The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

An Analysis of the Impact of Geospatial Data Sources on Mesh-Based Localisation Performance

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

[1] Austrian Institute of Technology, Center for Vision, Automation and Control, Unit Assistive and Autonomous Systems, Austria – (francesco.vultaggio, phillipp.fanta-jende)@ait.ac.at

[2] Technische Universität Braunschweig, Institute of Flight Guidance, Germany – a.kern@tu-braunschweig.de

[3] Technische Universität Braunschweig, Institute of Geodesy and Photogrammetry, Germany – m.gerke@tu-braunschweig.de

Keywords: Visual localisation, Mesh, Aerial Image, Smartphone Image, GNSS, Navigation

Abstract

This paper investigates how the provenance and resolution of geospatial data used to construct mesh maps affect the accuracy and robustness of mesh-based visual localisation. Mesh-based approaches offer significant advantages over traditional pipelines reliant on Structure from Motion (SfM) models, including the ability to scale to city-sized scenes—by leveraging large-scale data sources such as national mapping databases— and on-demand generation of arbitrary synthetic views. While prior work has focused on algorithmic improvements to mesh-based localisation, none has systematically analysed how different input data affect localisation outcomes. In this work, we evaluate three meshes—derived from aerial oblique imagery, combined aerial and ground mobile mapping data, and close-range ground imagery—across the egenioussBench Extended and House of Science query sets and four image matchers. We show that mesh quality is the dominant factor governing localisation performance. In the House of Science experiments, aerial meshes lack the resolution required to resolve façade detail, causing near-total localisation failure regardless of matcher. In the egenioussBench Extended experiments, augmenting an aerial mesh with ground data yields consistent but less dramatic improvements. We further introduce the Perceptual Detail Score (PDS), a viewing-condition-aware metric that proves to be a strong predictor of downstream pose accuracy across all experimental configurations.

Figure 1. Comparison between real images and rendered images from the same positions. On top sample image from the egenioussBench dataset, on the bottom sample images from the House of Science dataset. From left to right; real query image, Supermesh and ground mesh, aerial mesh.

1. Introduction

Visual localisation consists of estimating the position and orientation of a query image given a prior map of the environment the image was captured in. Visual localisation plays a crucial role in all applications where absolute pose is needed but GNSS might not be available, such as in indoor environments (Taira et al., 2018), or where GNSS is only available in a degraded state such as in urban canyons (More et al., 2024), adversarial environments (Radoš et al., 2024), or when the precision provided by GNSS is insufficient, such as in augmented reality (Sarlin et al., 2022).

The best way to understand the many different visual localisation techniques proposed in the literature is to categorise them based on the kind of *map* the technique uses to perform the localisation process. The term map as used by the visual localisation community is non-specific, encompassing representations varying from Structure from Motion (SfM) point clouds (Sattler et al., 2012), collections of posed images (Berton et al., 2022), the weights of a neural network (Brachmann et al., 2023), 3D meshes (Panek et al., 2022, Vultaggio et al., 2024), and more (Nadeem et al., 2019, Speciale et al., 2019, Loeper et al., 2024), for a broad review of the field we refer the reader to (Chen et al., 2023, Panek et al., 2025). Looking at the map representation as the key differentiator, one can distinguish four principal families of techniques.

Traditional methods (Sarlin et al., 2019, Sattler et al., 2012) rely on SfM point clouds and can be seen as a natural extension of incremental SfM reconstructions. Such approaches rely on global descriptors (Berton et al., 2022) to quickly index promising reference images, cluster them based on pre-computed covisibility analysis, match local features (DeTone et al., 2018, Lindenberger et al., 2023) with the triangulated 3D points visible in each reference view cluster, and finally obtain the query image pose through PnP algorithms (Lepetit et al., 2009). These are the most accurate methods (Sattler et al., 2018) but also the most difficult to scale due to their dependence on ground data and the large dimensions of the final map (Mera-Trujillo et al., 2020).

Visual Place Recognition (VPR) (Schubert et al., 2023) methods use collections of posed images and compute the position of an image by computing the visual similarity between reference views and the query image, the pose is then assumed to be a weighted average of their positions. These methods tend to be the fastest and most scalable (Keetha et al., 2023) at the cost of accuracy due to the lack of any explicit or implicit geometric information in the pose estimation process.

Regression methods train a neural network on a collection of posed images to predict the pose of query images in the scene. Some of them attempt to predict the pose itself, absolute pose regressors (Kendall et al., 2015), while others instead predict the coordinates of each of the pixels in the query image, scene coordinate regressors (Brachmann et al., 2023, Jiang et al., 2025). These methods are extremely fast and accurate but struggle to scale to large scenes and offer poor robustness compared to traditional SfM-based approaches.

Model-based techniques use the reference images to build an intermediate map representation such as a mesh (Oishi et al., 2020, Panek et al., 2022, Vultaggio et al., 2024), a Gaussian splat (Khatib et al., 2025), or a LOD model (Loeper et al., 2024). Building such intermediate models allows these techniques to compress the information encoded in the original reference images and generate necessary views for localisation from the 3D model itself. These methods compromise on absolute precision but offer more scalability.

Mesh-based localisation offers multiple advantages, in particular —thanks to modern photogrammetric pipelines— it is possible to obtain accurate and detailed meshes from aerial images enabling mesh-based approaches to seamlessly bridge the cross-view localisation barrier (Vultaggio et al., 2024), which remains an open question in the image domain (Sarlin et al., 2023), and to scale to city-sized scenes at a fraction of the effort required by techniques which demand ground reference data.

The literature on mesh-based localisation has thus far focused on the best algorithmic improvements to leverage meshes as maps for localisation systems (Vultaggio et al., 2024, Vultaggio et al., 2025, Oishi et al., 2020, Yan et al., 2023); in this work we intend to instead focus our attention on how different meshes can affect the localisation accuracy and robustness. To this end, we use the following data sources for generating meshes as reference:

- **Aerial Oblique Imagery**: Standard aerial surveying data; offers good compromise between availability, façade detail and resolution
- **Aerial Oblique Imagery and Ground Mobile Mapping Imagery and Mobile Laser scanning (MLS)**: Both aerial and ground datasets combined in a single mesh are referred to as *Supermesh*; this type of dataset is at an experimental stage
- **Close range Ground Imagery**: Highest resolution and detail but also the most difficult to obtain at city scale.

We use these different data to build mesh maps and use those to localise ground-level images captured from a smartphone whose ground truth has been obtained by Structure from Motion and bundle block adjustment involving Ground Control Points (GCPs), resulting in a highly accurate and map-independent geo-localisation of the query data, see Fig.1 for a comparison of the query images against the reference meshes.

This is the first work to analyse the impact that high-resolution meshes have on visual localisation outcomes. We follow the approach proposed by (Vultaggio et al., 2024) where, after an offline pre-rendering stage, the position of the query images is estimated via an iterative filtering approach followed by correspondence finding and pose estimation via robust PnP. We measure both the accuracy and robustness obtained by the different data sources. We will measure the accuracy in terms of threshold pose errors and global metrics in terms of median and RMSE translational and rotational errors. Our goal is to shed light on a thus far overlooked —but crucial— aspect of mesh-based visual localisation: the impact that reference data have on downstream localisation performances.

2. Related Works

The first works to use meshes for localisation purposes were indoor techniques in which the localisation process was used to correct for drift in camera tracking (Oishi et al., 2020). The work that first popularised the idea for larger-scale applications was MeshLoc (Panek et al., 2022), which used ground reference images to build a mesh and then rendered views from the reference poses to localise query ground images. This first work was limited insofar as it only relied on rendered images for the local feature matching stage while still using real reference images for the retrieval stage. Later works (Berton et al., 2024) trained global descriptors to retrieve real images using synthetic views, enabling fully mesh-based localisation pipelines.

Some works have explored using aerial meshes for mesh-based localisation. The work from (Vultaggio et al., 2024) proposes using an aerial mesh to localise ground imagery and evaluates different initialisation procedures, while (Yan et al., 2023) uses meshes obtained from imagery collected from a low-flying drone to localise query images captured from a smartphone. The main difference between the two is that (Vultaggio et al., 2024) pre-renders reference views around the city from which to localise ground queries, while (Yan et al., 2023) renders views on-demand based on pose priors obtained from GNSS, refined in an iterative render-and-compare fashion.

While several prior works have explored the relationship between 3D model and localisation accuracy, none have systematically varied the input data used to construct the mesh while holding the localisation pipeline constant. The work from (Panek et al., 2023) compared localisation outcomes across multiple 3D models, but their analysis focused on hand-sculpted and internet-sourced models without standard geodetic provenance and processing; no mesh generation metadata (e.g. Ground Sampling Distance, camera specifications) was reported. The work from (Zhang et al., 2024b) compared traditional meshes, colourised point clouds, and Gaussian splatting, but used the *same* input data for all representations, thereby isolating the effect of *model type* rather than *data quality*. In contrast, our work fixes the model type (textured mesh) and the localisation pipeline, and systematically varies the provenance and resolution of the input data, thereby directly quantifying the impact of data quality on downstream localisation performance.

Other mesh datasets popular in the visual localisation community include (Sarlin et al., 2022, Blum et al., 2025), which provide high-quality meshes obtained through mobile mapping rigs combining high-resolution LiDAR and camera systems. The query image poses in these works were acquired through accurate co-registration with the final 3D model. While these datasets of-

Nadir and Oblique Aerial Imagery	
Camera system	UltraCam Osprey 4.1
View Distance (AGL)	∼1550 m
GSD (nadir / oblique)	7.5 cm / 6.5 cm (centre)
Georeferencing accuracy	∼1 GSD (XY), 1.5 GSD (Z)

Table 1. Details on the egeniousBench reference data

House of Science Ground Data	
Camera system	Apple iPad (PIX4Dcatch)
View Distance (AGL)	ground level
GSD	0.2 cm (average)
Georeferencing accuracy	mean RMS 0.029 m (23 GCPs)

Table 2. Details on the House of Science reference data

fer some of the most accurate 3D data available in the literature, we argue they are not representative of standard large-scale 3D models. In this work we use the egeniousBench query set (Fanta-Jende et al., 2025) for the urban-scale experiments, and introduce the House of Science dataset, a new close-range façade benchmark acquired at the House of Science building in Braunschweig, for the close-range experiments.

3. Method

3.1 Reference Data Acquisition and Mesh Generation

The airborne imagery used to generate the aerial mesh was acquired in May 2023, see Tab.1 for details. The ground level imagery and LiDAR data used for the Supermesh generation has been provided by Cyclomedia[1]. The 3D reconstructions for both the oblique only and the supermesh have been carried out by Skyline[2].

For the close-range tests, ground imagery was captured using an iPad via PIX4Dcatch in November 2025 and processed with PIX4Dmatic (Pix4D SA, 2024), see Tab.2

3.2 Query data acquisition

The first set of ground-level query images was acquired in January 2024 in Braunschweig using a handheld smartphone rigidly mounted to a tactical-grade inertial navigation system (INS) (Fig. 2). Images were resampled to 960×1280 px, yielding a mean ground sampling distance (GSD) of approximately $4\,\mathrm{cm}$. The camera trajectory was estimated via Post-Processed Kinematic (PPK) GNSS and subsequently refined through Structure-from-Motion (SfM) with bundle adjustment, constrained by Ground Control Points (GCPs) and validated against Check Points (CPs) measured with RTK GNSS.

[1] https://www.cyclomedia.com/en
[2] https://www.skylinesoft.com/

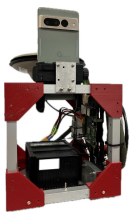

Figure 2. Image Capturing rig comprising the smartphone and PPK system.

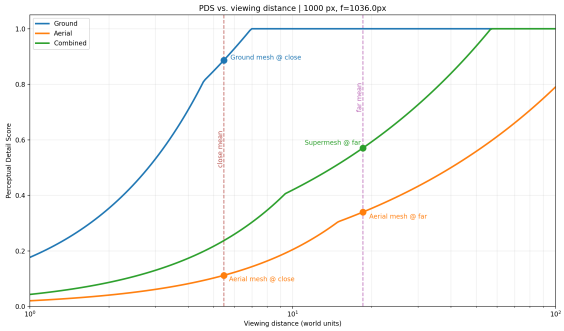

Figure 3. Perceptual Detail Score (PDS) versus viewing distance for the three reference meshes. Vertical dashed lines indicate the mean viewing distance of each query dataset, estimated from depth maps computed during ground-truth processing with PIX4Dmatic. PDS saturates at 1 once geometric and textural features become sub-pixel at the rendering resolution.

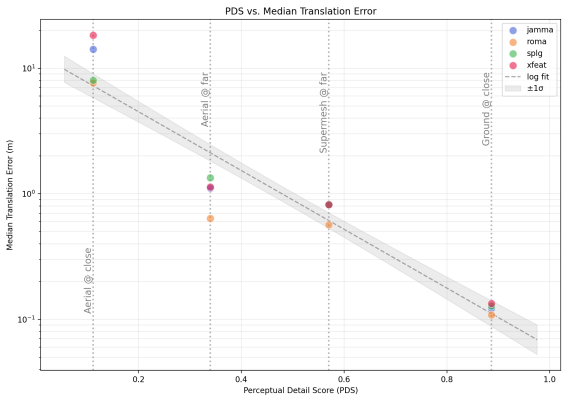

Figure 4. Mesh quality, as measured by the Perceptual Detail Score (PDS) versus median translation pose error of the Visual Localisation pipeline. Across all 16 experiment–matcher combinations, the PDS is a strong predictor of the final accuracy of the visual localisation pipeline.

The second set of query images, used in the House of Science experiments, was acquired in March 2026 at the House of Science building in Braunschweig to simulate a surveying scenario, in contrast to the pedestrian navigation trajectory of the first set. Images were captured using a smartphone running PIX4Dcatch at a resolution of 1920×1440 px, yielding a mean GSD of $0.6\,\mathrm{cm}$ over a covered area of approximately $600\,\mathrm{m}^2$. The position and orientation of the camera was estimated via SfM constrained by GCPs and validated against CPs measured with RTK GNSS. The resulting mean GCP RMS position error was $0.024\,\mathrm{m}$. The images were then rescaled to 960×720 px for the Visual Localisation experiments.

3.3 Experimental Design

To evaluate the impact of mesh data on localisation performance we use three reference meshes and two query datasets, yielding four experiments. To motivate their selection, we propose the *Perceptual Detail Score* (PDS), a scalar that quantifies how much of a mesh’s geometric and textural detail is visible under a given viewing condition. PDS projects the characteristic feature size of each channel into screen space via the pinhole model, yielding a per-channel score:

$$s = \left[\max\left(\frac{l\,f}{d}, 1\right)\right]^{-1}, \qquad \mathrm{PDS} = \sqrt{s_{\mathrm{geom}} \cdot s_{\mathrm{tex}}} \tag{1}$$

where f is the focal length in pixels and d is the viewing distance. l_{geom} is the RMS of the per-triangle mean edge length across all triangles in the mesh. l_{tex} is the area-weighted RMS world-space texel footprint:

$$l_{\mathrm{tex}} = \sqrt{\frac{\sum_i A_i\,\ell_i}{\sum_i A_i}}, \qquad \ell_i = \frac{A_i}{A_i^{(uv)} \cdot W \cdot H} \tag{2}$$

where A_i is the world-space area of triangle i, $A_i^{(uv)}$ its area in UV space, and $W \times H$ the texture resolution. Both l_{geom} and l_{tex} are in world-space units and enter Eq. 1 symmetrically. Each channel score $s \in [0, 1]$ reaches unity when features are finer than one rendered pixel; below that threshold it decays as the inverse of the projected feature size. It should be noted that the PDS is a theoretically derived measure and cannot account for local mesh deficits arising from the specific properties of the acquisition geometry. Localisation accuracy in practice will therefore be influenced by the degree to which the query viewpoints happen to coincide with well—or poorly—reconstructed regions of the map, a factor that the PDS alone cannot predict. In practice we will see in Sec.4 that the PDS is a strong predictor of the final position error. As shown in Fig. 3, the four experiments span two qualitatively distinct regimes, grouped into the egenioussBench Extended and House of Science experimental pairs. At far viewing distances the aerial and supermesh occupy a similar PDS range, isolating the marginal contribution of supplementary ground data at scale. At close range, the aerial and ground meshes diverge substantially, enabling analysis of the accuracy degradation incurred when the map resolution is mismatched to the query scale.

3.4 Pose estimation process

We adopt the localisation framework proposed by (Vultaggio et al., 2024) to localise terrestrial smartphone query images against mesh data; see Fig. 5. The pipeline comprises two main stages. In the offline stage, road network data is used to sample viewpoints from within the mesh; the resulting rendered views are then processed to extract both global (Berton et al., 2024) and local descriptors, which are subsequently used during the online localisation phase.

The online localisation stage proceeds in two steps. In the initialisation step, a multi-stage filtering approach is applied: first, the most visually similar database images are retrieved via *Global Feature Retrieval* (GFR) (Berton et al., 2024); these candidates are then spatially filtered to retain only those closest to the query's pose prior, *Spatial Nearest Neighbour re-ranking* (SNNr). The resulting shortlist of the most promising views is then matched against the query image using learned image matchers, Local Features Matching (LFM). In the subsequent *Relative Pose Estimation* (RPE) step, image-to-image correspondences are lifted to 3D using the depth buffers obtained during rendering. The resulting 2D-3D correspondences are used to estimate the camera pose via PnP (Vultaggio et al., 2025) within a RANSAC (Baráth et al., 2019) loop, after which the pose is refined via non-linear optimisation (Agarwal et al., 2023).

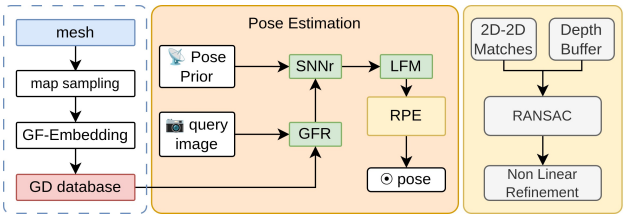

Figure 5. Overview of the visual localisation system. The mesh is sampled within the area of interest to generate a database of rendered views, which are used for retrieval and pose estimation against a query image.

In our experiments we evaluate multiple image matching strategies, covering both sparse —XFeat(Potje et al., 2024) and SuperPoint(DeTone et al., 2018)— and dense —JamMa(Lu and Du, 2025) and RoMaV2 (Edstedt et al., 2025)— matchers. Sparse matchers detect a set of keypoints and compute a corresponding descriptor for each. Both sparse models evaluated here employ CNN-based architectures: XFeat prioritises extraction speed, while SuperPoint serves as an established baseline. At matching time, the extracted descriptors are matched using LightGlue (Lindenberger et al., 2023), a transformer-based model that improves the robustness and accuracy of sparse feature matching by iteratively refining descriptors through alternating self- and cross-attention blocks while pruning unmatchable features.

We additionally evaluate two dense matchers: JamMa (Lu and Du, 2025) and RoMaV2 (Edstedt et al., 2025). Both networks take an image pair as input and produce a dense correspondence map. They operate by extracting dense feature representations from the input pair, which are then processed by a series of transformer blocks that generate and iteratively refine the correspondence set. Dense matchers are generally more robust than sparse approaches, albeit at the cost of higher inference time. JamMa addresses this trade-off by replacing conventional attention layers with a Mamba-based (Dao and Gu, 2024) architecture. Mamba was originally proposed in the Natural Language Processing (NLP) community to reduce the quadratic complexity of standard self-attention; a thorough treatment of state space models is beyond the scope of this work, and we refer the interested reader to (Zhang et al., 2024a) for a survey of Mamba models in vision applications.

Finally, we note that sparse features for the pre-rendered database images can be computed offline and cached, whereas dense matchers require the full image pair at inference time and therefore cannot benefit from this pre-computation.

4. Experiments

4.1 egenioussBench Extended

Table 3 compares the localisation performance of the aerial mesh and the Supermesh on the egenioussBench query set, and Fig. 6 summarises the accuracy–efficiency trade-off for each configuration.

Augmenting the aerial mesh with ground imagery and LiDAR yields a consistent improvement across all matchers and metrics. The most immediate effect is visible in the outlier rate, which drops by roughly one third: from 31.0% to 19.0% for XFeat and from 31.0% to 16.7% for SuperPoint. Similar gains are reflected in the threshold recall: at the 5 m, 10° threshold, XFeat improves from 69.0% to 81.0% and SuperPoint from

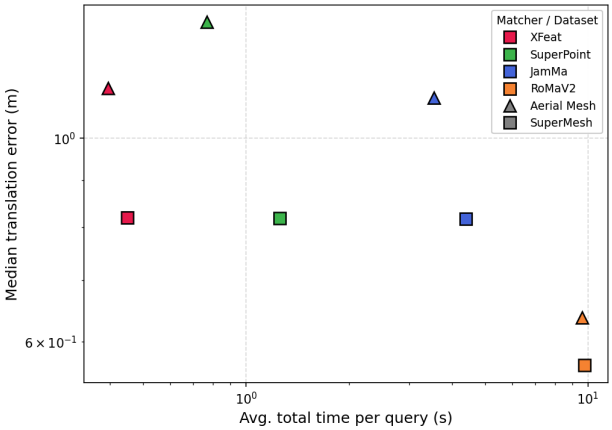

Figure 6. Median translation error versus average total localisation time for the egenioussBench Extended experiments. Marker shape denotes the mesh; colour denotes the matcher.

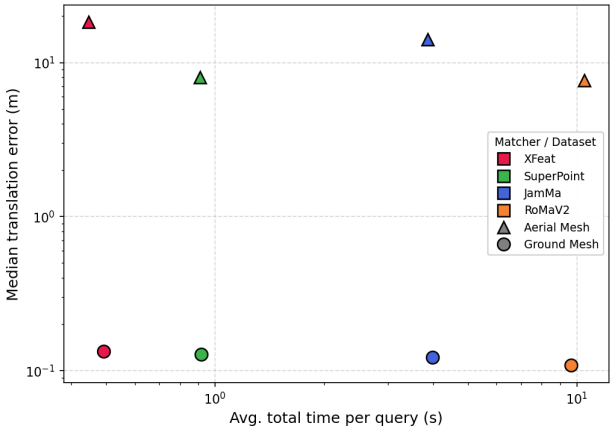

Figure 7. Median translation error versus average total query time for the House of Science experiments. Marker shape denotes the mesh; colour denotes the matcher.

69.0% to 83.3%. These improvements can be attributed to the enhanced façade detail provided by the ground-level data, which enables a greater number of successful feature matches for images that would otherwise fall outside the convergence basin of the pose estimator.

When considering the median pose error, the Supermesh narrows the gap between matchers. With the aerial mesh, median translation errors range from 0.6 m (RoMaV2) to 1.3 m (SuperPoint); with the Supermesh, all four matchers fall within the 0.6–0.8 m band. A similar compression is observed in rotational error, where the Supermesh brings all matchers to 0.5–0.6°. RoMaV2 achieves the lowest median error in both mesh configurations (0.6 m, 0.5° on the Supermesh) and the lowest RMSE among successfully localised images, but its matching stage is approximately 15–30× slower than the sparse alternatives. As shown in Fig. 6, in the egenioussBench Extended experiments, XFeat on the Supermesh achieves a median error within 35% of RoMaV2 while requiring less than 5% of the computation time, positioning it as the most practical choice when latency is constrained.

4.2 House of Science

Table 4 reports the results for the House of Science query set, and Fig. 7 and 6 show the corresponding accuracy-efficiency plot.

The aerial mesh is unable to adequately resolve façade geometry at short viewing distances, and the localisation pipeline fails for the majority of queries regardless of the matcher employed. Outlier rates range from 58.3% (RoMaV2) to 93.8% (XFeat), and even the best-performing configuration (RoMaV2) localises only 41.7% of images within 5 m, 10°. This is consistent with the low PDS predicted for the aerial mesh at close range (Fig. 3): the projected texel and triangle sizes are too coarse to produce reliable correspondences at sub-metre standoff distances.

Replacing the aerial mesh with the high-resolution ground mesh reverses this outcome. All matchers achieve outlier rates below 6.2%, and the tightest threshold (0.2 m, 1°) is met by 79.2–97.9% of queries. Median errors collapse to 0.1 m in translation for all matchers, with rotational errors of 0.2–0.3°. This represents a two-order-of-magnitude reduction in median translation error relative to the aerial mesh, confirming that in the House of Science experiments, mesh resolution is the dominant factor governing localisation accuracy. As Fig. 7 shows, once the mesh is adequate all matchers converge to the same accuracy floor, suggesting that below a certain PDS threshold the matcher choice is secondary to the map quality. RoMaV2 achieves the only zero-outlier result (100% recall at all thresholds) and the lowest RMSE (0.1 m, 0.3°), but the practical margin over the sparse alternatives is small.

5. Discussion

Across both operating regimes, improving mesh quality produces gains that no matcher upgrade alone can deliver and does so without increasing the latency, see Fig.7 and 6. In the egenioussBench Extended experiments, the Supermesh reduces outlier rates by up to 14 percentage points and compresses the accuracy spread across matchers, while in the House of Science experiments, the ground mesh transforms an essentially unusable configuration into sub-decimetre localisation. These results corroborate the PDS analysis of Sec. 3: when the mesh detail is well matched to the viewing distance the choice of matcher has a second-order effect, whereas when the mesh is under-resolved even the most robust matcher cannot compensate for the missing geometric and textural information. As shown in Fig. 4, the PDS is a strong predictor of the final pose error across all 16 experiment–matcher combinations.

The practical implication is that investments in reference data quality yield higher returns than investments in matcher sophistication. In the egenioussBench Extended experiments, the Supermesh closes the accuracy gap between a lightweight sparse matcher (XFeat) and the most expensive dense matcher (RoMaV2) to within 0.2 m, making the former the preferred choice under latency constraints. In the House of Science experiments, the matcher choice becomes almost irrelevant once the mesh resolution is adequate: all four matchers achieve sub-decimetre median errors on the ground mesh.

A limitation of the PDS metric in its current form is that it requires knowledge of the viewing distance d. In our experiments, d was estimated from depth maps computed during ground-truth processing with PIX4Dmatic, independently of the localisation pipeline. In a deployment scenario, d could be approximated from a coarse pose prior (e.g. GNSS height above

Matchers	0.5,2 / 2,5 / 5,10 [m,°] [1] ↑ %,%,%	Outliers [2] ↓ %	ME [3] ↓ m,°	RMSE [4] ↓ m,°	T_m [5] ↓ s	T_t [6] ↓ s
Aerial Mesh						
XFeat	19.05 / 61.90 / 69.05	30.95	1.13 , 0.77	1.38 , 1.21	**0.29**	**0.39**
SuperPoint	14.29 / 57.14 / 69.05	30.95	1.34 , 0.74	1.81 , 1.54	0.59	0.77
JamMa	14.29 / 66.67 / 73.81	**26.19**	1.11 , 0.66	1.30 , 1.06	3.37	3.54
RoMaV2	**35.71 / 69.05 / 69.05**	30.95	**0.64 , 0.61**	**0.62 , 0.55**	9.25	9.62
Supermesh						
XFeat	28.57 / 71.43 / 80.95	19.05	0.82 , 0.61	1.30 , 1.21	**0.30**	**0.45**
SuperPoint	26.19 / 73.81 / 83.33	**16.67**	0.82 , 0.59	1.18 , 1.17	0.61	1.25
JamMa	23.81 / 71.43 / 80.95	19.05	0.82 , 0.52	1.09 , 0.77	3.34	4.40
RoMaV2	**38.10 / 78.57 / 78.57**	21.43	**0.57 , 0.54**	**0.61 , 0.58**	9.08	9.76

Table 3. Quantitative comparison of visual localisation results in the egenioussBench extended dataset using the aerial mesh and Supermesh which integrates ground imagery and lidar data in the mesh generation process. The table presents: [1] The percentage of query images successfully localised within three pose error thresholds —0.5m, 2°, 2m, 5°, and 5m, 10°; [2] the percentage of images localised beyond the last threshold (5 m, 10°); [3] the Median Error (ME) across all images; [4] the root Mean Squared Error (RMSE) of the images successfully localised within the last pose error threshold; [5] the mean execution time for the image matching stage; [6] the mean total time required to compute the query pose.

Matchers	0.2,1 / 0.5,2 / 2,5 / 5,10 [m,°] [1] ↑ %,%,%,%	Outliers [2] ↓ %	ME [3] ↓ m,°	RMSE [4] ↓ m,°	T_m [5] ↓ s	T_t [6] ↓ s
Aerial Mesh						
XFeat	0.00 / 0.00 / 2.08 / 6.25	93.75	18.33 , 69.94	3.06 , 1.62	**0.30**	**0.45**
SuperPoint	0.00 / 6.25 / 10.42 / 29.17	70.83	**8.03 , 10.59**	2.76 , 4.13	0.59	0.91
JamMa	0.00 / 0.00 / 8.33 / 12.50	87.50	14.20 , 56.60	1.87 , 3.43	3.59	3.87
RoMaV2	**0.00 / 8.33 / 37.50 / 41.67**	**58.33**	7.64 , 29.69	**1.13 , 2.66**	9.68	10.48
Ground Mesh						
XFeat	79.17 / 87.50 / 91.67 / 93.75	6.25	0.13 , 0.34	0.31 , 1.25	**0.36**	**0.49**
SuperPoint	89.58 / 93.75 / 93.75 / 93.75	6.25	0.13 , 0.31	0.13 , 0.39	0.70	0.92
JamMa	93.75 / 93.75 / 95.83 / 95.83	4.17	0.12 , 0.26	0.13 , 0.51	3.68	3.98
RoMaV2	**97.92 / 100.00 / 100.00 / 100.00**	**0.00**	**0.11 , 0.21**	**0.11 , 0.34**	9.25	9.64

Table 4. Quantitative comparison of visual localisation results in the House of Science dataset using an aerial mesh and a high resolution ground mesh. The table presents: [1] The percentage of query images successfully localised within three pose error thresholds — 0.2m,1°, 0.5m, 2°, 2m, 5°, and 5m, 10°; [2] the percentage of images localised beyond the last threshold (5 m, 10°); [3] the Median Error (ME) across all images; [4] the root Mean Squared Error (RMSE) of the images successfully localised within the last pose error threshold; [5] the mean execution time for the image matching stage; [6] the mean total time required to compute the query pose.

the mesh surface), making PDS useful as a planning tool for determining where existing mesh data are sufficient and where higher-resolution acquisitions would yield meaningful accuracy improvements.

6. Conclusions

This work presented the first systematic analysis of how the provenance and resolution of the input data used to construct a mesh map affect the accuracy and robustness of mesh-based visual localisation. By evaluating three meshes—aerial, combined aerial and ground (Supermesh), and close-range ground—across the egenioussBench Extended and House of Science query sets and four image matchers, we showed that mesh quality is the dominant factor in localisation performance, outweighing the choice of matcher.

In the egenioussBench Extended experiments, augmenting an aerial mesh with ground imagery and LiDAR reduced outlier rates by up to 14 percentage points and compressed the median error spread across all matchers to within 0.2 m. In the House of Science experiments, replacing the aerial mesh with a high-resolution ground mesh reduced median translation errors by two orders of magnitude, from metres to sub-decimetre level. We further introduced the Perceptual Detail Score (PDS), a viewing-condition-aware metric that proved to be a strong predictor of downstream pose accuracy across all 16 experiment–matcher combinations.

These findings carry direct implications for the deployment of mesh-based localisation systems: in the egenioussBench Extended scenario, standard aerial oblique meshes augmented with ground data provide accuracy competitive with the most expensive dense matchers, while in the House of Science experiments, high-resolution ground meshes are essential regardless of the matcher employed. Future work will extend this analysis to additional scenes and investigate the use of PDS as a planning tool for targeted data acquisition.

Acknowledgements

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the EUSPA. Neither the European Union nor the EUSPA can be held responsible for them.

We thank Cyclomedia for the ground imagery and LiDAR data used to generate the Supermesh and Skyline for the support in the processing of the data.

References

Agarwal, S., Mierle, K., Team, T. C. S., 2023. Ceres Solver.

Baráth, D., Noskova, J., Ivashechkin, M., Matas, J., 2019. MAGSAC++, a Fast, Reliable and Accurate Robust Estimator. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13–19. doi: 10.1109/CVPR42600.2020.00138.

Berton, G., Junglas, L., Riccardo Zaccone, Thomas Pollok, B. C., Masone, C., 2024. Meshvpr: Citywide visual place recognition using 3d meshes. *European Conference on Computer Vision (ECCV)*.

Berton, G., Masone, C., Caputo, B., 2022. Rethinking visual geo-localization for large-scale applications. *IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*, 4878–4888. doi: 10.1109/CVPR52688.2022.00483.

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 (CVPR)*, 27424–27434.

Brachmann, E., Cavallari, T., Prisacariu, V. A., 2023. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. *IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*. doi: 10.1109/CVPR52729.2023.00488.

Chen, C., Wang, B., Lu, C. X., Trigoni, N., Markham, A., 2023. Deep learning for visual localization and mapping: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Dao, T., Gu, A., 2024. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. *IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*, 224–236. doi: 10.1109/CVPRW.2018.00060.

Edstedt, J., Nordström, D., Zhang, Y., Bökman, G., Astermark, J., Larsson, V., Heyden, A., Kahl, F., Wadenbäck, M., Felsberg, M., 2025. RoMa v2: Harder Better Faster Denser Feature Matching.

Fanta-Jende, P., Vultaggio, F., Kern, A., Loeper, Y., Gerke, M., 2025. egenioussBench: A New Dataset for Geospatial Visual Localisation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-1-W6-2025, 75–82. doi: 10.5194/isprs-archives-XLVIII-1-W6-2025-75-2025.

Jiang, X., Wang, F., Galliani, S., Vogel, C., Pollefeys, M., 2025. R-SCoRe: Revisiting Scene Coordinate Regression for Robust Large-Scale Visual Localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11536–11546.

Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K. M., Scherer, S., Krishna, M., Garg, S., 2023. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*. doi: 10.1109/LRA.2023.3343602.

Kendall, A., Grimes, M., Cipolla, R., 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. *Proceedings of the IEEE international conference on computer vision*, 2938–2946.

Khatib, F., Moran, D., Trostianetsky, G., Kasten, Y., Galun, M., Basri, R., 2025. GSVisLoc: Generalizable Visual Localization for Gaussian Splatting Scene Representations. *arXiv*. doi: 10.48550/arXiv.2508.18242.

Lepetit, V., Moreno-Noguer, F., Fua, P., 2009. EPnP: An Accurate O(n) Solution to the PnP Problem. *International Journal of Computer Vision*, 81(2), 155–166. doi: 10.1007/s11263-008-0152-6.

Lindenberger, P., Sarlin, P.-E., Pollefeys, M., 2023. LightGlue: Local Feature Matching at Light Speed. *IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 01–06. doi: 10.1109/ICCV51070.2023.01616.

Loeper, Y., Gerke, M., Alamouri, A., Kern, A., Bajauri, M. S., Fanta-Jende, P., 2024. Visual localization in urban environments employing 3D city models. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2-W8-2024, 311–318. doi: 10.5194/isprs-archives-XLVIII-2-W8-2024-311-2024.

Lu, X., Du, S., 2025. JamMa: Ultra-lightweight Local Feature Matching with Joint Mamba. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14934–14943.

Mera-Trujillo, M., Smith, B., Fragoso, V., 2020. Efficient Scene Compression for Visual-based Localization. *2020 International Conference on 3D Vision (3DV)*, 1–10. doi: 10.1109/3DV50981.2020.00111.

More, H., Cianca, E., De Sanctis, M., 2024. Comparing Positioning Performance of LEO Mega-Constellations and GNSS in Urban Canyons. *IEEE Access*, 12, 24465–24482. doi: 10.1109/ACCESS.2024.3364819.

Nadeem, U., Jalwana, M. A., Bennamoun, M., Togneri, R., Sohel, F., 2019. Direct image to point cloud descriptors matching for 6-dof camera localization in dense 3d point clouds. *International Conference on Neural Information Processing*, Springer, 222–234.

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. doi: 10.1109/LRA.2020.3007120. https://ieeexplore.ieee.org/abstract/document/9133282. 5 citations (Semantic Scholar/DOI) [2023-10-18] Conference Name: IEEE Robotics and Automation Letters.

Panek, V., Kukelova, Z., Sattler, T., 2022. MeshLoc: Mesh-Based Visual Localization. *European Conference on Computer Vision ECCV*, 589–609, doi: 10.1007/978-3-031-20047-2_34.

Panek, V., Kukelova, Z., Sattler, T., 2023. Visual localization using imperfect 3d models from the internet. *IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*, 13175–13186.

Panek, V., Zhou, Q., Ding, Y., Agostinho, S., Kukelova, Z., Sattler, T., Leal-Taixé, L., 2025. A Guide to Structureless Visual Localization. *arXiv*. doi: 10.48550/arXiv.2504.17636.

Pix4D SA, 2024. Pix4Dmapper. `https://www.pix4d.com`.

Potje, G., Cadar, F., Araujo, A., Martins, R., Nascimento, E. R., 2024. Xfeat: Accelerated features for lightweight image matching. *IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*. doi: 10.1109/CVPR52733.2024.00259.

Radoš, K., Brkić, M., Begušić, D., 2024. Recent Advances on Jamming and Spoofing Detection in GNSS. *Sensors*, 24(13), 4210. doi: 10.3390/s24134210.

Sarlin, P.-E., Cadena, C., Siegwart, R., Dymczyk, M., 2019. From coarse to fine: Robust hierarchical localization at large scale. *IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*. doi: 10.1109/CVPR.2019.01300.

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. *European Conference on Computer Vision*, Springer, 686–704.

Sarlin, P.-E., Trulls, E., Pollefeys, M., Hosang, J., Lynen, S., 2023. SNAP: Self-Supervised Neural Maps for Visual Positioning and Semantic Understanding. *Advances in Neural Information Processing Systems*, 36, 7697–7729.

Sattler, T., Leibe, B., Kobbelt, L., 2012. Improving Image-Based Localization by Active Correspondence Search. *European Conference on Computer Vision (ECCV)*, 752–765, doi: 10.1007/978-3-642-33718-5_54.

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J. et al., 2018. Benchmarking 6dof outdoor visual localization in changing conditions. *IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*, 8601–8610. doi: 10.1109/CVPR.2018.00897.

Schubert, S., Neubert, P., Garg, S., Milford, M., Fischer, T., 2023. Visual place recognition: A tutorial [tutorial]. *IEEE Robotics & Automation Magazine*, 31(3), 139–153.

Speciale, P., Schonberger, J. L., Kang, S. B., Sinha, S. N., Pollefeys, M., 2019. Privacy Preserving Image-Based Localization. [Online; accessed 30. Oct. 2025].

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. *IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*. doi: 10.1109/TPAMI.2019.2952114.

Vultaggio, F., Fanta-Jende, P., Gerke, M., 2025. Perspective-n-Point in Practice: Performance, Robustness, and Accuracy for Mesh-Based Localisation. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-1-W4-2025, 131–138. doi: 10.5194/isprs-archives-XLVIII-1-W4-2025-131-2025.

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*, 48, 447–454.

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. *IEEE International Conference on Multimedia and Expo (ICME)*, 2171–2176. doi: 10.1109/ICME55011.2023.00371.

Zhang, H., Zhu, Y., Wang, D., Zhang, L., Chen, T., Wang, Z., Ye, Z., 2024a. A Survey on Visual Mamba. *Applied Sciences*, 14(13), 5683. doi: 10.3390/app14135683.

Zhang, L., Tao, Y., Lin, J., Zhang, F., Fallon, M., 2024b. Visual localization in 3d maps: Comparing point cloud, mesh, and nerf representations. *arXiv preprint arXiv:2408.11966*.