The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W6-2025
3rd International Workshop on Evaluation and BENCHmarking of Sensors, Systems and
GEOspatial Data in Photogrammetry and Remote Sensing (GEOBENCH), 20–21 November 2025, Wroclaw, Poland

egenioussBench: A New Dataset for Geospatial Visual Localisation

Phillipp Fanta-Jende[1], Francesco Vultaggio[1,2], Alexander Kern[3], Yasmin Loeper[2], Markus Gerke[2]

[1] Unit Assistive and Autonomous Systems, Center for Vision, Automation and Control,
AIT Austrian Institute of Technology, Vienna, Austria - (phillipp.fanta-jende, francesco.vultaggio)@ait.ac.at
[2] Institute of Geodesy and Photogrammetry, Technische Universität Braunschweig,
Brunswick, Germany - (m.gerke, y.loeper)@tu-braunschweig.de
[3] Institute of Flight Guidance, Technische Universität Braunschweig,
Brunswick, Germany - a.kern@tu-braunschweig.de

Keywords: Visual Localisation, Geospatial Mesh, Deep Learning, City Model

Abstract

We present egenioussBench, a visual localisation benchmark built on geospatial reference data: a city-scale airborne 3D mesh and a CityGML LoD2 model. This pairing reflects deployable mapping assets and supports true scalability beyond traditional SfM-based approaches. The query data comprise smartphone images with centimetre-accurate, map-independent ground truth obtained via PPK and GCP/CP-aided adjustment. From 2,709 images, we derive a non-co-visible subset by estimating the full co-visibility matrix from rendered depth and selecting a maximum independent set; the released data include a test split of 42 non-co-visible images with withheld ground truth and a validation split of 412 sequential images with poses, e.g. for training of pose regressors and self-validation. The benchmark features a public leaderboard evaluated with binning metrics at multiple pose-error thresholds alongside global statistics (median, RMSE, outlier ratio), ensuring fair, like-for-like comparison across mesh- and LoD2-based methods. Together, these design choices expose realistic cross-view and cross-domain challenges while providing a rigorous, scalable path for advancing large-scale visual localisation. We make the evaluation code and data availeable at `https://github.com/fratopa/egenioussBench` and `https://www.egeniouss.eu/`

1. Introduction

Visual localisation - estimating a camera's position and orientation from images - underpins a wide variety of platforms and applications such as autonomous driving (Sattler et al., 2018), robotics (Maggio et al., 2023), Unmanned Aerial Vehicles (UAVs) (Wu et al., 2024), and augmented reality (Pang et al., 2023). It is especially critical when GNSS is unavailable or unreliable, for instance indoors, in urban canyons, or under jamming and spoofing.

Traditional approaches rely on Structure-from-Motion (SfM) maps (Sattler et al., 2017, Sarlin et al., 2019), but these suffer from limited scalability and heavy storage demands, often reaching hundreds of gigabytes (Mera-Trujillo et al., 2020). Deep learning alternatives such as scene coordinate regression (Wang et al., 2024) offer efficiency and speed, yet still struggle with large-scale deployment. Moreover, most localisation techniques assume access to ground imagery which poses a constraint to their ability to scale to large scenes. These limitations motivate the need for alternative reference representations.

3D meshes are a promising option (Panek et al., 2022, Brachmann et al., 2023). Unlike SfM point clouds, they are naturally smaller in size, allow rendering from arbitrary viewpoints, enabling cross-view matching between ground-level imagery and aerial reconstructions (Vultaggio et al., 2024). This is vital for scalable localisation, particularly as municipalities increasingly provide detailed meshes and city models (Syed Abdul Rahman et al., 2024).

Progress in this area has been limited by the nature of available datasets. Existing approaches have typically relied on two types of data: large-scale meshes with coarse geometry and imprecise ground-truth poses for the query images (Berton et al., 2024), or smaller datasets where accurate ground-truth poses are obtained by co-registering query images with dense, high-resolution meshes reconstructed from street-level imagery (Panek et al., 2022, Sarlin et al., 2022) (see Figure 1).

While low-resolution datasets are valuable for training Visual Place Recognition (VPR) models, their limited accuracy makes them unsuitable for evaluating visual localisation performance. Conversely, datasets that depend on co-referencing with high-resolution 3D models inherently require meshes of very fine detail to produce precise ground-truth poses. This dependence not only makes such datasets impractical for large-scale mapping applications, but it also renders the localisation task very easy and will lead to unrealistically accurate results.

In this work, we contribute a new dataset for mesh-based visual localisation with centimeter-accurate ground truth collected in a map-independent fashion, thus allowing us to pair ground-level query poses with a challenging aerial map.

In cases where a mesh might not be available an additional source for geospatial reference is 3D city models, commonly provided as City Geography Markup Language (CityGML) models[1]. They are freely available for many cities, do not require a large amount of memory, and the scene representation database does not have to be generated from images. However, since CityGML models consist purely of potentially low-detailed geometry compared to the mentioned textured mesh data, it represents a challenge for traditional feature-based localisation techniques. Simplified meshes without texture and general Level of Detail (LoD) meshes are comparable to CityGML models. Current work addresses the challenging

[1] https://www.ogc.org/standard/citygml/

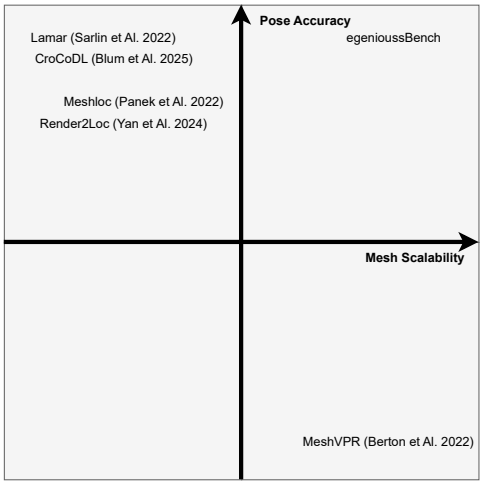

Figure 1. Comparison of egenioussBench to other state of the art mesh-based visual localisation datasets

yet promising reference data type (Zhu et al., 2024, Zhu et al., 2025, Panek et al., 2023). As with the work on large-scale meshes, the lack of precise and independent ground truth for the query data and georeferenced 3D models often presents an obstacle to evaluating visual localisation techniques.

Our contributions are threefold: (i) we introduce egenioussBench, the first benchmark coupling a high-resolution aerial 3D mesh with a CityGML LoD2 model; (ii) we release map-independent ground-level smartphone imagery with centimetre-accurate ground truth poses for cross-view evaluation; and (iii) we establish a benchmark protocol and leader-board to foster rigorous comparison of mesh- and object-based localisation methods. The dataset can be found at `https://www.egeniouss.eu/`

2. Related Work

2.1 Mesh-based Visual Localisation

Mesh-based visual localisation techniques can be broadly divided into two categories: iterative approaches and hierarchical ones. Iterative approaches assume to have access to an initial position and then adopt a render and compare strategy (Oishi et al., 2020, Yan et al., 2023) to refine the pose estimate of the query image. Instead, hierarchical approaches (Panek et al., 2022, Vultaggio et al., 2024) employ multi-stage retrieval methods to narrow down plausible images to a subset of candidate views and then determine the query position through accurate matches against these.

An often unspoken assumption of mesh-based localisation techniques is that the poses from which the reference images were captured, P_{ref}, are good poses from which to render images to be used in the localisation pipeline (Panek et al., 2022). While this is a sound assumption for 3D data collected from ground level, this assumption breaks in the case of an aerial dataset, where the large viewpoint difference makes reference poses harder – if not impossible – to localise from (Yan et al., 2023).

Another assumption in the mesh-based localisation literature is that a render-and-compare approach will naturally converge to the correct position if initialised in the neighbourhood of the ground-truth query pose, P_{gt} (Oishi et al., 2020, Yan et al., 2023). This assumption is grounded in the idea that real-world imagery is often trivially matched if captured from the same pose, defining an error cost function – $\mathcal{E}(P_r)$ – of matching a query image with a pose rendered at pose P_r,

$$\arg\min_{P} \mathcal{E}(P_r) = P_{\text{gt}}$$

This is not always the case: As shown in (Vultaggio et al., 2024), matching query images against renderings from the ground-truth query poses can often lead to substantial pose errors due to errors in the mesh reconstruction process, such as obstructions, blind spots, or inaccurate correspondences (see Figure 5).

With this work, we aim to challenge the assumptions underlying current mesh-based localisation techniques and highlight new areas for improvement, such as viewpoint selection, which has been recognised as a crucial aspect of localisation in the robotics community (Di Giammarino et al., 2025) but which is often overlooked in the visual localisation field.

2.2 Mesh-based Localisation Datasets

Most localisation datasets consist of posed reference images, posed query images, and a 3D model. Often these datasets assume the 3D model to take the form of a SfM point cloud in which each point is associated with a local visual descriptor. Among these, the popular ones are Aachen (Zhang et al., 2021), 12 scenes (Dong et al., 2022), inlooc(Taira et al., 2018), in these we find represented indoor (Taira et al., 2018, Dong et al., 2022) outdoor (Zhang et al., 2021), and challenging queries collected at different times of day and weather conditions.

Mesh-based localisation was first explored in indoor settings by (Oishi et al., 2020) as a loop closure and drift-correction technique. Full visual localisation using mesh maps was first proposed in MeshLoc (Panek et al., 2022). Here, the retrieval stage was conducted on real images, and the local matches and 3D lifting were performed on synthetic RGB and depth views. Other prominent datasets are Lamar (Sarlin et al., 2022) and CroCoDL (Blum et al., 2025), which provide high-resolution 3D meshes collected from a custom mobile mapping setup and generate the ground truth of the query images by co-registering them with the 3D map. Both MeshLoc, Lamar, and CroCoDL provide meshes with high-resolution textures and fine-grained geometry, made possible by careful ground-level data collection.

In this work, we argue that such datasets, while valuable, provide an unrealistically good reference mesh, which, while possible to obtain for small scenes, becomes prohibitively expensive to acquire for city-scale applications.

A dataset providing true city-scale data is MeshVPR (Berton et al., 2024), however here the query pose ground truth position has not been estimated to a level of precision sufficient for visual localisation as this dataset was originally developed to train Visual Place Recognition (VPR) models. Render2Loc (Yan et al., 2023) provides drone-captured data and accurate ground truth obtained through co-referencing of query images and a 3D model, however is not representative of true large-scale data collection, having been collected from a low-flying drone resulting in an estimated GSD of 1.03cm.

Another dataset with a similar ground truth query pose acquisition process to ours is the SLAM dataset recently released by (Krishnan et al., 2025), here the query pose is also obtained through the use of GCPs and image reconstruction, but being a SLAM dataset, it does not provide a reference map for visual localisation.

This work aims to fill a gap in datasets currently released by providing a true city-scale mesh dataset collected from airborne imagery and centimetre-accurate ground-level imagery. Both datasets underwent a professional photogrammetric bundle adjustment involving GCPs collected in the scene.

2.3 Object-based Visual Localisation

Object-based visual localisation uses an object's geometric characteristics as features to be matched in the query pose estimation process. The geometric features can be utilised in their original form, e.g., through projections or through rendered representations. The object data consists of simplified, textureless, and low-detail 3D models.

The approaches for the object data are similar to those that use 3D meshes as reference data. In hierarchical approaches, the 3D model is usually used to lift the 2D-2D feature matches of the query image with the retrieved reference images with known poses from the database into 3D (Panek et al., 2022). In the absence of reference images for image retrieval, sampling approaches are used to generate synthetic images that utilize the object data. The 2D-2D feature matching is then based on synthetic images instead of reference images (Panek et al., 2023).

In iterative pose estimation approaches, the correspondence between the features from the query image and the features from renderings of pose hypotheses is calculated and optimised (Zhu et al., 2024, Zhu et al., 2025, Loeper et al., 2024). A common problem with the use of synthetic images is the feature matching or the calculation of correspondence between the query image and the non-photorealistic renderings. Previous work has shown that cross-domain matching can be successful (Tomesek et al., 2022, Brejcha et al., 2020, Mikolka-Flöry et al., 2022). Nevertheless, there is a need for further development to calculate the correspondence between the query image and renderings of textureless and low-detailed geometry models, e.g. by developing feature matching approaches adapted to textureless 3D models.

Instead of explicitly using 3D models, other approaches implicitly use the models to train a neural network that predicts the pose (Acharya et al., 2022, Yao et al., 2024). In contrast to pose regressors that depend on SfM methods to generate training data or require images with ground truth poses, these approaches use synthetic images or features from synthetic images to train the neural network. Specifically, the synthetic image is rendered using 3D models and the ground truth query image poses. The features are then generated from the synthetic image, e.g. in the form of edge and segmentation maps. However, it must be emphasised that not only textureless and low-detailed models are used as training data. In addition, textureless models were either textured or rendered photorealistically before the edge and segmentation maps were generated. To date, research on using low-detailed, textureless models for pose regression remains limited.

2.4 Object-based Visual Localisation Datasets

With CADLoc (Panek et al., 2023), the authors provide a benchmark dataset for analysing imperfect 3D models from the Internet for use in visual localisation. In addition to the models, the dataset also contains query and reference images. Since the scale of the scenes is not known, the usual metrics of visual localisation cannot be used for evaluation. Instead, the mean and maximum Dense Re-Projection Error (Wald et al., 2020) are used.

For the visual localisation of UAVs, LoD-Loc provides two datasets consisting of LoD models and query images (Zhu et al., 2024). The query images of the datasets from LoD-Loc originate from the UAVD4L dataset (Wu et al., 2024) and from CrossLoc (Yan et al., 2021). The UAVD4L dataset contains rendered images, query images, their ground truth and digital surface model (DSM). The LoD model was generated in LoD3 from the oblique images of the UAVD4L dataset (Zhu et al., 2024). The LoD model for the query images from CrossLoc is in LoD2 and from the Swiss federal authorities.

Object-based visual localisation has not yet been extensively researched. In particular, there is a lack of datasets that contain both query images with precise, map-independent ground truth poses and georeferenced 3D models. However, LoD models are already available for many cities worldwide. Our contribution is to provide these two components in one dataset.

3. The Benchmark Challenge

egeniousBench evaluates state-of-the-art visual localisation under realistic, city-scale conditions. Participants are provided with two forms of geospatial reference data - a high-resolution aerial mesh and a CityGML LoD2 model - together with query images captured using a smartphone.

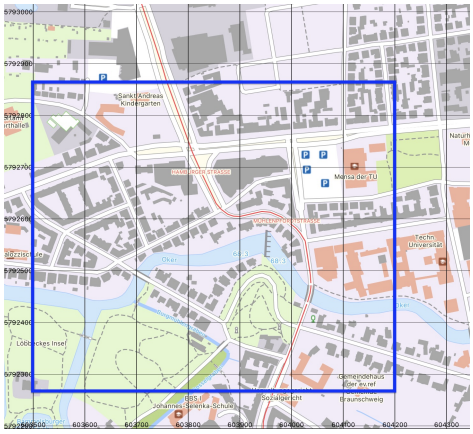

Figure 2. Overview of the area of interest in Braunschweig. Map source: basemap.de

Unlike previous benchmarks, the query images are non-overlapping in object space. This design prevents the use of multi-view geometry from image sequences and enforces true cold-start localisation. Methods must therefore rely either on implicit regressors (e.g. learning-based global pose estimation) or on explicit matching against the provided references (mesh or LoD2).

The query set is divided into two partitions: one with full ground-truth poses for training and self-validation, and one

with only approximate poses. The withheld ground-truth of the second partition is used for evaluation, with results published on a public leaderboard.

Participants will be asked to submit a CSV file with full 6DOF pose information per image. We will then evaluate each submission by binning according to thresholds compared to the ground truth (0.5m,2°/2m,5°/5m,10°), and overall median translation and rotation errors.

To ensure fair comparison, the challenge is evaluated separately for mesh-based and LoD2-based localisation. Examples of possible approaches are given in (Vultaggio et al., 2024) and (Loeper et al., 2024).

We aim to organise a scientific workshop or special issue featuring the results submitted by participants.

4. Datasets

egenioussBench covers an area within the city of Braunschweig, Germany, in Fig.2, a map is shown. The data is projected in UTM, zone 32N, and spans 570m in N-S and 700m in E-W direction. Corners at North West: [N: 5792850m, E: 603500m] and South East: [N: 5792280m, E: 604200m].

The dataset consists of three components:

1. Airborne-image-based 3D mesh (oblique imagery, 7.5 cm GSD) - reference for mesh-based localisation (Fig. 3);
2. CityGML LoD2 model - reference for object-based localisation (Fig. 4);
3. Smartphone query images — ground-level queries, with cm-level ground-truth poses (Fig. 5).

4.1 Airborne-image-based 3D mesh

Airborne oblique imagery acquired in May 2023 served as the basis for the 3D mesh reconstruction, see Fig. 3.

Camera system	UltraCam Osprey 4.1
Flying height (AGL)	∼1550 m
GSD (nadir / oblique)	7.5 cm / 6.5 cm (centre)
Georeferencing accuracy	∼1 GSD (XY), 1.5 GSD (Z)

4.2 CityGML LoD2

The second reference dataset is provided in the form of the 3D city model of Braunschweig (Landesamt für Geoinformation und Landesvermessung Niedersachsen, 2025), see Fig. 4. The 3D city model is given in LoD2. The dataset is supplied in OGC Standard CityGML format. A standardised process is used to create the 3D building models of Braunschweig. The process is based on the building outlines from cadastral maps, the DTM with 5 m grid resolution and 3D measurement data from laser scanning or matching point cloud (Landesamt für Geoinformation und Landesvermessung Niedersachsen, 2025).

The fundamentals for creating 3D building models are defined by product and quality standards (Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland, 2025). In accordance with these standards, the building footprints are extracted from the cadastre. The characteristic roof shape is modelled for the building on the basis

Figure 3. Geospatial mesh data; Data: Geofly, Processing: Skyline

Standard	CityGML, Level of Detail 2 (LoD2)
Geometry	Building footprints with prismatic roofs (generalised)
Source data	Cadastral outlines; 5 m DTM; LiDAR / image-matching point clouds
Horizontal ref.	Inherited from cadastre (city mapping frame)
Vertical ref.	Derived from point-cloud terrain heights
Empirical check	Most sampled building corners within ∼10 cm of mesh

of the building object of the cadastre and its roof shape attribute, provided that defined recording criteria are met. However, it may sometimes be necessary to model the roof manually. Using cadastre data as the basis for modeling, the 3D building models also obtain their positional accuracy. The accuracy of the coordinates from the cadastre varies, as the creation of the coordinates varies (Landesamt für Geoinformation und Landesvermessung, 2025). Therefore, no standardised value can be given for the positional accuracy. On average, according to an empirical check, the sampled building corners are within 10 cm of the airborne-image-based 3D mesh. The height accuracy can be given as approx. 1 m.

4.3 Query dataset

Ground-level query images were captured in January 2024 with a handheld smartphone rigidly mounted to a tactical-grade INS (Fig. 6). Images were resampled to 960x1280px, yielding an average GSD of ∼4 cm. See Fig. 5.

A precise trajectory was estimated via Post-Processed Kinematics (PPK) and refined with a Structure-from-Motion and bundle adjustment using Ground Control Points (GCPs) and Check Points (CPs) measured with RTK GNSS.

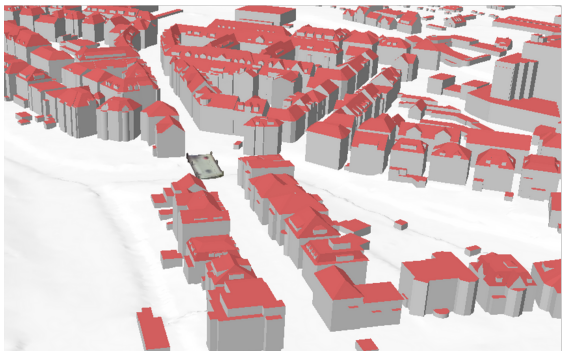

Figure 4. LoD2 model; Data: City of Braunschweig

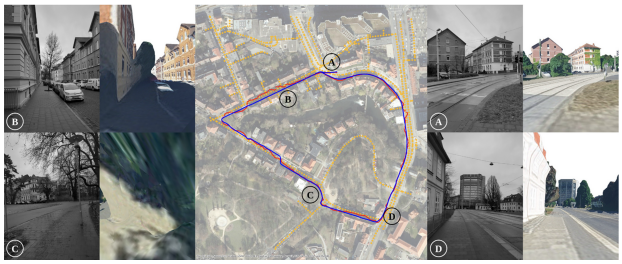

Figure 5. Left and right: examples of the query images and the view rendered from its reference Center: reference trajectory of the camera in blue, the smartphone's internal GNSS pose estimate in red, and the sampled views in orange, source: (Vultaggio et al., 2024).

GCP RMSE	(4, 4, 7) cm in X, Y, Z
CP mean error	(10, 10, 8) cm in X, Y, Z
Image pose mean error	8 mm (X,Y), 3 mm (Z)
Image pose std. dev.	7 mm (X,Y), 1 mm (Z)
Orientation mean error	0.04° ($\sigma = 0.03^\circ$)

The collected dataset comprises 2709 RGB images. To generate the query set, we estimate a subset of non-co-visible images by first rendering the complete set of views using the mesh data. The depth data are then used to compute the full co-visibility matrix between all image pairs. To identify the largest non-co-visible subset, the matrix is converted into a graph where image pairs with fewer than 10% co-visible pixels are considered non-co-visible. Finally, we compute the maximum independent set on this graph using the solver proposed by (Hespe et al., 2019). This set of non co-visible images has been further refined through manual inspection to filter out any spuriously selected image resulting from poor meshing. The final released dataset consists of a test set comprised of 42 non-co-visible images for which the ground-truth position is withheld, and a validation set of 412 consecutive images for which the ground truth poses are provided.

5. Benchmark Baseline

To establish a baseline for this dataset, we evaluate MeshLoc (Panek et al., 2022) and our previous Visual Localisation method (Vultaggio et al., 2024). MeshLoc assumes that suitable poses for rendering the reference views correspond to those from which the images used to generate the map were originally captured. In our case, this would imply rendering views from an altitude of 1550 m (i.e. acquisition altitude of the aerial dataset); therefore, we instead employ the initialisation procedure described in our earlier work.

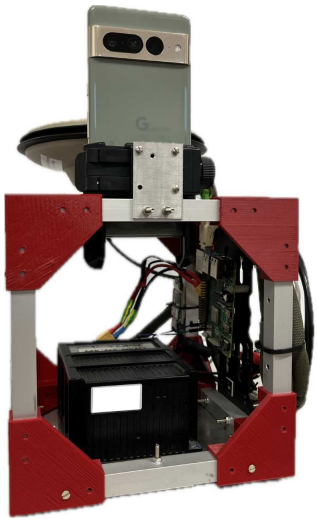

Figure 6. Smartphone rigidly mounted to INS system to capture query data.

To initialise both methods, we render views in the mesh using open street map data to extract the road network and then sample views along the path. In order to compare both methods in a fair way, we use the same retrieval procedure for both methods where we first retrieve the 500 most visually similar images from the pre-rendered views using CosPlace (Berton et al., 2022) global descriptors tuned to work on synthetic images (Berton et al., 2024) and then filter them to obtain the 50 closest ones to the smartphone's GNSS pose estimate. Subsequently, the query image is matched against the subset of promising images and the full set of 2D-3D correspondences is used to compute the final pose estimate. To have more details on the initialisation procedure, please refer to our earlier work (Vultaggio et al., 2024). Both methods use SuperPoint (DeTone et al., 2018) to extract local features.

Method	0.5m,2°/2m,5°/5m,10° ↑ **% / % / %**	ME ↓ m , °	Time ↓ s
MeshLoc	**19.05** / 66.67 / 76.19	0.97 , 0.63	290
Ours	**19.05 / 69.05 / 78.57**	**0.89** , **0.56**	**41**

Table 1. Results of MeshLoc and our method on the egenioussBench dataset. We report the percentage of the 42 query frames localised within each threshold and the median translation and rotational error.

The results are presented in table 1 and show the percentage of frames localised within three thresholds in accordance with previous visual localisation literature (Zhang et al., 2021, Vultaggio et al., 2024) and the median translation and rotation error over all the frames. We also report the execution time of both methods, although this will not be assessed in the final rankings, as not all methods will be run on the same hardware.

The results show that when using the same retrieved images, both methods perform equally well. Our earlier method performs slightly better due to the use of a more advanced pose estimation and refinement pipeline which includes a modern RANSAC (Barath et al., 2020) and PnP (Vultaggio et al., 2025). It is also apparent that this dataset is more challenging than any other available benchmark dataset present in the literature when comparing the accuracy achieved by MeshLoc on our dataset. For instance, employing MeshLoc on the Aachen dataset (Zhang et al., 2021), it becomes obvious that the perform-

ance has degraded as a result of the more challenging texture and possibly the automatic reference pose generation process.

6. Timeline

The following schedule is envisaged:

- November 2025: Release of data
- May 2026: deadline for submissions of solutions, coming along with a short description of the used method, if it should be included in the paper/workshop.

7. Conclusion

We introduce egenioussBench, a benchmark for mesh- and object-based visual localisation that combines an airborne-image-derived city mesh with a CityGML LoD2 model and centimetre-accurate, map-independent ground truth for smartphone queries. By constructing a non-co-visible query subset via a maximum independent set on a co-visibility graph, the benchmark enforces true cold-start localisation while the sequential validation split supports training and self-validation. Crucially, by relying on deployable and standardised geospatial assets rather than dense, photorealistic SfM maps, egenioussBench is designed to challenge truly scalable approaches - those that can operate with modest storage/compute footprints and be maintained at city or national scale. A public leaderboard with binning metrics across pose-error thresholds and global statistics ensures rigorous, like-for-like comparison across reference types.

Our baseline evaluations with MeshLoc and our approach are designed to make results directly comparable to the prevailing mesh-based literature. Using identical retrieval procedure and the same evaluation protocol, both methods achieve similar accuracy, while our approach attains substantially lower runtime. We emphasise that our method is early-stage and primarily included to establish a clear point of reference; speed is reported for context rather than ranking.

We hope egenioussBench will catalyse research on (i) viewpoint selection and rendering pose design for aerial-to-ground matching, (ii) robust cross-domain correspondence between real images and texture-less or low-detail geometric models (e.g. LoD2), (iii) retrieval and initialisation strategies that exploit weak priors, and (iv) learning-based pose regression that can leverage sparse or simplified 3D. We also encourage work that explicitly optimises storage, rendering, and compute budgets as first-class scalability objectives.

We release data, evaluation code, a public leaderboard, and plan a community event to consolidate progress. Future extensions will broaden geographic coverage and diversify query acquisition conditions (e.g. UAV, bicycle) to further stress-test generalisation and to evaluate methods at larger spatial scales.

ACKNOWLEDGEMENTS

Funded by the European Union

This work is part of the EU-Horizon project *egeniouss*[2], which received funding under the call HORIZON-EUSPA-2021-Space with the project number 101082128.

[2] https://www.egeniouss.eu/

References

Acharya, D., Tennakoon, R., Muthu, S., Khoshelham, K., Hoseinnezhad, R., Bab-Hadiashar, A., 2022. Single-image localisation using 3D models: Combining hierarchical edge maps and semantic segmentation for domain adaptation. 136, 104152. PII: S0926580522000255.

Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland, 2025. Produkt- und Qualitätsstandard für 3D-Gebäudemodelle. https://www.adv-online.de/AdV-Produkte/Standards-und-Produktblaetter/Standards-der-Geotopographie/ (last visited October 08, 2025).

Barath, D., Noskova, J., Ivashechkin, M., Matas, J., 2020. MAGSAC++, a Fast, Reliable and Accurate Robust Estimator. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Seattle, WA, USA, 1301–1309.

Berton, G., Junglas, L., Zaccone, R., Pollok, T., Caputo, B., Masone, C., 2024. MeshVPR: Citywide Visual Place Recognition Using 3D Meshes. *Computer Vision – ECCV 2022*.

Berton, G., Masone, C., Caputo, B., 2022. Rethinking Visual Geo-Localization for Large-Scale Applications. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4878–4888.

Blum, H., Mercurio, A., O'Reilly, J., Engelbracht, T., Dusmanu, M., Pollefeys, M., Bauer, Z., 2025. CroCoDL: Cross-device Collaborative Dataset for Localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27424–27434.

Brachmann, E., Cavallari, T., Prisacariu, V. A., 2023. Accelerated Coordinate Encoding: Learning to Relocalize in Minutes Using RGB and Poses. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Vancouver, BC, Canada, 5044–5053.

Brejcha, J., Lukáč, M., Hold-Geoffroy, Y., Wang, O., Čadík, M., 2020. Landscapear: Large scale outdoor augmented reality by matching photographs with terrain models using learned descriptors. A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (eds), *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, 12374, Springer International Publishing, 295–312.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. SuperPoint: Self-Supervised Interest Point Detection and Description. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Salt Lake City, UT, USA, 337–33712.

Di Giammarino, L., Sun, B., Grisetti, G., Pollefeys, M., Blum, H., Barath, D., 2025. Learning Where to Look: Self-supervised Viewpoint Selection for Active Localization Using Geometrical Information. A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (eds), *Computer Vision – ECCV 2024*, Springer Nature Switzerland, Cham, 188–205.

Dong, S., Wang, S., Zhuang, Y., Kannala, J., Pollefeys, M., Chen, B., 2022. Visual Localization via Few-Shot Scene Region Classification. *2022 International Conference on 3D Vision (3DV)*, IEEE Computer Society, 393–402.

Hespe, D., Schulz, C., Strash, D., 2019. Scalable Kernelization for Maximum Independent Sets. *ACM Journal of Experimental Algorithmics*, 24(1), 1.16:1–1.16:22. https://doi.org/10.1145/3355502.

Krishnan, A., Liu, S., Sarlin, P.-E., Gentilhomme, O., Caruso, D., Monge, M., Newcombe, R., Engel, J., Pollefeys, M., 2025. Benchmarking Egocentric Visual-Inertial SLAM at City Scale.

Landesamt für Geoinformation und Landesvermessung, 2025. ALKIS. "https://ni-lgln-opengeodata.hub.arcgis.com/apps/lgln-opengeodata::alkis/about" (last visited October 21, 2025).

Landesamt für Geoinformation und Landesvermessung Niedersachsen, 2025. Opengeodata niedersachsen. "https://ni-lgln-opengeodata.hub.arcgis.com/" (last visited October 21, 2025).

Loeper, Y., Gerke, M., Alamouri, A., Kern, A., Bajauri, M. S., Fanta-Jende, P., 2024. Visual localization in urban environments employing 3D city models. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W8-2024, 311–318. https://isprs-archives.copernicus.org/articles/XLVIII-2-W8-2024/311/2024/.

Maggio, D., Abate, M., Shi, J., Mario, C., Carlone, L., 2023. Loc-nerf: Monte carlo localization using neural radiance fields. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 4018–4025.

Mera-Trujillo, M., Smith, B., Fragoso, V., 2020. Efficient Scene Compression for Visual-based Localization. *2020 International Conference on 3D Vision (3DV)*, 1–10.

Mikolka-Flöry, S., Ressl, C., Schimpl, L., Pfeifer, N., 2022. Automatic orientation of historical terrestrial images in mountainous terrain using the visible horizon. 6, 100026. PII: S2667393222000151.

Oishi, S., Kawamata, Y., Yokozuka, M., Koide, K., Banno, A., Miura, J., 2020. C*: Cross-Modal Simultaneous Tracking and Rendering for 6-DoF Monocular Camera Localization Beyond Modalities. *IEEE Robotics and Automation Letters*, 5(4), 5229–5236.

Panek, V., Kukelova, Z., Sattler, T., 2022. MeshLoc: Mesh-Based Visual Localization. S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, T. Hassner (eds), *Computer Vision – ECCV 2022*, 13682, Springer Nature Switzerland, Cham, 589–609.

Panek, V., Kukelova, Z., Sattler, T., 2023. Visual Localization using Imperfect 3D Models from the Internet. *10.48550/arXiv.2304.05947*.

Pang, W., Xia, C., Leong, B., Ahmad, F., Paek, J., Govindan, R., 2023. Ubipose: Towards ubiquitous outdoor ar pose tracking using aerial meshes. *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, 1–16.

Sarlin, P.-E., Cadena, C., Siegwart, R., Dymczyk, M., 2019. From coarse to fine: Robust hierarchical localization at large scale. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12716–12725.

Sarlin, P.-E., Dusmanu, M., Schönberger, J. L., Speciale, P., Gruber, L., Larsson, V., Miksik, O., Pollefeys, M., 2022. Lamar: Benchmarking localization and mapping for augmented reality. *ECCV*.

Sattler, T., Leibe, B., Kobbelt, L., 2017. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. 39(9), 1744–1756.

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T., 2018. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8601–8610.

Syed Abdul Rahman, S. A. F., Abdul Maulud, K. N., Ujang, U., Wan Mohd Jaafar, W. S., Shaharuddin, S., Ab Rahman, A. A., 2024. The Digital Landscape of Smart Cities and Digital Twins: A Systematic Literature Review of Digital Terrain and 3D City Models in Enhancing Decision-Making. *SAGE Open*, 14(1), 21582440231220768.

Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A., 2018. InLoc: Indoor Visual Localization with Dense Matching and View Synthesis.

Tomesek, J., Cadik, M., Brejcha, J., 2022. Crosslocate: Cross-modal large-scale visual geo-localization in natural environments using rendered modalities. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2193–2202.

Vultaggio, F., Fanta-Jende, P., Gerke, M., 2025. Perspective-n-Point in Practice: Performance, Robustness, and Accuracy for Mesh-Based Localisation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-1-W4-2025, 131–138.

Vultaggio, F., Fanta-Jende, P., Schörghuber, M., Kern, A., Gerke, M., 2024. Investigating Visual Localization Using Geospatial Meshes. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W8-2024, 447–454. https://isprs-archives.copernicus.org/articles/XLVIII-2-W8-2024/447/2024/.

Wald, J., Sattler, T., Golodetz, S., Cavallari, T., Tombari, F., 2020. Beyond Controlled Environments: 3D Camera Re-Localization in Changing Indoor Scenes. *https://arxiv.org/abs/2008.02004*.

Wang, F., Jiang, X., Galliani, S., Vogel, C., Pollefeys, M., 2024. GLACE: Global Local Accelerated Coordinate Encoding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Seattle, WA, USA, 5819–5828.

Wu, R., Cheng, X., Zhu, J., Liu, X., Zhang, M., Yan, S., 2024. UAVD4L: A Large-Scale Dataset for UAV 6-DoF Localization. *10.48550/arXiv.2401.05971*.

Yan, Q., Zheng, J., Reding, S., Li, S., Doytchinov, I., 2021. CrossLoc: Scalable Aerial Localization Assisted by Multimodal Synthetic Data. *10.48550/arXiv.2112.09081*.

Yan, S., Cheng, X., Liu, Y., Zhu, J., Wu, R., Liu, Y., Zhang, M., 2023. Render-and-Compare: Cross-view 6-DoF Localization from Noisy Prior. *2023 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE Computer Society, 2171–2176.

Yao, D., Zhu, H., Ren, B., Zhuang, X., 2024. Improving single image localization through domain adaptation and large kernel attention with synthetic data. 137, 108951. PII: S0952197624011096.

Zhang, Z., Sattler, T., Scaramuzza, D., 2021. Reference Pose Generation for Long-term Visual Localization via Learned Features and View Synthesis. *International Journal of Computer Vision*, 129(4), 821–844.

Zhu, J., Peng, S., Wang, L., Tan, H., Liu, Y., Zhang, M., Yan, S., 2025. LoD-Loc v2: Aerial Visual Localization over Low Level-of-Detail City Models using Explicit Silhouette Alignment. *https://doi.org/10.48550/arXiv.2507.00659*.

Zhu, J., Yan, S., Wang, L., Zhang, S., Liu, Y., Zhang, M., 2024. LoD-Loc: Aerial Visual Localization using LoD 3D Map with Neural Wireframe Alignment. *10.48550/arXiv.2410.12269*.