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

4D World Viewers as Multi-user Content Management Systems

Sander Münster[1], Jonas Bruschke[1], Vaibhav Rajan[1], Dávid Komorowicz[1], Rebecca Preßler[1], Dominik Ukolov[1]

[1] Digital Humanities, Friedrich-Schiller-Universität Jena, Germany - sander.muenster@uni-jena.de

Keywords: Digital History, Cultural Heritage, Digital 3D Modelling, Image Matching, Virtual City Tours.

Abstract

Since 2016, our group has developed a modular 4D world web-visualisation software framework for both mobile and desktop devices. A major challenge is to obtain and compile 4D world models at large scale. This article proposes to (a) highlight our data retrieval pipeline and (b) geolocalisation as both rough and fine positioning as well as (c) 3D data processing to blend 3D assets with digital elevation models for seamless 3D environments. Main results include a validation of different approaches for rough and fine level geolocalisation by using Large Language and Vision Language Models as well as key point matchers and a preprocessing of 3D assets to generate 3D environments. The article shows results from the validation of these pipelines with ground-truth material taken from our datasets of 75,000 3D mesh models and 1.5 M images.

1. Introduction

Since 2016, our group has developed a modular 4D web-visualisation software framework for 4D city and 4D content browsing to test and validate design hypotheses in virtual, augmented, and 2.5D visualisation on mobile and desktop devices (Münster, 2018). As user acceptance of native applications is decreasing (Bender, 2020), especially for specific and short-term use, as relevant for most cityscape scenarios (Reips, 2002), this is a browser-based web application.

Multiple strategies are applied to gather, process, and serve different types of data. On the one hand, users can contribute and upload content manually. On the other hand, data should also be queried and processed from various data sources automatically as much as possible to in order to create a virtual world that the application is to visualise. In former articles we highlighted the prospected research agenda (Münster et al., 2020) as well as technological venues (Münster et al., 2021a,b, 2024c), but also conceptual challenges to automatically reconstruct the past (Münster et al., 2022, 2024a).

This article proposes to (a) highlight our data retrieval pipeline and (b) geolocalisation as both rough and fine positioning as well as (c) 3D data processing to blend 3D assets with digital elevation models for seamless 3D environments.

2. State of the Art

During the past two decades, numerous digital image archives containing vast numbers of photographs have been set up (Münster et al., 2018; Capurro et al., 2024). This comprises collections of user-generated contemporary photographs, but also historic photo collections and image bases with geographic coverage as, e.g., Google Street View. For the 3D worlds, large-scale datasets such as Objaverse including 10.2 million 3D models (Deitke et al., 2023) or ShapeNet including 50k 3D models (Chang et al., 2015) and repositories such as Sketchfab hosting several 100k heritage items (Flynn, 2022) have been amassed. As an overlapping area, there are several automated 3D model creation processes that utilise existent imagery (Snavely et al., 2007; Stathopoulou et al., 2019; Wu et al., 2021; Maiwald et al., 2023b) with large-scale 2D/3D datasets compiled such as MVImgNet2.0 (Wu et al., 2024) and MegaScenes (Tung et al., 2025). Despite various approaches, a still open major task is the provision of sufficient metadata to spatialise and temporalise this material (Münster, 2023). Varied approaches are used for AI-based metadata enrichment (e.g. Orzechowski et al., 2025), but also for creating multimodal 3D representations (e.g. Rusnak and Kaplan, 2025).

Geo-based data is visualised in various web-based portals, increasingly using a 3D approach. Schindler and Dellaert (2012) were among the first to make historical photographs accessible on the web in relation to 3D models. Other applications introduced time-evolving 3D city models incorporating multi-media such as photographs and other historical documents (Blettery et al., 2020; Jaillot et al., 2021), or augmented realities on mobile devices (Hasselman et al., 2023). Though, precise spatial and temporal location of the data remains challenging, requiring either a rich set of metadata or manual work involved. With regards to spatialisation of photographs, there are several applications to support this task by either clicking corresponding points (Blanc et al., 2018) or using a rephotography approach (Schaffland et al., 2020). However, both approaches still require manual input. Other approaches include the geolocalisation via images (Pramanick et al., 2022; Vivanco Cepeda et al., 2023; Wang et al., 2024; Kulkarni et al., 2024; Xu et al., 2024) and texts (e.g. Singh and Aneja, 2024).

Figure 1. Browser-based mobile application showing textured 3D models of historical buildings and points of interest.

3. Multi-user interfaces

The 4D visualisation framework consists of two main applications: 1) The 4Dcity app is a mobile application covering cultural tourism and education (Figure 1). Used in situ, the user can explore how their surrounding might have looked in the past. 2) the 4D Browser targets scholars interested in architectural history and urban development. The user browses historical images on a city-scale (Figure 2). In both applications, the user can use a time slider to filter the data that also affects the visualisation.

Figure 2. Graphical user interface of the 4D Browser application.

The applications share the same backend and database that has been fed with data from different repositories in the past. However, there are many historical photographs that cannot be found in those big repositories, as they are held by local residents or small-town archives. To exploit this potential, several user contests have been organised to gather additional, hitherto unseen photographs. In this regard, the applications have been enhanced. Users can upload historical images, but also rephotograph existing ones. To this end, user accounts have been introduced, including a dashboard where the users can check their contributions. A moderator validates all uploaded content in order to make it available for everyone.

While the contribution by individual users is a great chance to acquire historical data that otherwise would remain hidden, the amount of data is rather marginal and only relevant on a small scale. To this end, an automatic ingestion and processing of publicly available data is required in order to present appropriate datasets on a larger scale.

4. Data Retrieval and Enrichment

4.1 3D Data

The initial dataset utilised in 3DBigDataSpace stems from different data collections and was compiled between 11/2023 and 04/2025 (Table 1). To retrieve legally accessible content, we selected CC-0 or CC-BY licensed content only. For data retrieval, we used a series of server-side scripts in Python and PHP feeding into an SQL database and Unix file storage.

4.2 Images Data

Various large-scale data resources are available nowadays. The Google Landmark v2 dataset from 2020 contains 4.7 M images of landmarks (Weyand et al., 2020). For the Landmark dataset, we translated each location into coordinates by (a) requesting Wikimedia coordinates, and in case of no results (b) retrieving OSM coordinates or (c) Google Place coordinates via geopy, and selected those landmarks with keywords referring to built structures. Europeana holds 1.3 M images tagged with “building” for which we retrieved images with provided location coordinates. Other data sources include the already compiled set of images in the 4D Browser and collections made available by Fortepan (CC-BY-SA-3.0) and Pol Meyert – a Belgian photographer (Table 2).

Data source	No.	Description
Europeana	8,708	The Europeana 3D dataset contains validated metadata and is utilised to provide ground truth data. The metadata retrieval has been conducted via the Europeana Python framework.[1]
Objaverse 1.0	55,614	The Objaverse 1.0 dataset includes 800,000 3D objects with those selected by us which are classified as Cultural Heritage (Deitke et al., 2023). It has been compiled by the Paul Allen institute. The datasets are mainly retrieved from open-licensed content held by SketchFab.
5DCulture	8,406	The 5DCulture dataset was compiled in the eponymous project in 2024. The dataset includes various mainly low-poly models of single buildings from different age in the cities of Trento, Sion, Amsterdam, Dresden and Jena. The dataset was used to test the Zenodo pipeline.
Objaverse XL-Smithsonian	2,407	A set of models from the Smithsonian museum is included in the ObjaverseXL dataset of 10.2 M 3D meshes.
LiDAR 3D Buildings	1,374	LiDAR 3D buildings were segmented via OSM ground plots into single building models. These models are used for mapping and multi-LOD approaches.

Table 1. 3D data sources.

4.3 Data Storage

For data storage, the Objaverse 1.0 Cultural Heritage dataset and the 5DCulture dataset were ingested in Zenodo via the Zenodo Toolbox (Münster et al., 2024b).[2] To ensure long-term availability and citability, building models and image sources have been uploaded to Zenodo, which are accompanied by Europeana Data Model (EDM) and METS/MODS XMLs (Figure 3). In processing the visual data, particular attention has been paid to privacy concerns. A two-stage process involving person detection and detailed segmentation has been implemented to mask individuals while preserving as much architectural visual material as possible.

Thumbnails of the 3D models were rendered from five perspectives and in various resolutions using an automated pipeline to address different immediate use cases. To manage this amount of data and the different versions, a database is updated in real-time with each upload or update process.

[1] https://github.com/europeana/rd-europeana-python-api
[2] https://github.com/Digital-Humanities-Jena/zenodo-toolbox

Data source	No.	Description
Wikimedia Commons	124,896	A subset of the Google Landmark v2 dataset from 2020 containing 4.7 M images of landmarks of which 980,000 images have been georeferenced and 125,000 refer to built structures.
Europeana Buildings	1,318,086	Photographs tagged as built structures in the Europeana.
Fortepan	18,816	The Fortepan dataset was compiled from various data sources from 1860 till 2008.
4D Browser dataset	34,486	This is an already integrated collection of positioned and oriented historical and contemporary city photographs.
Architectural photo dataset	75,000	The Belgian photographer Pol Mayert donated a set of architectural photographs which were processed via the Zenodo pipeline.

Table 2. 2D data sources.

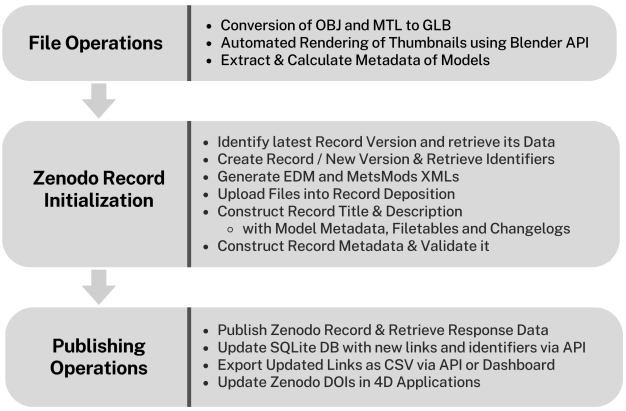

Figure 3. General process flow for using the Zenodo API.

This database is accompanied by a dashboard and an API, enabling the retrieval of the latest direct links to the model or image sources on Zenodo in addition to contextual information like data types, previous version DOIs or dataset identifiers. Furthermore, the latest DOIs are stored in the object data of the 4D applications and will therefore be available through its API.

Each Zenodo record contains extracted and aggregated metadata, tables with direct links to main files and thumbnails as well as changelogs. Using the dashboard, directly exporting CSVs of query results allow for direct implementations of the data hosted on Zenodo without any additional queries. While retrievals from Zenodo are limited, the current rate is sufficient for positional field-of-view applications in real-time. This approach not only secures the data, but also facilitates its use in scientific, educational and creative contexts.

5. Geolocalisation

For retrieving a geolocalisation from descriptions, we benchmarked several LLMs with the Europeana dataset which has already approved coordinates. A specific comparison included (a) Spacy with a large-scale English language model (en_core_web_lg) as transformer-based model, (b) Llama 3.2-3B and (c) DeepSeek R1 Distill Qwen 1.5B as lightweight LLMs and (d) DeepSeek R-1 as full-scale LLM. The best results were achieved by DeepSeek-R1 with 82 % recognition rate for countries and 46 % for cities (Figure 4). An important finding was that the other models were significantly less accurate with only 30 % correct countries identified by Spacy.

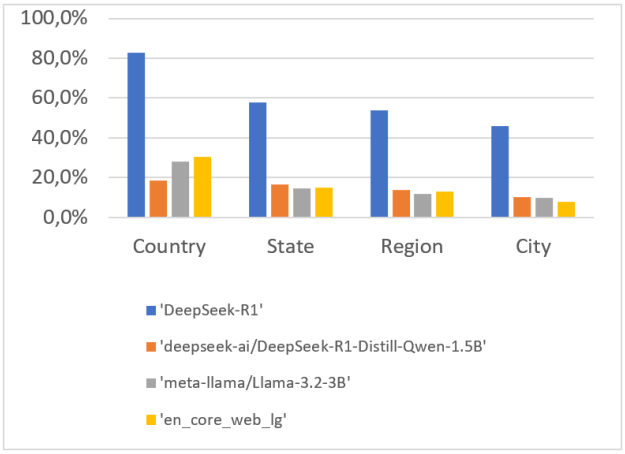

Figure 4. Matching between text-based model retrieved and human-assigned location information from the Europeana dataset ($n = 2,465$).

Another step involves identifying the specific object shown in the 3D model. To accomplish this, we render the model into a series of images and employ a content-based image retrieval (CBIR) script to identify similar images. We benchmarked three geolocalisation frameworks with the Europeana 3D renderings. Those are PLONK (Dufour et al., 2024), GeoCLIP (Vivanco Cepeda et al., 2023) and OrienterNet (Sarlin et al., 2023).

With regards to its performance GeoCLIP and PLONK as image-only approaches performed not well with less than 5 % retrieval rate at country level (Figure 5). OrienterNet uses given textual location information – in our case taken from the Deepseek R-1 results from the text-based geolocalisation. Not surprisingly, the retrieval rate is comparably high, although containing a high level of locations not retrieved.

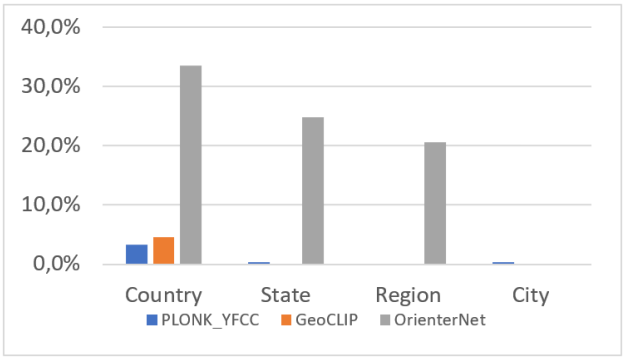

Figure 5. Matching between image-based model retrieved and human-assigned location information from the Europeana dataset ($n = 242$).

We tested the recognition of places for historical images of Dresden from the 4D Browser dataset. In that case, PLONK

reached 60.4 % retrieval rate at city level, with GeoCLIP reaching 41.5 %. OrienterNet did also successfully retrieve both city and borough information at 55.1 % with city name given (Figure 6).

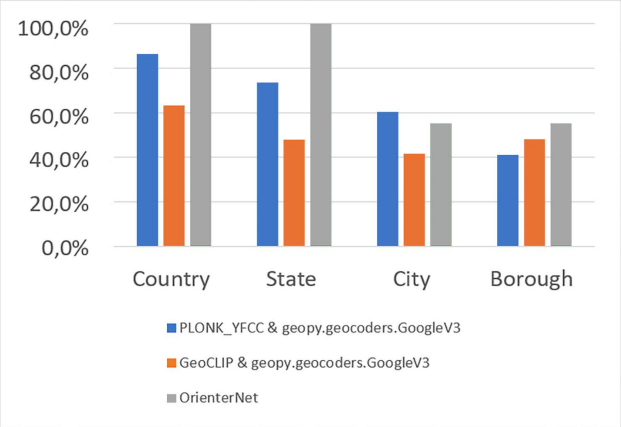

Figure 6. Matching between historical images and human-assigned location information for data from the 4D Browser dataset ($n = 1.168$).

For selecting images of architectural exteriors, we use a VGG-16-based classifier. The classifier was extended from a previous version (Münster et al., 2024c) and now trained with 6,830 manually oriented photographs and other images. The classified files belonging to two classes with 2,715 images showing architectural exteriors and 4,115 images not showing architectural exterior. For training, 5,464 files were used – including nine variants by data augmentation per file – and for validation we used 1,366 files.

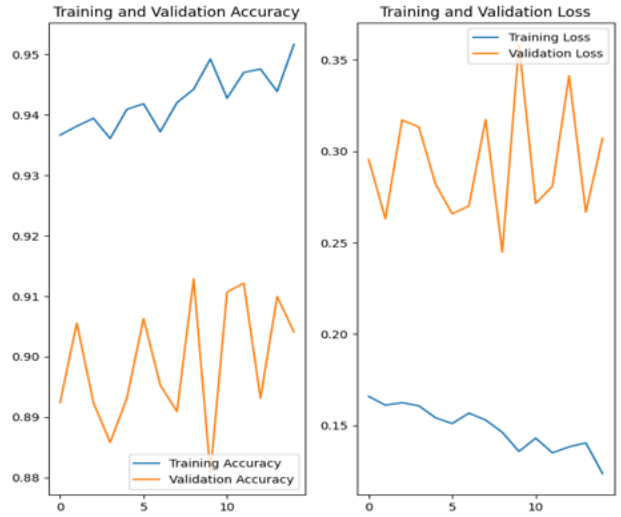

Figure 7. VGG-16-based training and validation accuracy and loss for classification of architectural exteriors/others.

The accuracy of the validation dataset is above 0.90 with loss of 0.30 (Figure 7). The classifier is used to automate the step to identify architectural exteriors as a prerequisite for further processing in the pipeline.

To gain a good ratio between true and false positives as well as negatives, we currently only judge on an image if the predictor certainty is 90 % or higher. For validation, we tested the classifier with a subset of the Dresden images from the 4D Browser dataset which exclusively contains exteriors. Correct detection rate is 86.2 % with 12.2 % false negatives and 1.6 % with predictor certainty below threshold (Table 3).

Accepted	506
Declined	72

Table 3. Validation with 4D Browser dataset ($n = 587$).

6. Large-scale Image Pose Estimation

In order to display historical images within the 3D/4D environment, their poses need to be estimated. Feature extraction and image matching for pose estimation has been possible for quite some time, even for a very large photographic dataset on a city scale Agarwal et al. (2011). The vast majority of approaches work with contemporary images in this regard. For historical images, this however proves highly challenging. Existing approaches either needed a lot of manual adjustments or were only applicable to a small set of images on a much smaller scale Maiwald et al. (2023a). However, recent advances in image matching approaches bring us closer to the goal of spatialising historical images on a city scale.

Our proposed pipeline combines the repeatability of DISK keypoints (Tyszkiewicz et al., 2020) and the pairwise matching performance of the MASt3R model (Leroy et al., 2024). We extract all available DISK keypoints without limitation, which helps with cases where the limited number of points do not cover the whole image. MASt3R extracts a dense point map representation, which results in significantly more matches. We keep 10k matches on a coarse grid. We use the DISK keypoints as anchor points to make it manageable for Structure from Motion (SfM). Moreover, we filter the matching image pairs based on a high inlier match number (1000). This way, we obtain models with relatively few false positive matches. A downside of such a large number of keypoint matches is that SfM becomes very slow for $>$ 1000 images. To alleviate this problem, we partition the city model into $N = 30$ clusters. Within these clusters, we use an ensemble of image content-based global retrieval methods – AnyLoc (Keetha et al., 2023) and MegaLoc (Berton and Masone, 2025). We use the available approximate GPS coordinates for clustering.

The Fortepan photo archive contains over 200,000 photos across many countries and cities. The collection includes indoor, outdoor, portrait, group photo and document type photos, among others. Out of these, only the outdoor images are usable for cityscape reconstruction, with little distractors covering the buildings. Using image classification is a straightforward way to filter out the unusable classes, thus reducing processing time. In practice, however, the boundary between these classes is blurry. Most in-the-wild photos contain people or vehicles as the main object and the buildings to be reconstructed are considered the background. That means that it’s difficult to distinguish between portraits and or group photos with a blank background and in-the-wild photos with a usable background.

Another option is to apply semantic segmentation on the dynamic classes (people, vehicles, vegetation, sky, etc.) and calculate the ratio of these pixels. We find that filtering based on this ratio would discard images that depict crowded scenes but contain strong geometric information. Instead, we use these semantic masks to remove the feature points from dynamic objects as seen in in Figure 8.

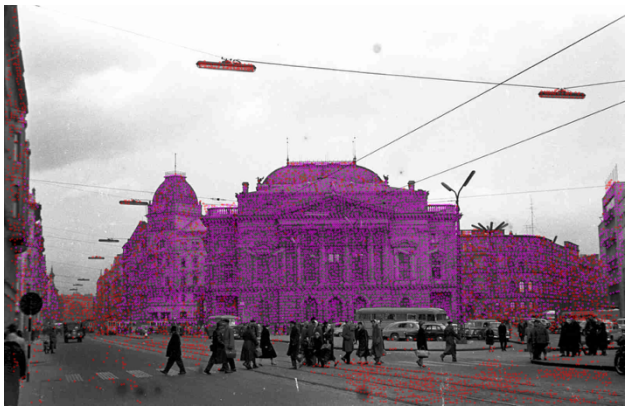

Figure 8. Masked keypoints: purple points correspond to 3D points in scene, red keypoints are unmatched 2D points. Dynamic objects do not have keypoints.

The Fortepan photo archive contains a combination of human annotated place names, rough GPS coordinates and image description. Often the precise street/square address is written in the description without the corresponding GPS metadata. We use a Large Language Model (LLM) to parse location information from the description if it is available and use a geocoding service to turn them into GPS coordinates. Images without any metadata can be geolocalised at different granularity based on Section 5.

We use OrienterNet (Sarlin et al., 2023) to further refine the positions. This achieves two goals: Firstly, refine location information from the (incomplete) street level to a more fine-grained location. Secondly, perturb the GPS coordinates that are at the exact same coordinates, often caused by using geocoding services. This helps with the nearest neighbour query. Finally, we keep images of Budapest in a 6 km radius from the city centre to include only the densely (photographed), older part of the city. This results in 41,671 images.

We use KMeans clustering to divide the images into $N = 30$ clusters. The resulting clusters can be seen in Figure 9. The clusters have approximately the same geometrical size which means that the number of images in each cluster varies depending on the density of the photos.

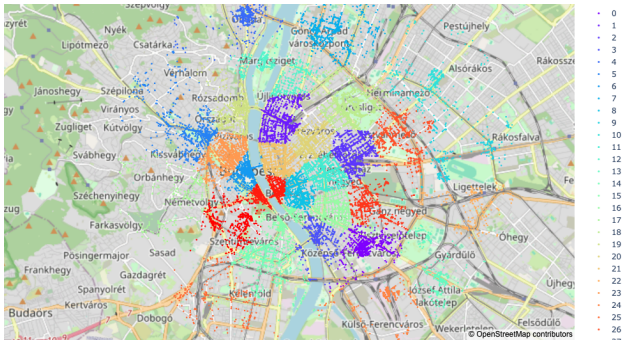

Figure 9. Location and distribution of the 30 clusters. Note that the refined GPS coordinates align well to the streets.

In each cluster we query the top 50 most similar images within the cluster using both models. We use a threshold t_{anyloc} and $t_{megaloc}$ to discard pairs with a low similarity score. We always keep at least 10 pairs per image regardless of the threshold to account for hard to match cases which is common in the historic setting. We obtain the thresholds by calibrating the retrieval models to maximise the true positive rate while minimizing the false positive rate using a RoC curve. We consider an image pair a true positive if their distance is less than $d = 150\text{m}$. We calculate the thresholds on both the Budapest subset as well as the whole dataset and achieve similar thresholds. This means that the retrieval methods work similarly at city and world scale. The obtained thresholds are $t_{anyloc} = 0.40$ and $t_{megaloc} = 0.25$.

Doppelgangers are still a challenge in the historical domain. The inherent ambiguity of appearance/temporal changes do not necessarily mean a different place. Or inversely, different looking places are not necessarily different. Most notable examples are bridges, symmetric churches and large hotel buildings with similar looking sides. Partitioning the dataset into clusters also helps with this problem. As opposed to having a single large scene with multiple incorrect matches with doppelgangers, the result is a set of smaller components with fewer false positive matches (doppelgangers). It does not accumulate, making it easier to discover and separate.

We compare the proposed pipeline to the baseline method using 10k DISK feature points and the LightGlue (Lindenberger et al., 2023) matcher by listing the number of reconstructed images, number of models and number of images in the largest model for the top 5 clusters in Table 4. Figure 10 shows the qualitative comparison of the largest scene from cluster 12.

Cluster Id	6	17	9	28	12
DISK + LightGlue (Baseline)					
reconstructed models	11	11	3	7	4
reconstructed images	733	770	599	447	471
images in biggest model	310	227	531	154	376
DISK + MASt3R (Ours)					
reconstructed models	20	13	6	17	9
reconstructed images	1406	1356	968	979	904
images in biggest model	548	764	800	316	688

Table 4. Statistics of the 5 biggest clusters ignoring models with fewer than 20 images.

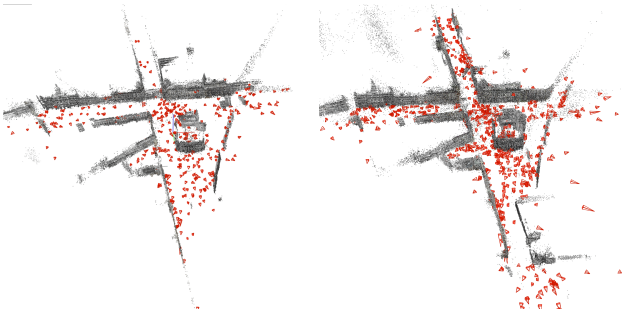

Figure 10. Automatic reconstruction of cluster 12. Left: DISK + LightGlue (Baseline). Right: DISK + MASt3R (Ours).

In conclusion, our proposed pipeline manages to reconstruct on average twice as many images from the dataset, compared to the baseline, process them faster by filtering the image pair candidates. The accuracy is slightly lower which can be easily improved. It is more robust to doppelgangers and avoids matching dynamic objects at different places.

7. 4D World Model Generation

Within the described 4D applications, different types of data are visualised: terrain data, 3D models of buildings, spatialised images, and points of interest augmenting buildings with additional data. The latter are queried at runtime from Wikidata and other databases. The terrain is retrieved from public APIs. The majority of the 3D building models are generated from data queried from OpenStreetMap (OSM). By simple extrusion of the building footprints, 3D geometries of LOD-1 (Level of Detail) can be generated only having flat roofs.

Another option is the integration of 3D models from other repositories. Photogrammetric or manual modeling processes are capable of producing a high level of detail and visual quality. However, the detailed scan of the building and its environment often does not fit to the approximated terrain. To address this challenge, a straightforward algorithm has been formulated to facilitate the adaptation of the photogrammetric model’s environment to the ground. The following steps must be assessed:

1. Detection of building vertices
2. Shifting of the building’s ground vertex to the terrain
3. Projection of the environment to the terrain plane

The initial step involves the integration of the floor plane of the corresponding building queried from OSM. The vertices of the model are iterated and a ray casting is used to check an encountering with the OSM plane. If this is the case, the vertex is declared as belonging to the building; otherwise, it is part of the environment. With the building vertices known, the deepest building vertex can be calculated to align the building part of the model with the ground first. In the last step, an iteration on the environment vertices and a ray casting again is performed to determine the distances to the ground. Each vertex is then shifted to the ground. The result is a mapping of the environment to the ground (Figure 11).

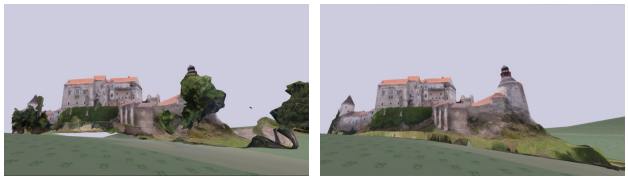

Figure 11. Screenshots of the visualised Pernštejn Castle. Left: Without ground projection – the model floats above the ground, the sky is visible and the scanned trees cover parts of the building. Right: With ground projection – alignment to the ground and no coverage.

Furthermore, two more optimisation techniques were tested: In order to reduce the model size, it is also possible to discard every environment vertex, if the building is only needed in the visualisation. However, if the environment is also to be shown in the visualisation, and the structure of it should also be kept, a weighted shifting of the environment vertices is possible. The displacement of vertices is determined by their distance from the building component of the model and the edges. Vertices in closer proximity to the edge will undergo a greater displacement than those in closer proximity to the building (Figure 12).

Figure 12. Left: Projection of the environment directly onto the terrain. Right: Blending the environment for a softer connection to the ground.

8. Application

For testing this functionality we are collaborating with several institutions, resulting in curated scenes in currently 7 countries:

- Amsterdam (NL): Valkenburg at 1,800 modelled by University of Amsterdam
- Dresden (DE): 3D model with various time layers and 5000+ images
- Pernštejn (CZ): 3D scan of Castle Pernštejn
- Valais (CH): 3D city model with 3 temporal layers reconstructed by EPFL
- Trento (IT): 3D reconstruction of the historical states by FBK
- Budapest (HU): 10k images of historic Budapest
- Leuven (BE): university quarter 3D model created by KU Leuven
- Jena (DE): 5,000 images and 3 temporal layers of cadastral data

9. Future Prospects

The 4D Browser and 4D City applications have been developed since 2016 via various projects at national and European scale. With regards to next steps, visual parameters and designs for 3D/4D visualisations of past architecture are yet rarely empirically validated (Münster et al., 2024b). Consequently, a currently starting next task will be to investigate and enhance the design of 3D/4D representations particularly considering the sparsity of historic data. Concerning 4D modelling, open research tasks are to reduce the number and quality of historical images necessary to enable the use of dense matching and the creation of historical (generalised) 3D models.

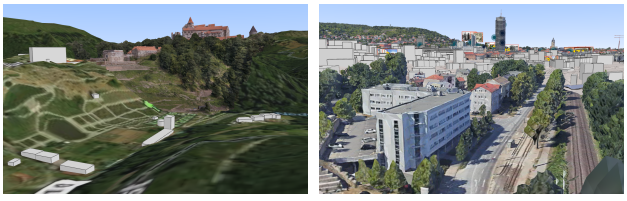

Figure 13. Screenshot of the 4D Browser of Castle Pernštejn with ingested 3D scan (Mikhail Volkov, CC-BY) (left) and Google 3D tile extension (right) in the 4D Browser.

Another approach currently under development is a large-scale automated workflow for orientation including contemporary data and 3D models. For data handling, the ingestion of LiDAR data and Google 3D Tiles as well as the parametric re-meshing seems promising to enhancing a contemporary visual experience (Figure 13).

Finally, we currently conduct tests to enhance the model coherence by AI-generated content as predicted façade textures or roof features. Beside the technical and design challenges, this also contains various methodical challenges about if and how AI can generate valid hypothesis of the past (Münster et al., 2024d).

Acknowledgments

The research which this paper is based on was carried out in the EU projects INDUX-R (Grant No. 101135556), 5DCulture (Grant No. 101100778) and 3DBigDataSpace (Grant No. 101173385).

References

Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., Szeliski, R., 2011. Buidling Rome in a Day. *Communications of the ACM*, 54(10), 105–112.

Bender, B., 2020. The Impact of Integration on Application Success and Customer Satisfaction in Mobile Device Platforms. *Bus Inf Syst Eng*, 62, 515–533.

Berton, G., Masone, C., 2025. MegaLoc: One Retrieval to Place Them All. *arXiv preprint arXiv:2502.17237*.

Blanc, N., Produit, T., Ingensand, J., 2018. A semi-automatic tool to georeference historical landscape images. *PeerJ Preprints*, 6, e27204v1.

Blettery, E., Lecat, P., Devaux, A., Gouet-Brunet, V., Saly-Giocanti, F., Brédif, M., Delavoipière, L., Conord, S., Moret, F., 2020. A Spatio-temporal Web Application for the Understanding of the Formation of the Parisian Metropolis. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, VI-4/W1-2020, 45–52.

Capurro, C., Plets, G., Verheul, J., 2024. Digital heritage infrastructures as cultural policy instruments: Europeana and the enactment of European citizenship. *International Journal of Cultural Policy*, 30(3), 304–324.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F., 2015. ShapeNet: An Information-Rich 3D Model Repository. *arXiv preprint arXiv:1512.03012*.

Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S. Y., VanderBilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A., 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663*.

Dufour, N., Picard, D., Kalogeiton, V., Landrieu, L., 2024. Around the World in 80 Timesteps: A Generative Approach to Global Visual Geolocation. *arXiv preprint arXiv:2412.06781*.

Flynn, T., 2022. Over 100,000 cultural heritage models on Sketchfab. https://sketchfab.com/nebulousflynn/collections/over-100000-cultural-heritage-models-on-sketchfab, accessed 29 January 2022.

Hasselman, T., Lo, W. H., Langlotz, T., Zollmann, S., 2023. ARephotography: Revisiting historical photographs using augmented reality. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, ACM, Article 40.

Jaillot, V., Rigolle, V., Servigne, S., Samuel, J., Gesquière, G., 2021. Integrating multimedia documents and time-evolving 3D city models for web visualization and navigation. *Transactions in GIS*, 25(3), 1419–1438.

Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K. M., Scherer, S., Krishna, M., Garg, S., 2023. AnyLoc: Towards Universal Visual Place Recognition. *arXiv preprint arXiv:2308.00688*.

Kulkarni, P. P., Nayak, G. K., Shah, M., 2024. CityGuessr: City-level video geo-localization on a global scale. *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXIII*, Springer-Verlag, Berlin, Heidelberg, 293–311.

Leroy, V., Cabon, Y., Revaud, J., 2024. Grounding Image Matching in 3D with MASt3R. *arXiv preprint arXiv:2406.09756*.

Lindenberger, P., Sarlin, P.-E., Pollefeys, M., 2023. LightGlue: Local Feature Matching at Light Speed. *ICCV*.

Maiwald, F., Bruschke, J., Schneider, D., Wacker, M., Niebling, F., 2023a. Giving Historical Photographs a New Perspective: Introducing Camera Orientation Parameters as New Metadata in a Large-Scale 4D Application. *Remote Sensing*, 15(7), 1879.

Maiwald, F., Komorowicz, D., Munir, I., Beck, C., Münster, S., 2023b. Semi-automatic generation of historical urban 3D models at a larger scale using Structure-from-Motion, neural rendering and historical maps. S. Münster, A. Pattee, C. Kröber, F. Niebling (eds), *Research and Education in Urban History in the Age of Digital Libraries*, Springer Nature Switzerland, Cham, 107–127.

Münster, S., 2018. Cultural heritage at a glance : Four case studies about the perception of digital architectural 3D models. *2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018)*, 1–4.

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), 9781.

Münster, S., Apollonio, F. I., Blümel, I., Fallavollita, F., Foschi, R., Grellert, M., Ioannides, M., Jahn, P. H., Kurdiovsky, R., Kuroczyński, P., Lutteroth, J.-E., Messemer, H., Schelbert, G., 2024a. *Handbook of Digital 3D Reconstruction of Historical Architecture*. Springer, Cham.

Münster, S., Bruschke, J., Dworak, D., Komorowicz, D., Rajan, V., Ukolov, D., 2024b. 4D geo modelling from different sources at large scale. *Proceedings of the 6th Workshop on the AnalySis, Understanding and ProMotion of HeritAge Contents*, SUMAC ’24, ACM, 13–17.

Münster, S., Bruschke, J., Hoppe, S., Maiwald, F., Niebling, F., Pattee, A., Utescher, R., Zarriess, S., 2022. Multimodal AI support of source criticism in the humanities – work in progress. *Digital Humanites 2022 – Conference Abstracts*, 527–529.

Münster, S., Bruschke, J., Maiwald, F., Kleiner, C., 2021a. Software and content design of a browser-based mobile 4D VR application to explore historical city architecture. *3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents*, ACM, 13–22.

Münster, S., Kamposiori, C., Friedrichs, K., Kröber, C., 2018. Image libraries and their scholarly use in the field of art and architectural history. *Int J Digit Libr*, 19, 367–383.

Münster, S., Lehmann, C., Lazariv, T., Maiwald, F., Karsten, S., 2021b. Toward an automated pipeline for a browser-based, city-scale mobile 4D VR application based on historical images. F. Niebling, S. Münster, H. Messemer (eds), *Research and Education in Urban History in the Age of Digital Libraries*, Springer International Publishing, Cham, 106–128.

Münster, S., Maiwald, F., Bruschke, J., Kröber, C., Sun, Y., Dworak, D., Komorowicz, D., Munir, I., Beck, C., Münster, D. L., 2024c. A Digital 4D Information System on the World Scale: Research Challenges, Approaches, and Preliminary Results. *Applied Sciences*, 14(5), 1992.

Münster, S., Maiwald, F., di Lenardo, I., Henriksson, J., Isaac, A., Graf, M. M., Beck, C., Oomen, J., 2024d. Artificial Intelligence for Digital Heritage Innovation: Setting up a R&D Agenda for Europe. *Heritage*, 7(2), 794–816.

Münster, S., Maiwald, F., Lehmann, C., Lazariv, T., Hofmann, M., Niebling, F., 2020. An automated pipeline for a browser-based, city-scale mobile 4D VR application based on historical images. *Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents*, ACM, 33–40.

Orzechowski, M., Łukasz Opioła, Martínez, I. L., Ioannides, M., Panayiotou, P. N., Łukasz Dutka, Słota, R. G., Kitowski, J., 2025. Integrated data, metadata, and paradata management system for 3D Digital Cultural Heritage objects: Workflow automation, federated authentication, and publication. *Future Generation Computer Systems*, 107964.

Pramanick, S., Nowara, E. M., Gleason, J., Castillo, C. D., Chellappa, R., 2022. Where in the world is this image? transformer-based geo-localization in the wild. *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, Springer-Verlag, Berlin, Heidelberg, 196–215.

Reips, U.-D., 2002. Internet-Based Psychological Experimenting: Five Dos and Five Don’ts. *Social Science Computer Review*, 20(3), 241–249.

Rusnak, A., Kaplan, F., 2025. HAECcity: Open-Vocabulary Scene Understanding of City-Scale Point Clouds with Superpoint Graph Clustering. *arXiv preprint arXiv:2504.13590*.

Sarlin, P.-E., DeTone, D., Yang, T.-Y., Avetisyan, A., Straub, J., Malisiewicz, T., Bulo, S. R., Newcombe, R., Kontschieder, P., Balntas, V., 2023. OrienterNet: Visual Localization in 2D Public Maps with Neural Matching. *CVPR*.

Schaffland, A., Bui, T. H., Vornberger, O., Heidemann, G., 2020. New interactive methods for image registration with applications in repeat photography. *Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents*, ACM, 41–48.

Schindler, G., Dellaert, F., 2012. 4D Cities: Analyzing, Visualizing, and Interacting with Historical Urban Photo Collections. *Journal of Multimedia*, 7(2), 124–131.

Singh, A., Aneja, S., 2024. NewsCaption: Named-Entity aware Captioning for Out-of-Context Media. *arXiv preprint arXiv:2403.12618*.

Snavely, N., Seitz, S. M., Szeliski, R., 2007. Modeling the World from Internet Photo Collections. *International Journal of Computer Vision*, 80(2), 189–210.

Stathopoulou, E.-K., Welponer, M., Remondino, F., 2019. Open-source Image-based 3D Reconstruction Pipelines: Review, Comparison and Evaluation. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W17, 331–338.

Tung, J., Chou, G., Cai, R., Yang, G., Zhang, K., Wetzstein, G., Hariharan, B., Snavely, N., 2025. MegaScenes: Scene-level view synthesis at scale. A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (eds), *Computer Vision – ECCV 2024*, Springer Nature Switzerland, Cham, 197–214.

Tyszkiewicz, M., Fua, P., Trulls, E., 2020. DISK: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33.

Vivanco Cepeda, V., Nayak, G. K., Shah, M., 2023. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization.

Wang, Z., Xu, D., Khan, R. M. S., Lin, Y., Fan, Z., Zhu, X., 2024. LLMGeo: Benchmarking Large Language Models on Image Geolocation In-the-wild. *arXiv preprint arXiv:2405.20363*.

Weyand, T., Araujo, A., Cao, B., Sim, J., 2020. Google Landmarks Dataset v2 – a large-scale benchmark for instance-level recognition and retrieval. *Proc. CVPR’20*.

Wu, X., Averbuch-Elor, H., Sun, J., Snavely, N., 2021. Towers of Babel: Combining images, language, and 3D geometry for learning multimodal vision. *ICCV*.

Wu, Y., Shi, L., Liu, H., Liao, H., Qiu, L., Yuan, W., Gu, X., Dong, Z., Cui, S., Han, X., 2024. MVImgNet2.0: A Larger-scale Dataset of Multi-view Images. *ACM Trans. Graph.*, 43(6), Article 173.

Xu, S., Zhang, C., Fan, L., Meng, G., Xiang, S., Ye, J., 2024. AddressCLIP: Empowering vision-language models for city-wide image address localization. *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXVIII*, Springer-Verlag, Berlin, Heidelberg, 76–92.