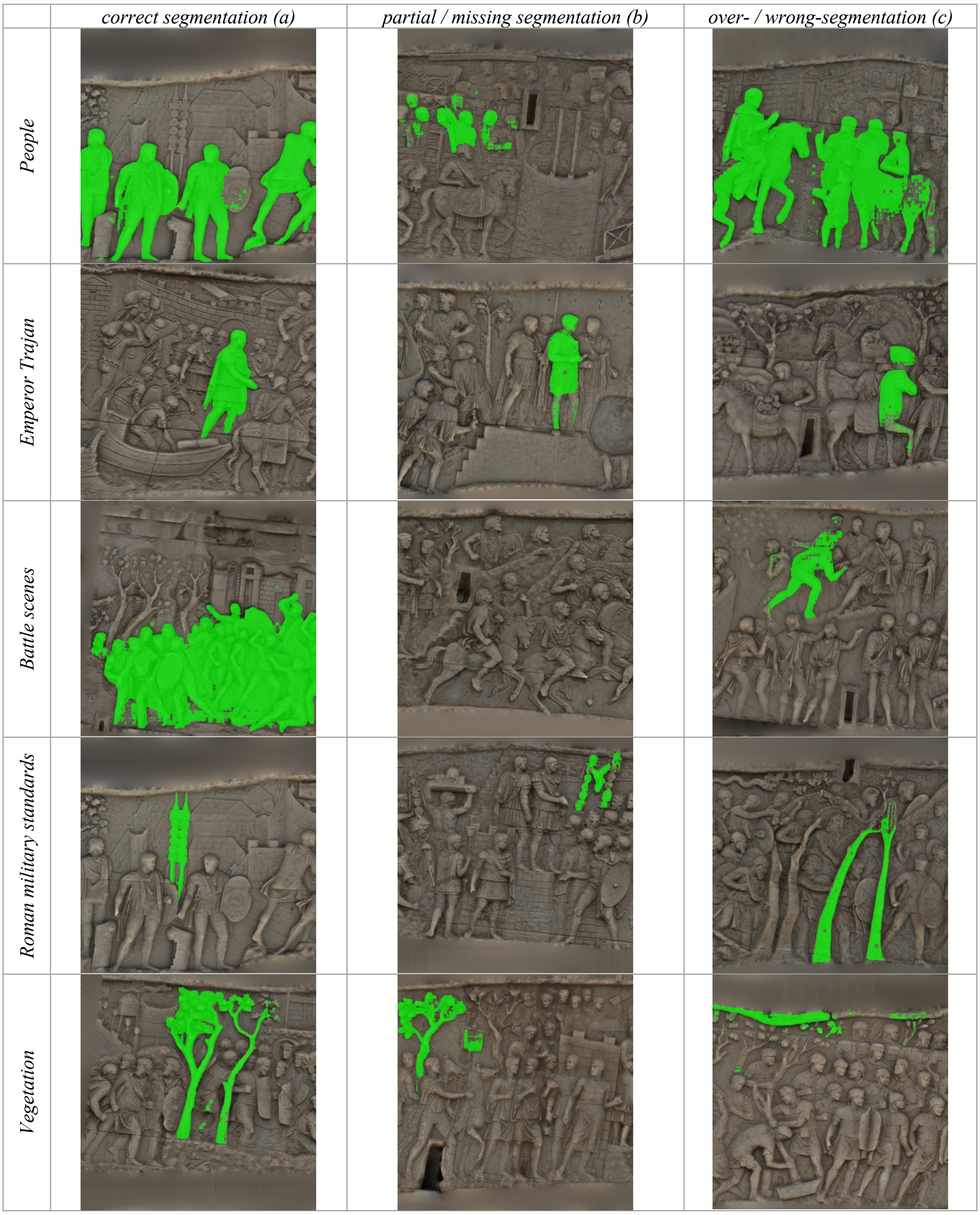

Figure 7. Visual results of the MLLM-based segmentation of the frieze’s texture for the selected classes. Metrics in Table 3.

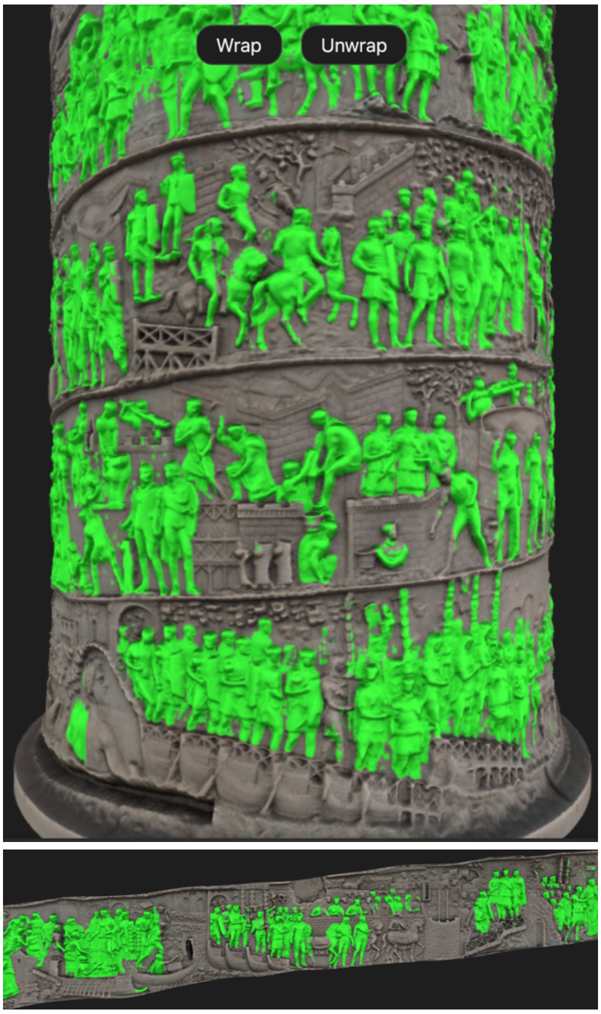

Figure 8. Interactive online viewer displaying the wrapped (top) and unwrapped (bottom) 3D model of Trajan’s Column with segmentation results for the *people* class.

6. Conclusions

This study shows that semantic segmentation of complex 3D heritage models, guided by natural language and enabled by multimodal large language models (MLLMs) in a zero-shot configuration, is technically feasible and produces meaningful results without domain-specific training or manual annotations. Despite promising outcomes, some limitations persist due to the domain representation gap in existing foundation models and the iconographic complexity of the heritage case study”

- Limited domain representation: cultural heritage, especially ancient sculpture, is poorly covered in most training datasets, making it challenging for models to process textures that are monochromatic, stylised, or eroded - very different from typical modern training images.
- Iconographic complexity: figures are often occluded, overlapping, ruined by the time or shown in non-standard perspectives, which complicates automatic recognition. Semantic boundaries can appear fragmented or ambiguous, with errors even in non-eroded and clearly legible areas.
- Fragmented context: the frieze’s length required generating more than 1000 texture tiles. Even with overlapping tiles and a voting mechanism, the system analyses only partial scenes at a time, limiting its ability to recognise relationships between adjacent figures. This affects semantic coherence but the workflow still performs robustly under these constraints.

Nevertheless, achieved results point to clear opportunities for further refinement, which would require extending beyond a pure zero-shot approach. Introducing a small set of manually segmented examples could indeed adapt the model to the unique visual and semantic features of ancient friezes, improving its sensitivity to carved contours, symbolic motifs, and stylistic conventions. Moreover, adding structured historical or iconographic information - such as figure roles, attributes, or scene hierarchies - would support clearer semantic disambiguation and more consistent labelling across complex scenes.

In conclusion, the complete workflow - from 3D data acquisition to automated texture generation and zero-shot AI-based segmentation - demonstrates a scalable and adaptable pipeline for digital heritage documentation and analysis. By further refining the AI methods integrating domain-specific training and contextual data, the reported workflow can be replicated and reused across a broad range of decorated 3D artefacts: spiral columns, sarcophagi, temple reliefs, narrative friezes, or any context where figurative content is distributed on curved or morphologically complex surfaces. To maximise accessibility and scholarly reuse, the segmentation results are also being deployed through an interactive web-based 3D viewer (Figure 8) developed with Three.js, supporting detailed exploration, annotation, and cross-disciplinary collaboration.

Acknowledgments

The data acquisition in Rome was carried out by INCEPTION s.r.l. as part of the works of Cooperativa Archeologia, commissioned by Metro C and authorized by Roma Metropolitane.

References

Attene M., Robbiano F., Spagnuolo M., Falcidieno B., 2007. Semantic Annotation of 3D surface meshes based on feature characterisation. *Lecture Notes in Computer Science*, Vol. 4816, pp. 126-139.

Bassier, M., Mazzacca, G., Battisti, R., Malek, S., Remondino, F., 2024. Combining image and point cloud segmentation to improve heritage understanding. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-2/W4-2024, 49-56.

Beard, M., 2007: *The Roman Triumph*. Harvard University Press, Cambridge, MA and London, England.

Boykov, Y., Jolly, M.-P., 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: *Proc. ICCV*, 105-112.

Coarelli, F., 2000. *The Column of Trajan*. Colombo, Roma.

Conti, C., Moffa, F., 2022. *Lectures on Trajan’s Column and its Architect Apollodorus of Damascus*. In “L'Erma di Bretschneider”.

De Luca, L., Busayarat, C., Stefani, C., Veron, P., Florenzano, M., 2013. A semantic-based platform for the digital analysis of the architectural heritage. *Computers & Graphics*, Vol. 35(2).

De Luca, L., 2013. 3D Modelling and Semantic Enrichment in Cultural Heritage. Proc. Photogrammetric Week, pp. 323-333.

Farella, E.M., Morelli, L., Rigon, S., Grilli, E., Remondino, F., 2022. Analysing Key Steps of the Photogrammetric Pipeline for Museum Artefacts 3D Digitisation. *Sustainability*, 14, 5740.

Fei, H., Wu, S., Zhang, H., Chua, T. S., & Yan, S., 2024. Vitron: A unified pixel-level vision LLM for understanding, generating, segmenting, editing. Proc. *NeurIPS*.

Grilli, E., Dininno, D., Petrucci, G., and Remondino, F., 2018. From 2D to 3D supervised segmentation and classification for cultural heritage applications, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLII-2, 399-406.

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

Grilli, E., Farella, E. M., Torresani, A., Remondino, F. 2019. Geometric features analysis for the classification of cultural heritage point clouds. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. 42, 541-548.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. Proc. *ICCV*, 2961–2969.

Hölscher, T., 2004. *The Language of Images in Roman Art*. Cambridge University Press, Cambridge.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., & Girshick, R. 2023. Segment anything. Proc. *ICCV*, pp. 4015-4026.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. ImageNet classification with deep Convolutional Neural Networks. Proc. *NeurIPS*.

Lepper, F., Frere, S., 1988. *Trajan's Column: A New Edition of the Cichorius Plates*. Alan Sutton Publishing, Gloucester.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C., 2024. LLaVA-onevision: Easy visual task transfer. *arXiv preprint arXiv*:2408.03326.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., & Chen, B., 2018. PointCNN: Convolution on x-transformed points. *Advances in neural information processing systems*, 31.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv*:2303.05499, 2023.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. Proc. *CVPR*, 3431-3440.

Maietti, F., Medici, M., Ferrari, F., Ziri, A. E., & Bonsma, P., 2018. Digital cultural heritage: Semantic enrichment and modelling in BIM environment. *Lecture Notes in Computer Science*, Vol 10605. Springer, pp. 104-118.

Manferdini, A.M., Remondino, F., Baldissini, S., Gaiani, M., Benedetti, B., 2008. 3D modeling and semantic classification of archaeological finds for management and visualization in 3D archaeological databases. Proc. *VSMM*, pp. 221-228.

Matrone, F., Grilli, E., Martini, M., Paolanti, M., Pierdicca, R., Remondino, F., 2020a. Comparing Machine and Deep Learning methods for large 3D heritage semantic segmentation. *International Journal of Geo-Information*, 9, 535.

Matrone, F., Lingua, A., Pierdicca, R., Malinverni, E. S., Paolanti, M., Grilli, E., Remondino, F., Murtiyoso, A., Landes, T., 2020b. A benchmark for large-scale heritage point cloud semantic segmentation. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B2-2020, 1419-1426.

Poux, F., Neuville, R., Van Wersch, L., Nys, G.-A., Billen, R., 2017. 3D point clouds in archaeology: Advances in acquisition, processing and knowledge integration applied to quasi-planar objects. *Geosciences*, *7*, 96.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. PointNet: Deep learning on point sets for 3D classification and segmentation. Proc. *CVPR*, 652-660.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 5105-5114.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I., 2021. Learning transferable visual models from natural language supervision. Proc. *ICML*, pp. 8748-8763.

Ravi, N., Gabeur, V., Hu, Y. T., Hu, R., Ryali, C., Ma, T., ... & Feichtenhofer, C., 2024. SAM2: Segment anything in images and videos. *arXiv preprint arXiv:*2408.00714.

Réby, K., Guilhelm, A., & De Luca, L. 2023. Semantic segmentation using foundation models for cultural heritage: An experimental study on Notre-Dame de Paris. Proc. *ICCV*, pp. 1689-1697.

Remondino, F., 2011. Heritage Recording and 3D Modeling with Photogrammetry and 3D Scanning. *Remote Sensing*, 3(6), pp. 1104-1138.

Remondino, F., Nocerino, E., Toschi, I., Menna, F., 2017. A critical review of automated photogrammetric processing of large datasets. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. XLII-2/W5, pp. 591-599.

Robert, D., Raguet, H., & Landrieu, L., 2023. Efficient 3D semantic segmentation with superpoint transformer. Proc. *ICCV*.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. Proc. *MICCAI*.

Secord, J., & Zakhor, A., 2007. Tree detection in urban regions using aerial lidar and image data. *IEEE Geoscience and Remote Sensing Letters*, 4(2), 196-200.

Serna, S.P., Schmedt, H., Ritz, M., Stork, A., 2012. Interactive Semantic Enrichment of 3D Cultural Heritage Collections. Proc. *VAST*, Eurographics Association, pp. 33-40.

Stathopoulou, E.K., Welponer, M., Remondino, F., 2019. Open-source image-based 3D reconstruction pipeline: review, comparison and evaluation. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. XLII-2/W17, pp. 331-338.

Teruggi, S., Grilli, E., Fassi, F., Remondino, F., 2021. 3D surveying, semantic enrichment and virtual access of large cultural heritage. *ISPRS Annals Photogramm. Remote Sens. Spatial Inf. Sci.,* VIII-M-1-2021, 155–162

Thomas, H., Qi, C. R., Deschaud, J. E., Marcotegui, B., Goulette, F., & Guibas, L. J., 2019. Kpconv: Flexible and deformable convolution for point clouds. Proc. *ICCV*, pp. 6411-6420.

Weinmann, M., Jutzi, B., Hinz, S., Mallet, C., 2015. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 286-304.

Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., ... & Gavves, E., 2024. Visa: Reasoning video object segmentation via large language models. Proc. *ECCV*, pp. 98-115.

Yang, S., Hou, M., Li, S., 2023. Three-Dimensional Point Cloud Semantic Segmentation for Cultural Heritage: A Comprehensive Review. *Remote Sensing*, 15(3):548

Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., ... & Yang, M. H., 2025. Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos. *arXiv preprint arXiv*:2501.04001.

Zhang, T., Li, X., Fei, H., Yuan, H., Wu, S., Ji, S., ... & Yan, S. 2024. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in Neural Information Processing Systems*, *37*, 71737-71767.

Zhao, H., Jiang, L., Jia, J., Torr, P. H., & Koltun, V., 2021. Point transformer. Proc. ICCV, pp. 16259-16268.