3DGeoRef: an automated framework for georeferencing heritage 3D models

Simone Rigon[1], Elisa Mariarosaria Farella[1], Luca Morelli[1], Gianluca Bertolasi[1], Fabio Remondino[1], Sander Münster[2]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Email: (srigon, elifarella, lmorelli, gbertolasi, remondino)@fbk.eu

[2] Digital Humanities, Friedrich-Schiller-Universität Jena, Germany – Email: sander.muenster@uni-jena.de

KEY WORDS: Data Space for Cultural Heritage, 3D heritage models, georeferencing, VLM, multimodal models.

ABSTRACT:
Geolocalization is the process of determining the precise geographical coordinates and orientation of a device or person or object. Georeferencing is the same process but referred to images, maps or 3D models. Pinpointing where in the world an image was acquired or a 3D model is located, down to a decimetre or meter error, remains a challenge in photogrammetry and computer vision, especially when no priors are available or not touristic locations are considered. Methods have evolved from simple GNSS tagging to complex computer vision and AI-driven spatial reasoning, including the recent LLM/VLM/MLLM approaches. This work presents an automated pipeline to georeference heritage 3D models lacking geolocation metadata. The developed method combines synthetic views generation of a not-georeferenced 3D model, VLM and multimodal-based location estimation, satellite imagery retrieval and learning-based image matching techniques to determine the transformation to align the 3D model with real-world coordinates. Results are below the meter accuracy if a substantial amount of surrounding data is available to support the inference of the initial rough location. The final aim of the presented pipeline is to supplement the Cultural Heritage European Data Space with enriched 3D models.

Figure 1: Example of 3D models retrieved from public repositories and used for testing the proposed automatic georeferencing tool.

1. Introduction

In 2021, with the launch of the Common European Data Space for Cultural Heritage (CNECT/LUX/2021/OP/0070), new policy and technical frameworks have been designed to boost the sharing, interoperability, and reuse of digital heritage assets across Europe. This European Commission initiative, part of the broader European Data Space ecosystems and led by Europeana Foundation, paved the way for several related projects and programs, targeting to expand and support European cultural and creative infrastructure. Within this framework, the 3DBigDataSpace project[1] was established to contribute to the common Cultural Heritage European Data Space[2] infrastructure and solicit enhanced access to high-quality 3D heritage contents. The project aims at overcoming the lack of 3D models for various applications, including XR, by exploiting large-scale public 3D repositories offered by providers and aggregators, while enhancing their geometric and radiometric quality, and enriching their descriptive semantics. A set of tools and frameworks is currently under development for the data optimization and enrichment, which will be tested and applied on thousands of retrieved 3D heritage models, with a specific focus on three main tasks:

- Geometry optimization (simplification for XR uses and remeshing) and texture enhancement (through the generation of PBR - Physically Based Rendering - maps);
- 3D models classification (category/domain identification and point clouds segmentation and classification);
- 3D heritage models georeferencing.

For these activities, both manual and automatic methodologies are set up, as complementary/alternative approaches to guarantee, respectively, high-quality control and ground truth references, and faster and more efficient processing of large volumes of data.
With a focus on the georeferencing and spatialization task, two different tools have been designed: an interactive tool, based on a 3D geospatial engine to visualize and interact with Earth-scale geospatial data and to manually adjust the scale and position of the imported models; an end-to-end AI-based pipeline, described in this article, to automatically derive positional information of a not-georeferenced 3D model and transform its coordinates and scale based on real-world data.

[1] https://3dbigdataspace.eu/

[2] https://www.dataspace-culturalheritage.eu/en

Georeferencing assigns spatial meaning (coordinates, scale, orientation and, eventually, projection) to archival and modern digital data so they can be integrated into geographic information systems (GIS) or online 3D viewers, compared across time or fused with other spatial information. In the cultural heritage sector, a broad variety of input data is available, including historic maps and plans, archival photographs, terrestrial and aerial imagery or LiDAR data. The core challenge is matching these heterogeneous sources, often of differing scale, accuracy, and completeness, compared to a common, real-world coordinate system.

Recent advances in AI offer some opportunities to address some of these challenges, reducing reliance on manual input or missing metadata. In particular, Vision Language Models (VLMs) and multimodal large language models (MLLMs) are emerging as powerful tools for geospatial reasoning, with the ability to infer approximate locations directly from the visual and semantic content of images (Minaee et al., 2025; Danish et al., 2026). Unlike traditional methods, relying on explicit metadata or GNSS tags, AI-driven approaches can leverage contextual cues for retrieving missing geographic information. To refine these course geolocalization outputs, learning-based image matching techniques can be exploited. Recently AI-based solutions are demonstrating a growing robustness compared to handcrafted methods for co-registration, overcoming well-known limitations of traditional approaches when data feature significant differences in scale, perspective or lighting.

1.1 Aim of the work

This article focuses on the implementation and validation of a new AI-based automatic pipeline to georeference and assign real-world scale to thousands of 3D heritage building models retrieved from public open repositories of 3D models. The aim is to offer a solution for effectively processing large volumes of not-georeferenced 3D data, featuring heterogeneous quality and completeness levels, and creating metrically consistent and spatially coherent 3D products. The geolocated outputs will be ingested into the Cultural Heritage European Data Space but also enrich the 4D Browser[3] (Münster et al., 2025).

2. Related works

Recent research on the georeferencing problem has explored and proposed different manual and automatic solutions to assign spatial meaning to archival and modern digital data, with different approaches when handling 2D or 3D sources.

On the 2D and photographs side, research on georeferencing is especially addressing the co-registration of historical and modern data, exploiting geometric and AI solutions for image matching with archival terrestrial (Maiwald et al., 2021; Morelli et al., 2022) and aerial (Craciun and Le Bris, 2022; Farella et al., 2022; Maiwald et al., 2023) data. Other works are specifically tailored to the georeferencing of historical aerial photo indexes (Malek et al., 2025) or topographic maps (Luft, 2020; Milleville et al., 2022). In computer vision, the term image localization is often used (i.e. given a certain image, predict where in the world it was taken), with its sub-tasks Visual Place Recognition - VPR, i.e. determining where an image was acquired given a certain database or known map, and visual localization, i.e. estimating the camera poses (position and orientation).

Img2GPS (Hays and Efros, 2008) is one of the first attempt to estimate a distribution over geographic locations from a single image using a purely data-driven scene matching approach. Authors leveraged a dataset of over 6 million GPS-tagged images from the Internet. PlaNet (Weyand et al., 2016) solved the localization problem as a classification one by subdividing the surface of the earth into thousands of multi-scale geographic cells and train a deep network using millions of geotagged images.

Image georeferencing			
Method	*Principle*	*Accuracy*	*Scalability*
Metadata-based	Use sensors metadata, e.g. GNSS, INS or timestamp	Low-Medium	High
Feature-based	Match features through handcrafted or global descriptors (e.g. NetVLAD, CosPlace)	Low-Medium	High
Learning-based	End-to-end localization with deep neural networks and large labelled datasets	Medium	High
Cross-modal	Match semantic structures (e.g. images with building footprints or 3D models)	Medium	Medium
3D model georeferencing			
Method	*Principle*	*Accuracy*	*Scalability*
Metadata-based	Inherit georeferenced information from 3D modeling process	Medium	High
Feature-based	Match geometric features across 3D data, e.g. using 3D descriptors or manually selected features (e.g. SHOT, ISS)	High	Medium
Learning-based	Learn place descriptors in point clouds (e.g. PointNetVLAD3D)	Medium	High
Cross-modal	Match semantic structures (e.g. CrossLoc3D)	Medium	Medium

Table 1: Image and 3D model georeferencing methods.

	Gemini	**Llama 3.2 via Ollama**	**Geo-clip**	**Scaling-Geoloc**	**PlaNet**	**Img2Loc**	**HLoc**
Type of model	Multimodal foundation model	General purpose LLM	VLM specialized for geolocation	Hybrid large-scale retrieval	Classification	LLM-Hybrid	Geometric
Localization method	Visual context and global knowledge-base	Landmark recognition and/or text understanding	CLIP-based geographic coordinate regression	Contrastive retrieval	CNN-based	Image retrieval and regression	Feature matching + SfM
Weakness	Closed ecosystem, limited transparency, Internet required, privacy concerns	No dedicated global map, no native vision or spatial grounding	Research product, task-specific vision-language model	High computational needs	Low resolution (and outdated)	Requires large training data	Not global, it requires a reference 3D model
Deployment	Cloud-based	Offline / local execution	Offline / local execution	Offline / local execution	Offline / local execution	Offline / local execution	Offline / local execution
Code availability	No, API	yes	yes	No	Yes	Yes	No

Table 2: Main characteristics of some image localization methods.

[3] https://4dbrowser.org/

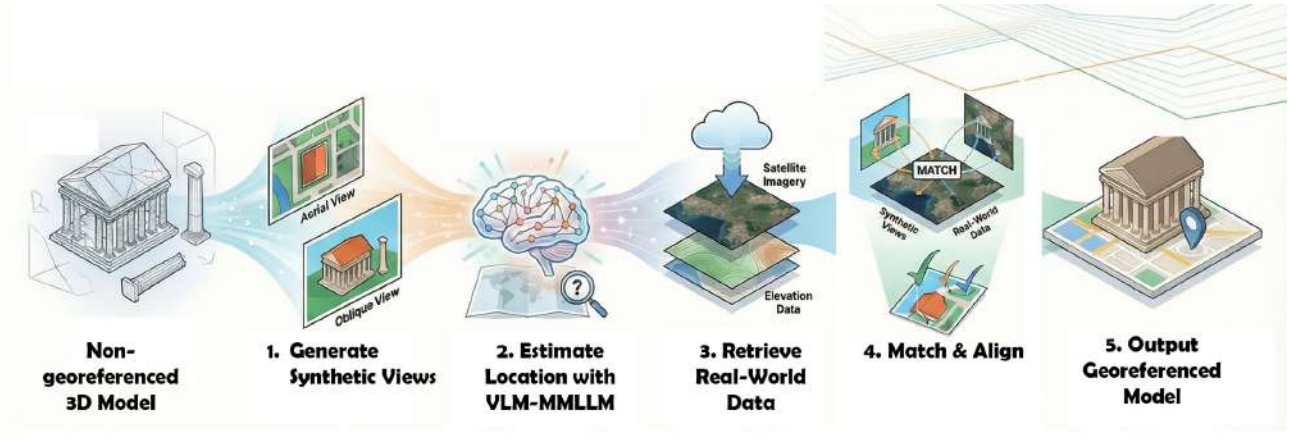

Figure 2: The proposed 3DGeoRef pipeline to georeference a heritage 3D model lacking geolocation metadata.

GeoCLIP (Vivaco et al., 2023) addresses the challenges of worldwide image geo-localization by introducing a novel CLIP-inspired approach that aligns images with geographical locations. Lindenberger et al. (2025) fuses direct image geo-localization via classification, and cross-view retrieval against aerial imagery. Other methods determine the geographic location using cross-modality data (Zhang et al., 2021; Tomesek et al., 2022; Panek et al., 2023).

When coming to 3D models, a consistent part of the recent literature is dedicated to the georeferencing of BIM models for GIS integration (Azari et al., 2025), considering the challenges of the interaction between different domains, georeferencing definitions and coordinate systems. Guan et al. (2023) presented CrossLoc3D, a novel 3D place recognition method that solves a large-scale point matching problem in a cross-source setting.

Panek et al. (2022) presents a feature-based visual localization approach based on 3D meshes. Loeper et al. (2024) presented a visual localization approach for urban environments based on the boundary representation (BREP) of the CityGML models.

Hybrid approaches (e.g. HLoc) integrate feature matching, SFM and image retrieval (Sarlin et al., 2019), e.g. based on NetVLAD (Arandjelovic et al., 2016).

Recently, researchers proposed tools and methodologies to simplify the georeferencing process when dealing with reality-based 3D models (Shaji et al., 2025), also using on Transformers (Pramanick et al., 2022), natural language (Ye et al., 2025) and VLM (Vivanco et al., 2023; Zhou et al., 2024; Cai et al., 2025; Liu et al., 2025; Xu et al., 2025) to regress coordinates of a given image.

3. Methodology

The implemented 3DGeoRef is a framework designed to georeference 3D heritage assets (buildings and open-air monuments) in a fully automated manner and without requiring any prior geographic information (Figure 2). The tool combines synthetic views generation, VLM/MLLM-based location estimation, satellite imagery and elevation data retrieval, and learning-based image matching techniques to determine the transformation to align the not-georeferenced 3D model with real-world coordinates. The proposed end-to-end AI-based pipeline starts from a 3D reality-based heritage model and relies on a sequence of modules responsible for specific tasks, as detailed in the following subsections.

3.1 Synthetic images generation

As an initial step of the automated georeferencing pipeline, a set of synthetic images is generated from the 3D heritage model. Blender is leveraged to render both a nadir and multiple oblique images, simulating aerial views uniformly distributed around the 3D model. The camera positions are automatically defined based on the bounding box of the model and predefined rendering parameters, such as camera height, field of view, and viewing angles. This strategy ensures an adequate visual coverage of the 3D heritage geometry and its surrounding context. HDRI lighting is also applied to enhance the realism of the rendering, simulate natural illumination and reduce the domain gap between synthetic and real imagery. Although nadir views offer limited semantic cues for visual localization (Section 3.2), they are particularly useful for estimating an initial scale of the 3D model, and this coarse scaling step facilitates the subsequent image matching with satellite tiles (Section 3.4).

3.2 Approximate geolocation information

Once obtained the synthetic object views, an approximate geographic location must be estimated to reduce the search space and make the subsequent processing computationally feasible (Section 3.3). This task is addressed by exploiting Vision Language Models (VLMs) and multimodal large language models (MLLMs), which infer geographic information based on the visual appearance of the rendered images (Section 3.1). These images are queried against different AI-based localization models. Three alternatives are currently included and supported within the 3DGeoRef pipeline: the VLM-based GeoCLIP and the multi-modal OLLAMA with L-LAMA (v.3.2) and Gemini models. GeoCLIP is a VLM-based model leveraging contrastive language-image pretraining to match visual features of the rendered views with a large geotagged image database. OLLAMA is a multimodal large language model that integrates both visual and textual cues. Gemini is a multimodal model combining global visual recognition with contextual reasoning. Each of these models can be independently leveraged to provide approximate geographic information (latitude and longitude), allowing flexibility and redundancy within the pipeline. Nadir renderings are excluded from the queries since they typically provide limited contextual cues for the localization, while oblique views are favored, since they better preserve architectural and environmental features, as well as spatial relationships.The estimated coarse position and region are exploited for retrieving the corresponding satellite tiles and the Digital Elevation Model (DEM) (Section 3.3) used for the image matching step (Section 3.4) and final georeferencing (Section 3.5). By constraining satellite and elevation data search and retrieval to a plausible region, this AI-based localization task significantly improves the overall efficiency of the pipeline, reducing both computational costs and the risks of incorrect matches in the subsequent image matching and georeferencing steps.

3.3 Satellite image and elevation data retrieval

Mapbox[4] services are used to retrieve satellite imagery tiles for the region identified by the AI-based geolocalization models (Section 3.2). The retrieval process is iterative and adaptive. It starts from a zoom level corresponding to a target ground sampling distance of approximately 30 cm, suitable for capturing relevant urban and architectural features and details for the image matching. If imagery at this resolution is not available for the location of interest, the pipeline automatically reduces the zoom level, progressively decreasing the spatial resolution until a maximum threshold of 5 m is reached. This strategy ensures the availability of reference satellite data even in less-documented or remote areas. However, while this adaptive strategy increases the pipeline robustness, lower-resolution tiles may negatively impact the performance of the image matching step (Section 3.4), as a fewer distinctive features can be extracted and reliably matched with the rendered synthetic views, with consequences on the final georeferencing accuracy In addition to satellite imagery, elevation data are retrieved using OpenTopoData[5], which provides access to a Digital Elevation Model (DEM) with a spatial resolution of 30 m. While not highly detailed, this DEM dataset is freely available and open-source, ensuring global coverage and enabling full automated processing within the pipeline. The DEM data are integrated into the final 3D model transformation stage (Section 3.5) to ensure a consistent vertical alignment of the 3D heritage asset with real-world terrain.

3.4 Image matching and 2D transformation

Once the satellite imagery is retrieved, the image matching between the synthetic renderings generated from the 3D model (Section 3.1) and the corresponding satellite tiles is performed, exploiting the DIM (Dense Image Matching) library (Morelli et al., 2024)[6]. This library integrates state-of-the-art learning-based feature detection and descriptor methods. In 3DGeoRed, two complementary matching strategies are adopted: LoFTR (Local Feature Transformer) (Sun et al., 2021), which enables detector-free dense correspondence estimation, and a keypoint-based pipeline combining SuperPoint (DeTone, 2018) for feature detection and description with SuperGlue (Sarlin et al., 2020) for robust feature matching. Keypoints are independently extracted from both rendered images and satellite images and matched to identify a set of candidate correspondences (Figure 3). Corresponding feature points are then used to compute an affine 2D transformation, which models translation, rotation, and uniform scaling.

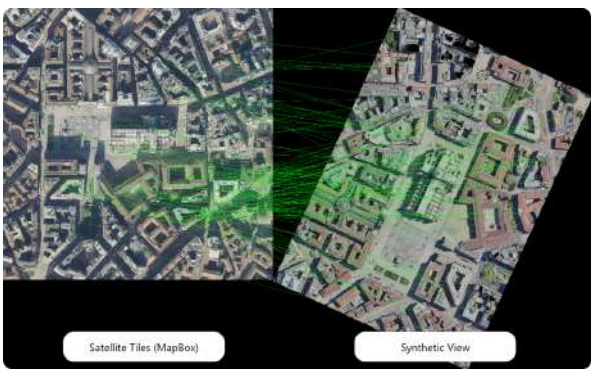

Figure 3. Example of image-matching between a satellite imagery tile (right) and a synthetic model view (left).

3.5 3D model transformation

The transformation parameters estimated during the image matching stage are finally applied to the 3D asset, adjusting its position, orientation, and scale to align with real-world coordinates. Additionally, vertical positioning is refined by integrating elevation corrections obtained from the OpenTopoData DEM (Section 3.3), ensuring a consistent alignment with the with real-world terrain. Once the transformation is applied, the output consists of a fully georeferenced 3D model that can be exported in standard geospatial 3D format, and visualized within common 3D environments, such as Google Earth, Cesium, of further GIS and 3D visualization platforms (Figure 4).

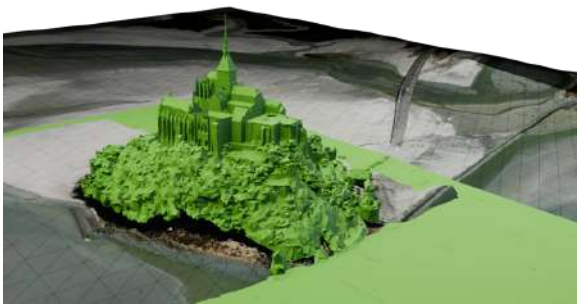

Figure 4. The georeferenced 3D model overlaid with the DEM and orthophoto.

4. Data and experiments

4.1 Data

The employed 3D heritage models (Figure 1) are found in various repositories and offers opportunities to semantically enrich digital assets and increase their accessibility for immersive experiences, research and preservation purposes. At the same time, their integration and reuse in digital infrastructures pose numerous challenges due to the frequent lack of descriptive and contextual information, which limits their value and prevents their effective and meaningful visualization in web or XR-based platforms (Münster et al., 2023). The 3D models considered in our experiments are primarily those generated with surveying reality-based techniques.

Dataset	GEMINI		GEOCLIP		OLLAMA	
	GEMINI	DIM	GEOCLIP	DIM	OLLAMA	DIM
Neuschwanstein Castle (Germany)	✓	✓	✓	✓	✓	✓
Milan Cathedral (Italy)	✓	✓	✗	✗	✓	✓
Parthenon (Greece)	✓	✓	✓	✓	✓	✓
Trento Cathedral (Italy)	✓	✓	✗	✗	✗	✗
Sagrada Familia (Spain)	✓	✓	✓	✓	✓	✓
Torre Eiffel (France)	✓	✓	✓	✓	✓	✓
Odawara Castle (Japan)	✗	✗	✗	✗	✗	✗
Rovereto Castle (Italy)	✗	✗	✗	✗	✗	✗
Loz Feliz – California (USA)	✗	✗	✗	✗	✗	✗
Dinkins Palace-Manhattan (USA)	✓	✓	✓	✓	✓	✓

Table 3. GeoRef results testing different AI-based localization solutions: succeeded (V) and failed (X).

[4] https://www.mapbox.com/
[5] https://www.opentopodata.org/
[6] https://github.com/3DOM-FBK/deep-image-matching

4.1 Results and evaluation

The proposed 3DGeoRef framework is evaluated on a heterogeneous dataset of well-known landmarks (Figure 1, Table 3) to assess its capability to automatically georeference 3D models lacking positional and scale information. The experiments compared the considered AI-based localization solutions - GEMINI, GEOCLIP and OLLAMA - for the rough localization of the models and evaluated the performance of the successive DIM library in refining the georeferencing through the matching of synthetic views with corresponding satellite imagery. The overall results for each dataset and VLM method are presented in Table 3, 4 and 5. Figure 5 reports some examples of georeferenced 3D models visualized within Cesium[7].

5. Ablation study

In order to assess the robustness of the proposed AI-based georeferencing method, two ablation studies are presented.
The first aims to evaluate the stability and repeatability of the VLM/MLLM methods by executing 10 independent runs on the same datasets and analyzing the variability of the estimated geographic coordinates. An example of the results, representative of the behaviour observed across most of the datasets, is shown in Figure 6. The results highlight different levels of variance across the evaluated models, with OLLAMA exhibiting the highest instability, while GeoCLIP and Gemini reporting more stable and consistent behaviour across runs.

	Approximate georeferencing			Refined georeferencing			
	Gemini	DIM	Distance	Map resol. [px/m]	dX [m]	dY [m]	RMSE [m]
Neuschwanstein Castle	Success	Success	0.02 km	0.6	1.4	0.4	1.5
Milan Cathedral	Success	Success	0.09 km	0.6	-0.3	-0.2	0.5
Parthenon	Success	Success	0.37 km	1.2	-5.7	-5.4	8.8
Trento Cathedral	Success	Success	0.03 km	0.6	-2.1	-1.3	3.4
Sagrada Familia	Success	Success	0.02 km	0.6	-0.5	-0.6	1.2
Torre Eiffel	Success	Success	0.01 km	0.6	-0.8	-5.2	2.9
Odawara Castle	Failed	-	605.3 km	-	-	-	-
Rovereto Castle	Failed	-	243.6 km	-	-	-	-
Loz Feliz, California	Failed	-	29.1 km	-	-	-	-
Dinkins Manhattan Palace	Success	Success	0.09 km	0.5	-1.1	-0.8	1.5
Mont-Saint-Michel Abbey	Success	Success	0.01 km	0.6	13.4	1.1	13.5
Vaticano	Success	Success	0.36 km	0.6	5.2	7.6	26.4
Statue of Liberty	Success	Success	0.01 km	0.6	0.7	2.1	2.6
Pantheon	Success	Success	0.01 km	0.6	1.1	-0.3	10.8

Table 4: Georeferencing performances (approximate and refined steps) for the proposed methodology based on Gemini.

	Approximate georeferencing			Refined georeferencing			
	GEOCLIP	DIM	Distance	Map resol. [px/m]	dX [m]	dY [m]	RMSE [m]
Neuschwanstein Castle	Success	Success	0.3 km	0.6	-4.0	-1.7	4.2
Milan Cathedral	Failed	-	340.8 km	-	-	-	-
Parthenon	Success	Success	0.3 km	1.2	-1.7	-6.5	7.9
Trento Cathedral	Failed	-	343.5 km	-	-	-	-
Sagrada Familia	Success	Success	0.04 km	0.6	-0.7	-1.6	2.7
Torre Eiffel	Success	Success	0.2 km	0.6	-0.4	-1.5	2.6
Odawara Castle	Failed	-	408.8 km	-	-	-	-
Rovereto Castle	Failed	-	787.7 km	-	-	-	-
Loz Feliz, California	Failed	-	521.3 km	-	-	-	-
Dinkins Manhattan Palace	Success	Success	0.3 km	0.5	-0.9	-0.6	1.4
Mont-Saint-Michel Abbey	Success	Success	0.03 km	0.6	3.9	2.6	13.3
Vaticano	Success	Success	0.2 km	0.6	5.7	1.3	20.1
Statue of Liberty	Success	Success	0.07 km	0.6	-0.1	0-5	3.7
Pantheon	Failed	-	882.8 km	-	-	-	-

Table 5: Georeferencing performances (approximate and refined steps) for the proposed methodology based on GEOCLIP.

	Approximate georeferencing			Refined georeferencing			
	OLLAMA	DIM	Distance	Map resol. [px/m]	dX [m]	dY [m]	RMSE [m]
Neuschwanstein Castle	Success	Success	0.02 km	0.6	-3.0	-3.5	8.2
Milan Cathedral	Success	Success	0.09 km	0.6	-0.8	-1.2	2.2
Parthenon	Success	Success	0.01 km	1.2	-9.8	-13.9	15.2
Trento Cathedral	Failed	-	462.9 km	-	-	-	-
Sagrada Familia	Success	Success	0.01 km	0.6	-0.4	-0.9	1.5
Torre Eiffel	Success	Success	0.01 km	0.6	-1.0	-2.9	3.1
Odawara Castle	Failed	-	408.5 km	-	-	-	-
Rovereto Castle	Failed	-	457.3 km	-	-	-	-
Loz Feliz, California	Failed	-	509.1 km	-	-	-	-
Dinkins Manhattan Palace	Success	Success	0.04 km	0.5	-0.6	-2.4	2.0
Mont-Saint-Michel Abbey	Success	Success	0.01 km	0.6	4.8	2.1	0.6
Vaticano	Success	Success	0.03 km	0.6	5.7	6.1	27.1
Statue of Liberty	Success	Success	0.01 km	0.6	-1.0	-0.1	3.1
Pantheon	Success	Success	0.08 km	0.6	1.5	-0.5	11.4

Table 6: Georeferencing performances (approximate and refined steps) for the proposed methodology based on OLLAMA.

[7] https://cesium.com/

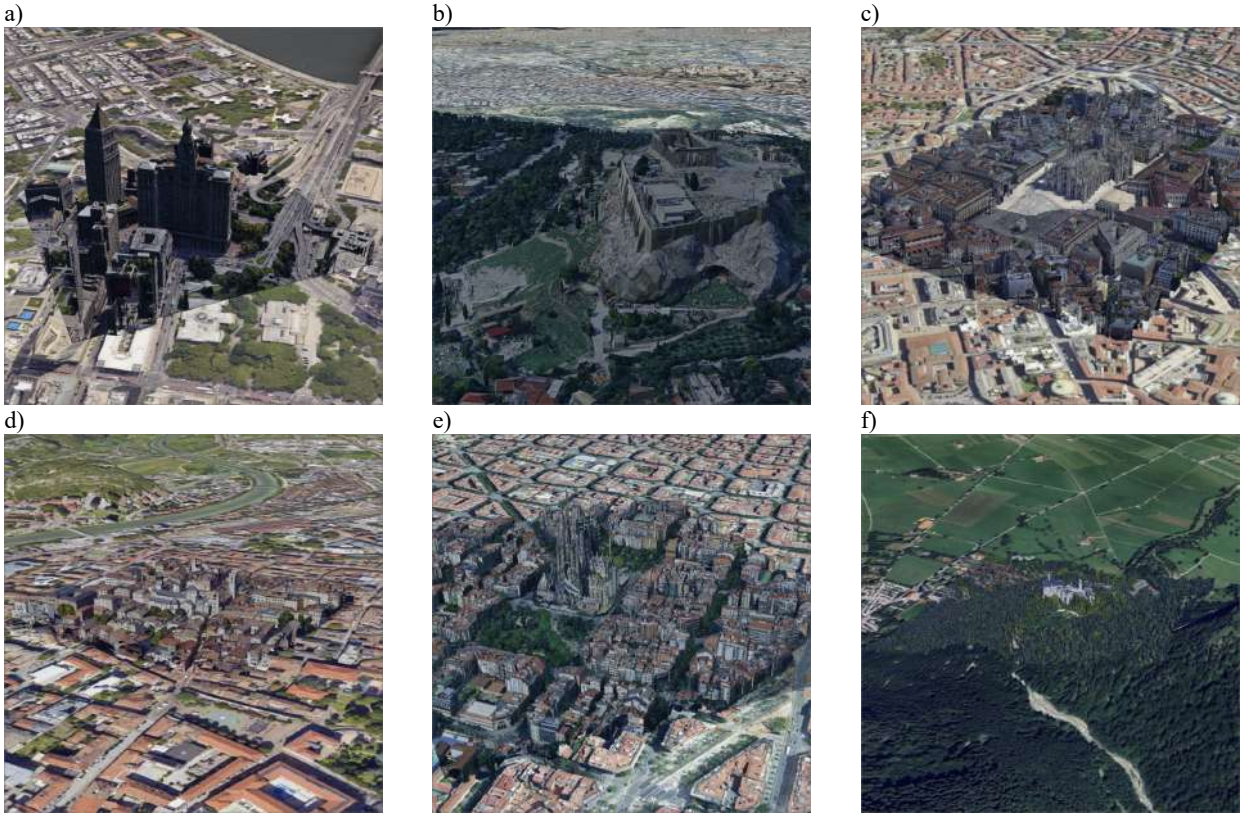

Figure 5: Examples of georeferenced 3D models obtained with 3DGeoRef and visualized within Cesium: a) Dinkins Palace; b) P arthenon; c) Milan Cathedral; d) Trento Cathedral; e) Sagrada Familia; f) Neuschwanstein Castle.

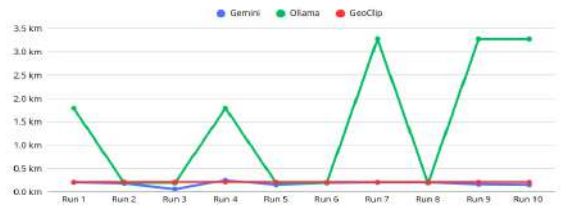

Figure 6: Distance results variation for the different VLM-LLMMs with 10 different runs for the Vatican City dataset.

Neuschwanstein Castle	✗	Milan Cathedral	✓
Parthenon	✗	Trento Cathedral	✗
Sagrada Familia	✓	Torre Eiffel	✓
Odawara Castle	✗	Rovereto Castle	✗
Loz Feliz, California	✗	Dinkins Manhattan Palace	✓
Mont-Saint-Michel	✓	Vatican City	✓
Statue of Liberty	✓	Pantheon	✓

Table 7: Gemini geolocalization performance focusing on target objects and with limited contextual data.

Moreover, when considering the full set of evaluated datasets, GeoCLIP shows lower variability but also reduced localization accuracy compared to Gemini.

The second ablation study investigates the impact of the surrounding visual context on the performance of the proposed geolocalization procedure. Experiments (reported in Table 7) are repeated considering a reduced amount of contextual information and focusing on the target objects. The results demonstrates that the tool is still able to identify plausible location in many cases, but the failure cases indicate that the robustness of the pipeline is considerably dependent on the availability of sufficient surrounding context. The datasets correctly identified by the VLM/MLLM model, when reducing the object's contexts, are shown in Figure 7.

6. Conclusions

The work presented an AI-based pipeline to automatically georeference a 3D model stored in a web repository but missing georeferencing metadata. The reported results show promising capability of automatically identifying and assigning spatial information to not-georeferenced 3D heritage models available in public repositories. This is an important pre-requisite for reusing and ingesting enriched 3D assets, e.g., into the European Data Space for Cultural Heritage. The performance of the proposed AI-based procedure heavily relies on a substantial amount of surrounding data to allow a good guess of the VLM/MLLM process and the image-matching stage.

Future developments include a preliminary coarse localization of the render on the orthophoto using dense DINO descriptors, which can constrain or replace the DIM stage in challenging cases.The developed procedure is available for research purposes at https://github.com/3DOM-FBK/3DGeoRef.

Acknowledgments

The presented research activities are developed as part of the EU project 3DBigDataSpace project GA 101173385, funded under

the Digital Europe Programme (DIGITAL), Call DIGITAL-2022-CULTURAL-02-HERITAGE.

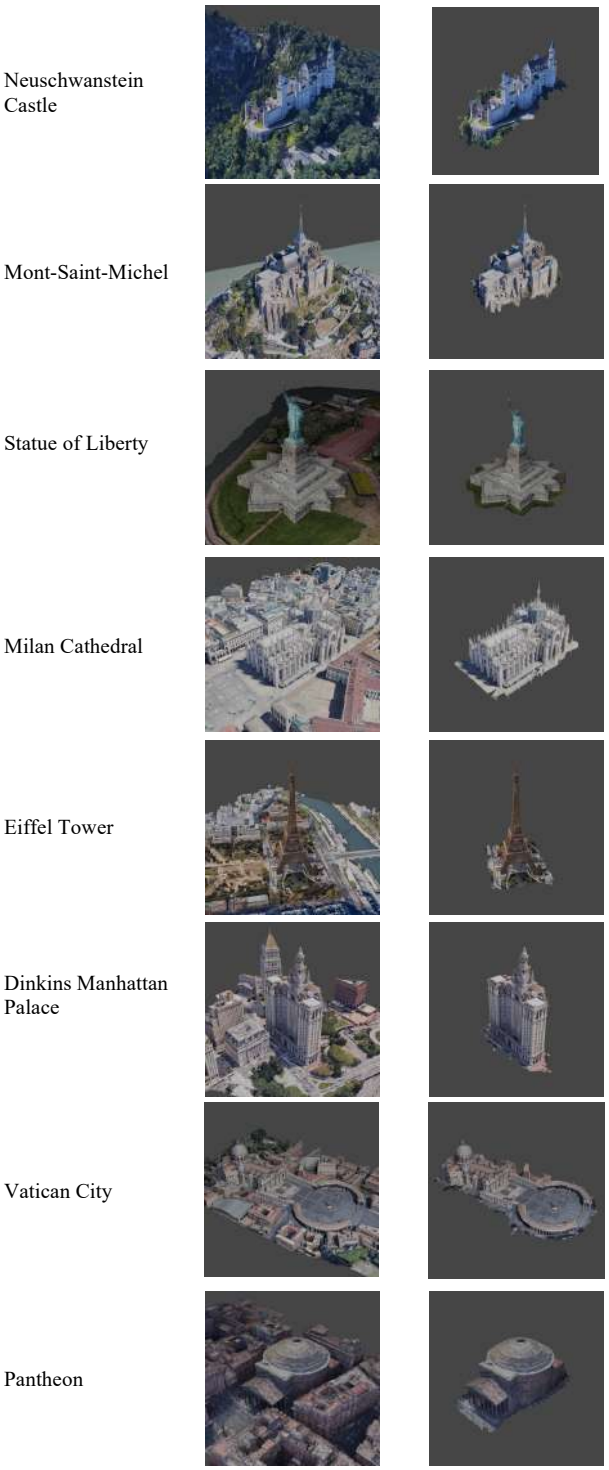

Figure 7: Examples of datasets correctly georeferenced using the proposed approach, both using the full and a reduced amount of contextual surrounding information.

References

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., 2016. NetVLAD: CNN architecture for weakly supervised place recognition. Proc. *CVPR*.

Azari, P., Li, S., Shaker, A. and Sattar, S., 2025. Georeferencing Building Information Models for BIM/GIS integration: a review of methods and tools. *ISPRS International Journal of Geo-Information*, 14(5), p.180.

Cai, Z, Wang, R., Gu, C., Pu, F., Xu, J., et al., 2025. Scaling Spatial Intelligence with Multimodal Foundation Models. *arXiv:2511.13719*.

CNECT/LUX/2021/OP/0070 - Deployment of a common European data space for cultural heritage: *PM. Annual report M12. Europeana Foundation*. September 2023. URL: https://pro.europeana.eu/page/data-space-deployment. CC BY-SA.

Craciun, D. and Le Bris, A., 2022. Automatic algorithm for georeferencing historical-to-nowadays aerial images acquired in natural environments. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., 43*, pp.21-28.

Danish, S., Sadeghi-Niaraki, A., Khan, S.U., Dang, L.M., Tightiz, L., Moon, H., 2026. A comprehensive survey of Vision–Language Models: Pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets. *Information Fusion*, Vol. 126, 103623.

DeTone, D., Malisiewicz, T. and Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. Proc. *CVPR*, pp. 224-236.

Farella, E.M., Morelli, L., Remondino, F., Mills, J.P., Haala, N. and Crompvoets, J., 2022. The EuroSDR TIME benchmark for historical aerial images. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci, 43*, pp.1175-1182.

Guan, T., Muthuselvam, a., Hoover, M., Wang, X., Liang, J., Sathyamoorthy, A.J., Conover, D., Manocha, D., 2023. CROSSLOC3D: Aerial-Ground cross-source 3D place recognition. Proc. *ICCV*, 11335-11344.

Hays, J., Efros, A.A., 2008. IM2GPS: Estimating geographic information from a single image. Proc. CVPR.

Lindenberger, P., Sarlin, P.E., Hosang, J., Balice, M., Pollefeyes, M., Lynen, S., Trulls, E., 2025. Scaling image geo-localization to continent level. Proc. *NeurIPS.*

Liu, Z., Du, Y., Fu, T., Su S., Ho, C., Wang, C., 2025. Vision-Language Memory for Spatial Reasoning. *arXiv:2511.20644*.

Loeper, Y., Gerke, M., Alamouri, A., Kern, A., Bajauri, M. S., Fanta-Jende, P., 2024 Visual localization in urban environments employing 3D city models. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-2/W8-2024, 311–318.

Luft, J., 2020. Automatic georeferencing of historical maps by geocoding. *Automatic vectorisation of historical maps, 13*, p.75.

Maiwald, F., Feurer, D. and Eltner, A., 2023. Solving photogrammetric cold cases using AI-based image matching: New potential for monitoring the past with historical aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing, 206*, pp.184-200.

Maiwald, F., Lehmann, C. and Lazariv, T., 2021. Fully automated pose estimation of historical images in the context of

4D geographic information systems utilizing machine learning methods. *ISPRS Int. Journal of Geo-Information, 10*(11), p.748.

Malek, S., Farella, E.M., Perda, G., Remondino, F., Cantoro, G., 2025. Towards automatic vector extraction from scanned historical aerial photo indexes. *Photogrammetric Engineering and Remote Sensing*, in press.

Marcondes, F.S., Gala, A., Magalhães, R., Perez de Britto, F., Durães, D. and Novais, P., 2025. Using OLLAMA. In: *Natural Language Analytics with Generative Large-Language Models.* Springer Briefs in Computer Science, pp. 23-35.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J., 2025. Large Language Models: a survey. *arXiv:2402.06196v3*.

Morelli, L., Bellavia, F., Menna, F. and Remondino, F., 2022. Photogrammetry now and then–from hand-crafted to deep-learning tie points. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLVIII-2/W1-2022, pp.163-170.

Morelli, L., Ioli, F., Maiwald, F., Mazzacca, G., Menna, F. and Remondino, F., 2024. Deep-image-matching: a toolbox for multiview image matching of complex scenarios. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLVIII-2/W4-2024, pp. 309-316.

Milleville, K., Verstockt, S. and Van de Weghe, N., 2022. Automatic georeferencing of topographic raster maps. *ISPRS International Journal of Geo-Information, 11*(7), p.387.

Münster, S., 2023. Advancements in 3D heritage data aggregation and enrichment in Europe: implications for designing the Jena Experimental Repository for the DFG 3D viewer. *Applied Sciences, 13*(17), p.9781.

Münster, S., Bruschke, J., Rajan, V., Komorowicz, D., Preßler, R. and Ukolov, D., 2025. 4D world viewers as multi-user content management systems. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., 48*, pp.1043-1050.

Panek, V., Kukelova, Z., Sattler, T., 2023. Visual localization using imperfect 3d models from the Internet. Proc. *CVPR*.

Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M., 2019. From coarse to fine: robust hierarchical localization at large scale. Proc. *CVPR*, 12716-12275.

Sarlin, P.E., DeTone, D., Malisiewicz, T. and Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. Proc. *CVPR*, pp. 4938-4947.

Shaji, R., Bouveyron, T. and Willmes, C., 2025. 3DTiler: A tool to georeference 3D models and generate 3D tiles. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., 48*, pp.157-163.

Sun, J., Shen, Z., Wang, Y., Bao, H. and Zhou, X., 2021. LoFTR: Detector-free local feature matching with transformers. In Proc. *CVPR*, pp. 8922-8931.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., Silver, D., 2023. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*.

Tombari, F., Remondino, F., 2013. Feature-based automatic 3D registration for cultural heritage applications. Proc. *IEEE Conference Digital Heritage*, Vol. 1, pp. 55-62.

Tomesek, J., Cadık, M., Brejcha, J., 2022. CrossLocate: Cross-modal large-scale visual geo-localization in natural environments using rendered modalities. Proc. *WACV*, pp. 2193-2202

Vivanco Cepeda, V., Nayak, G.K., Shah, M., 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems, 36*, pp.8690-8701.

Xu, C., Yu, F., Bianco, M.J., Kovarskiy, J., et al. 2025. Geo-R1: Unlocking VLM Geospatial Reasoning with Cross-View Reinforcement Learning. *arXiv:2510.00072.*

Ye, J., Lin, H., Ou, L., Chen, D., Wang, Z., Zhu, Q., He, C., Li, W. 2025. Where am I? Cross-View geo-localization with natural language descriptions. Proc. *ICCV*, pp. 5890-5900.

Weyand, T., Kostrikov, I., Philbin, J., 2016. PlaNet - Photo geolocation with convolutional neural networks. Proc. *ECCV*.

Zhang, L., Rupnik, E., Pierrot-Deseilligny, M., 2021. Feature matching for multi-epoch historical aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 182, pp. 176-189.

Zhou, Z., Zhang, J., Guan, Z., Hu, M., Lao, N., Mu, L., Li, S., Gai, G., 2024. Img2Loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. Proc. *SIGIR*, pp. 2749-2754.