Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Accuracy potential of visual localization exploiting high-end street-level imagery

**Jonas Meyer**

jonas.meyer@fhnw.ch

FHNW University of Applied Sciences and Arts Northwestern Switzerland

**Stephan Nebiker**

FHNW University of Applied Sciences and Arts Northwestern Switzerland

**Pascal Theiler**

iNovitas AG

**Norbert Haala**

University of Stuttgart

**Research Article**

**Keywords:** Visual localization, Accuracy investigation, Dataset, Image retrieval, Feature matching, Structure-from-motion

**Posted Date:** November 23rd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11647096/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Accuracy potential of visual localization exploiting high-end street-level imagery

**Jonas Meyer · Stephan Nebiker · Pascal Theiler · Norbert Haala**

Received: date / Accepted: date

**Abstract** Many applications, such as surveying, autonomous navigation, robotics and augmented reality, require accurate pose information with respect to a reference frame. Visual localization can complement GNSS positioning, whose availability and accuracy are often limited. However, the accuracy potential of visual localization has not yet been systematically investigated against survey-grade requirements, and no publicly available large-scale outdoor dataset provides ground-truth poses accurate enough for such an investigation. We address both gaps. First, we present the *FHNW Muttenz* dataset. It covers a contiguous 10 km street network, mapped in two mobile mapping campaigns about 1.5 years apart, and comprises high-resolution reference imagery as well as query sequences from four cameras in five scenes. The two campaigns are co-registered to each other with an accuracy of 5 mm. Co-registration of the query images to the reference imagery yields ground-truth poses accurate to 6–9 mm in position and well below 0.1° in orientation. All images are georeferenced with an accuracy of about 2.5 cm. Second, we use this dataset to systematically evaluate sparse feature-based visual localization with a modular pipeline built from established components, using the high-resolution reference imagery directly as the scene representation. For daytime imagery and pre-calibrated cameras, the median pose errors with respect to the reference imagery are 1–5 cm in translation and 0.03–0.1° in rotation. Given an accurately georeferenced scene representation, visual localization not only approaches the accuracy of survey-grade GNSS positioning but also provides full orientation information. The dataset is publicly available at `https://fhnw-muttenz-vl-dataset.github.io/`.

**Keywords** Visual localization · Accuracy investigation · Dataset · Image retrieval · Feature matching · Structure-from-motion

J. Meyer · S. Nebiker
Institute of Geomatics, FHNW University of Applied Sciences and Arts Northwestern Switzerland, 4132 Muttenz, Switzerland
J. Meyer (corresponding author): E-mail: jonas.meyer@fhnw.ch
S. Nebiker: E-mail: stephan.nebiker@fhnw.ch

J. Meyer · N. Haala
Institute for Photogrammetry and Geoinformatics, University of Stuttgart, 70174 Stuttgart, Germany
N. Haala: E-mail: norbert.haala@ifp.uni-stuttgart.de

P. Theiler
iNovitas AG, 5400 Baden, Switzerland
P. Theiler: E-mail: pascal.theiler@inovitas.ch

## 1 Introduction

Estimating the 6 degree-of-freedom (DoF) pose of a sensor with respect to a reference frame is a prerequisite for many applications, including (autonomous) navigation, surveying, robotics, and augmented reality (AR). Typically, GNSS sensors are used in combination with IMUs for this task. Since this procedure heavily relies on good GNSS reception, it is applicable only outdoors. Even there, occlusions and multi-path effects significantly limit the applicability of GNSS in complex urban environments. Visual localization, which estimates the 6-DoF pose of an image with respect to a previously mapped scene, provides a complementary positioning modality: objects that occlude GNSS signals typically exhibit a high density of salient visual features.

Over the last two decades, many visual localization methods have been proposed, which can be divided into 2D image-based and 3D geometry-based approaches. 2D

image-based methods represent the scene as a database of oriented images, from which image retrieval selects the reference images most similar to the query (Sivic and Zisserman 2003). 3D geometry-based methods, in contrast, store the scene geometry in a globally consistent manner. This allows them to establish 2D–3D correspondences and to estimate the query pose with a Perspective-n-Point (PnP) solver (Haralick et al. 1994) within a RANSAC framework (Fischler and Bolles 1981). Because no 3D geometry has to be computed or stored, 2D image-based approaches are more scalable and flexible, whereas 3D geometry-based approaches benefit from more extensive geometric reasoning and generally achieve higher accuracy (Panek et al. 2026). In both cases, the performance further depends on how much the query conditions differ from those during reference mapping (Sattler et al. 2018). Yet scalability and accuracy are both needed in practice: to complement GNSS effectively and bridge GNSS-denied areas, visual localization must scale to entire cities or regions without sacrificing accuracy.

Visual localization methods in outdoor scenarios are generally considered accurate if they achieve translation errors less than 25 cm and rotation errors less than 2° (Sattler et al. 2018). This level of accuracy is sufficient for autonomous driving or AR applications (Sarlin et al. 2019). However, to complement survey-grade GNSS positioning, which achieves accuracies of a few centimeters (approximately 5 cm in 3D) under favorable conditions, visual localization should yield comparable results. This requirement applies at two levels: firstly, the pose relative to the scene representation, and secondly, the georeferencing of the scene representation itself, which bounds the achievable absolute accuracy.

To the best of our knowledge, there exists no study that systematically evaluates the accuracy potential of visual localization in street scenes with respect to survey-grade accuracy demands. This is a direct consequence of the lack of publicly available large-scale datasets that contain sufficiently accurate ground-truth poses. In this work, our objective is to address both research gaps. We investigate the following question: what pose accuracy relative to precisely georeferenced street-level reference imagery can sparse feature-based visual localization achieve under favorable conditions? We refer to this as the accuracy potential of visual localization and address it through four sub-questions: (i) how does the selection of reference candidates, guided by pose priors, affect the pose accuracy; (ii) how do image resolution and the choice of local features affect it; (iii) how does it depend on the query camera and on the epoch of the reference imagery; and (iv) how much accuracy is lost by using 2D image-based visual localization compared to a 3D geometry-based method?

Answering these questions requires ground-truth poses whose uncertainty is well below the expected localization errors, which no existing large-scale outdoor dataset provides (Section 2.2). We therefore created a dataset in which reference imagery from two mobile mapping campaigns and query sequences from four different cameras are co-registered with quantified uncertainty. On this dataset, we evaluate the accuracy potential of visual localization using a modular visual localization pipeline built from established components. It uses the precisely georeferenced high-resolution imagery directly as scene representation, selects reference candidates guided by pose priors, and estimates the query pose from a local Structure-from-Motion (SfM) model reconstructed on the fly. Reference imagery acquired by professional vision-based mobile mapping systems is what we refer to as high-end street-level imagery. Our main contributions are summarized as follows:

- The FHNW Muttenz dataset[1]: high-resolution reference imagery of a contiguous 10 km street network from two mobile mapping campaigns about 1.5 years apart, and query sequences from four cameras in five scenes. The two campaigns are co-registered with a 3D RMSE of 5 mm at independent tie points. Co-registration of the query images to the reference imagery yields ground-truth poses accurate to 6–9 mm in position and well below 0.1° in orientation. A sensitivity analysis confirms that these poses are determined by the image observations rather than by the ground control point configuration.
- A systematic evaluation of the accuracy potential of sparse feature-based visual localization with respect to reference candidate selection, image resolution, local features, query camera and reference epoch, including a comparison with a representative 3D geometry-based visual localization method. Evaluations at error thresholds down to 1 cm and 0.1° reveal differences between local features that are not resolved at the coarse thresholds of existing benchmarks.

The remainder of this article is organized as follows. Section 2 presents related work on visual localization methods and on the datasets used to evaluate them. Section 3 introduces the dataset and the visual localization pipeline. Section 4 reports the experimental setup and the results. Section 5 discusses the results and the limitations of this study and identifies directions for future research. Section 6 concludes the article.

[1] `https://fhnw-muttenz-vl-dataset.github.io/`

## 2 Related work

### 2.1 Visual localization

As outlined in the introduction, visual localization methods can be categorized into *2D image-based* and *3D geometry-based* methods according to their scene representation. The latter are commonly referred to as 3D structure-based (Sattler et al. 2018), which implies that the 3D geometry is derived from SfM. Since many current 3D scene representations are not derived from SfM, we use the term 3D geometry-based instead. As image retrieval (IR) is crucial for both categories, we begin with a brief summary of this area.

#### *2.1.1 Image retrieval*

IR retrieves images from a database that depict the same scene as a given query image, either to select reference candidates or to approximate the query pose. The image content is typically encoded as a global image descriptor, and similar images are retrieved via a nearest neighbor search in the descriptor space. As shown by Humenberger et al. (2022), the appropriate image description method strongly depends on the subsequent pose estimation technique. IR methods originate from both the landmark retrieval (LR) and visual place recognition (VPR) domains, but they are typically tailored to specific environments and tasks and therefore lack generalization (Keetha et al. 2024; Berton and Masone 2025).

Early methods based on global image statistics were replaced by methods that aggregate local features (Jegou et al. 2010). NetVLAD (Arandjelovic et al. 2016) established a first baseline for deep global image descriptors, and subsequent work improved on it through novel losses, large-scale datasets and high-level CNN features while yielding smaller descriptors (Yu et al. 2020; Ali-Bey et al. 2022, 2023). CosPlace adopted a training strategy inspired by face recognition and achieved state-of-the-art results (Berton et al. 2022). More recently, vision foundation models such as DINOv2 (Oquab et al. 2024) have been used to design universal solutions (Keetha et al. 2024; Izquierdo and Civera 2024), at the cost of considerably longer descriptors. Most trained VPR models, however, learn from positives that share the viewpoint of the query image and therefore often fail to retrieve reference images that depict the same scene from different viewpoints.

EigenPlaces (Berton et al. 2023) addresses this limitation with a viewpoint-aware data sampling technique, while also producing compact descriptors. MegaLoc (Berton and Masone 2025) combines this sampling with training on several task-specific datasets and outperforms state-of-the-art VPR and LR models in retrieving reference images with diverse viewpoints. In large-scale environments, retrieval can additionally be constrained by pose priors from mobile sensors, such as GNSS positions and headings (Yan et al. 2023).

#### *2.1.2 2D image-based visual localization*

Visual localization methods relying on this representation can be further subdivided into pose approximation, relative pose estimation and regression, and local SfM model construction. In pose approximation, the pose of the query image is adopted from the most similar reference image or interpolated from the poses of the top-$k$ retrieved images (Torii et al. 2015; Sattler et al. 2019). Relative pose estimation first establishes 2D–2D correspondences between the query and the retrieved images and estimates their relative geometry. Since the reference poses are known, the absolute pose of the query image can then be derived (Zhou et al. 2020; Dong et al. 2023). Relative pose regression, in contrast, directly regresses the pose between two input images (Laskar et al. 2017; Zhou et al. 2020). Recent feed-forward 3D reconstruction models (Wang et al. 2024; Leroy et al. 2025; Wang et al. 2025) show competitive performance in image matching (Bellavia et al. 2025) and visual localization, and can estimate the query pose even with little visual overlap, where other methods fail (Panek et al. 2026).

Building a local SfM model on the fly from the top-$k$ retrieved images and estimating the query pose from 2D–3D matches (Sattler et al. 2018; Torii et al. 2021; Humenberger et al. 2022) combines the properties of 2D image-based and 3D geometry-based methods (Panek et al. 2026). According to Panek et al. (2026), such approaches achieve the highest pose accuracies among 2D image-based methods owing to their stronger geometric reasoning, but remain slightly less accurate than 3D geometry-based methods. In addition, triangulating 3D points on the fly causes considerable computational overhead.

ImLoc (Jiang et al. 2026) is, to our knowledge, the first 2D image-based approach that surpasses state-of-the-art methods, including those that explicitly exploit 3D scene geometry. It augments the reference images with depth maps and lifts dense RoMa correspondences (Edstedt et al. 2024) to 2D–3D correspondences, thereby enabling geometric reasoning without on the fly reconstruction while keeping the scene representation simple. Dense matching, however, is computationally demanding at high image resolutions (Section 3.2.2).

*2.1.3 3D geometry-based visual localization*

Traditionally, scenes for 3D geometry-based visual localization are represented by SfM point clouds (Arth et al. 2009; Irschara et al. 2009; Sattler et al. 2018), and the camera pose is estimated from 2D–3D correspondences obtained by matching query features against the descriptors of the 3D points. To cope with scalability and descriptor ambiguity in large scenes, hierarchical approaches confine matching to the 3D points visible in the top-$k$ retrieved reference images (Sattler et al. 2012; Sarlin et al. 2019). Besides SfM point clouds, meshes (Panek et al. 2022, 2023; Vultaggio et al. 2024), 3D city models (Majdik et al. 2014; Loeper et al. 2024), 3D Gaussian Splatting (Matteo et al. 2025; Zhai et al. 2025) and Neural Radiance Fields (NeRFs) (Liu et al. 2023; Zhou et al. 2024) serve as scene representations, while the pose is still estimated from 2D–3D correspondences derived from feature matching.

Scene coordinate regression methods instead train a network to directly regress the 3D coordinates of image patches, which are then used for RANSAC-based PnP pose estimation (Shotton et al. 2013; Brachmann and Rother 2018; Revaud et al. 2024). Absolute pose regression methods predict the query pose directly from the image (Kendall et al. 2015; Shavit et al. 2021), but most of them are not substantially more accurate than pose approximation and generalize poorly to unseen views (Sattler et al. 2019). Map-relative variants improve generalization to new scenes (Chen et al. 2024b). Render-and-compare methods iteratively refine a coarse initial pose, for example from IR, by aligning the query image with renderings of the scene representation (Yen-Chen et al. 2021; Sarlin et al. 2021; Chen et al. 2024a), but remain inferior to hierarchical feature-based methods in accuracy and robustness (Panek et al. 2026).

All these methods exploit the 3D geometry of the scene, represented explicitly or implicitly, in a globally consistent manner. Precomputing this geometry is time-consuming and resource-intensive, and ensuring global consistency is particularly challenging in large-scale environments (Sattler et al. 2018). Once built, such representations are static and costly to update when the scene changes. Moreover, meshes, city models and learned representations abstract geometric and visual information, which degrades the accuracy of the estimated pose (Panek et al. 2023; Bieringer et al. 2024; Gaisbauer et al. 2025; Sattler et al. 2019).

## 2.2 Datasets

Visual localization datasets are typically collected with specific applications and tasks in mind and therefore vary in scene type, sensor configuration and intended use. Table 1 provides an overview of popular datasets and benchmarks currently used in the field. Challenging conditions, which can severely degrade the performance of visual localization methods, occur more frequently outdoors than indoors, where they are largely limited to dynamics and changing camera intrinsics.

Reliable evaluation requires precise ground-truth poses, whose estimation is itself limited by challenging conditions and by the spatial extent of the scene. For small- to medium-scale environments, a common strategy is to register the reference and query sequences against high-precision LiDAR data (ETH3D (Schops et al. 2017), NAVER Labs (Lee et al. 2021), LaMAR (Sarlin et al. 2022), CroCoDL (Blum et al. 2025)). Matches to the scanner's panoramic images or to rendered views of the colorized point cloud are lifted to 2D–3D correspondences through the associated 3D points, which then drive pose estimation and refinement (Schops et al. 2017; Taira et al. 2021; Sarlin et al. 2022). Alternatively, multi-sensor platforms are combined with SLAM algorithms, whose maps provide an initial alignment that is refined in post-processing (Yan et al. 2023; Huang et al. 2024). Large-scale outdoor datasets such as San Francisco (Chen et al. 2011), CMU Seasons (Badino et al. 2011; Sattler et al. 2018) and RobotCar (Maddern et al. 2017; Sattler et al. 2018) are typically acquired with mobile mapping systems (MMS) (Puente et al. 2013). Their ground-truth poses are usually obtained by orienting individual sequences via SfM and co-registering them via ICP or manually measured tie points, incorporating relative sensor orientations and approximate poses from direct sensor orientation where available (Chen et al. 2011; Sattler et al. 2018).

Although query images are often captured with cameras that differ from those used for reference mapping, cross-device localization has long been neglected. Most datasets reproject images from non-perspective cameras to the classical pinhole model (Carlevaris-Bianco et al. 2016; Maddern et al. 2017; Wang et al. 2026). 360Loc (Huang et al. 2024) provides query images following the perspective, fisheye and spherical camera models, but its pinhole and fisheye images are generated with virtual cameras rather than recorded by physical devices.

High-resolution imagery is also underrepresented. Since computational resources and latency are often crucial in visual localization applications, only ETH3D (Schops et al. 2017), InLoc (Taira et al. 2021), NAVER Labs (Lee et al. 2021) and 360Loc (Huang et al. 2024) provide high-resolution images, none of which covers a large-scale outdoor environment. Moreover, most datasets consist of several non-overlapping submaps. While this design is reasonable when aggregating dis-

**Table 1** Comparison of visual localization datasets: environment (Env.), number of locations (# Loc.), spatial scale, challenging conditions, camera models, high-resolution imagery, and accuracy of the image poses. Challenges: weather (W), season (S), strong viewpoint changes (V), day/night (N), permanent or temporary construction (C), dynamics such as people, vehicles or rearranged furniture (D), and camera intrinsics (I). Camera models: perspective (P), fisheye (F), spherical (360). Pose accuracy reflects relative accuracy to the reference data, as reported in the original publications.

| Dataset | Env. | # Loc. | Scale | Challenging conditions | | | | | | | Camera model | High res. | Pose |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | W | S | V | N | C | D | I | ref / query | images | acc. |
| 7-Scenes (Shotton et al. 2013) | I | 7 | S | | | | | | | | P / P | | ≈ cm |
| Baidu mall (Sun et al. 2017) | I | 1 | M | | | | | | ✓ | ✓ | P / P | | ≈ dm |
| InLoc (Taira et al. 2021) | I | 5 | M | | | | | | ✓ | | P / P | ✓ | > dm |
| NAVER Labs (Lee et al. 2021) | I | 5 | M | | | | | | ✓ | ✓ | P / P | ✓ | ≈ dm |
| NCLT (Carlevaris-Bianco et al. 2016) | I+O | 1 | M | ✓ | ✓ | | | ✓ | ✓ | | P / P | | ≈ dm |
| ETH3D (Schops et al. 2017) | I+O | 25 | S | | | | | | | | P / P | ✓ | ≈ mm |
| LaMAR (Sarlin et al. 2022) | I+O | 3 | M | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | P / P | | ≈ cm |
| 360Loc (Huang et al. 2024) | I+O | 4 | M | ✓ | | | ✓ | ✓ | ✓ | ✓ | 360 / P+F+360 | ✓ | ≈ cm |
| CroCoDL (Blum et al. 2025) | I+O | 10 | L | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | P / P | | ≈ dm |
| Oxford Day & Night (Wang et al. 2026) | I+O | 5 | L | | | ✓ | ✓ | | ✓ | | P / P | | < dm |
| San Francisco (Chen et al. 2011) | O | 1 | L | | | | | ✓ | ✓ | ✓ | P / P | | ≈ m |
| CMU Seasons (Badino et al. 2011; Sattler et al. 2018) | O | 17 | L | ✓ | ✓ | | | | ✓ | ✓ | P / P | | > dm |
| Aachen Day Night (Sattler et al. 2012, 2018) | O | 1 | L | ✓ | | | | | ✓ | ✓ | P / P | | > dm |
| Cambridge (Kendall et al. 2015) | O | 3 | S | ✓ | | | | | ✓ | | P / P | | > dm |
| RobotCar Seasons (Maddern et al. 2017; Sattler et al. 2018) | O | 49 | L | ✓ | ✓ | | ✓ | | ✓ | | P / P | | ≈ dm |
| CrowdDriven (Jafarzadeh et al. 2021) | O | 26 | M | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | P / P | | < dm |
| SenseLoc (Yan et al. 2023) | O | 1 | M | ✓ | ✓ | | ✓ | ✓ | ✓ | | P / P | | < dm |
| **FHNW Muttenz (Ours)** | O | 1 | L | | ✓ | ✓ | | ✓ | ✓ | ✓ | P / P+F | ✓ | < cm |

tinct environments and estimating ground-truth poses (Sattler et al. 2018), real-world environments cannot be partitioned in this way without introducing edge cases at the boundaries.

Long-term localization performance is likely to decrease as the scene gradually changes (Sattler et al. 2018; Zhang et al. 2026), which makes updating the reference map critical, regardless of the scene representation. However, available datasets usually provide a single reference sequence and therefore do not support investigations on how to update reference maps efficiently.

Further datasets address specific scene representations. The cadloc dataset (Panek et al. 2023) and egenioussBench (Fanta-Jende et al. 2025) provide benchmarks for mesh-based visual localization, relying on 3D models from the internet that represent reality only imperfectly and on geospatial meshes, respectively. The TUM2TWIN dataset (Wysocki et al. 2026) provides multi-sensor data for universal digital twins, co-registered with an accuracy in the upper centimeter to decimeter range, and is therefore suitable for investigations on scene representations such as city models or meshes.

In summary, a diverse set of datasets is available for developing and assessing visual localization methods. However, none of them combines a large-scale street-scene setting with ground-truth poses accurate enough to assess the accuracy potential of visual localization.

## 3 Materials and methods

### 3.1 Dataset

The dataset is designed to evaluate the accuracy potential of visual localization in large-scale outdoor environments and addresses the limitations identified in Section 2.2.

#### *3.1.1 Study area*

The study area, depicted in Figure 1, encompasses various districts of Muttenz, a suburban municipality near Basel, Switzerland. The areas covered include parts of a historic town center, commercial areas, residential areas with detached houses, multi-story apartment buildings, diverse vegetation, and a mixed industrial area with production facilities and brownfield sites with parked vehicles and trucks. The road network, with a length of approximately 10 km, exhibits substantial heterogeneity, encompassing multi-lane roads, intersections with tram




**Fig. 1** Study area in Muttenz, Switzerland. Street network of the two mobile mapping campaigns, GCPs, TPs and ITPs used for georeferencing and co-registration, and the five query scenes (circled numbers). Example images illustrate the conditions on site. Background map: © OpenStreetMap contributors.

lines, wide pedestrian sidewalks, and ordinary two- and single-lane roads in residential areas (Figure 1).

### *3.1.2 Reference imagery*

To efficiently acquire reference imagery along the street network, we employ vision-based mobile mapping. The entire study area was mapped twice. The first campaign was conducted in autumn (October) 2023. The weather conditions during the first campaign were predominantly sunny, resulting in shadow patterns, and the trees showed slightly tinted foliage. The second campaign was conducted in early spring (March) 2025, under overcast skies, resulting in diffuse light, with trees free of foliage (Figure 1). During both campaigns, extensive construction work was in progress, including the complete redesign of entire road sections.

**Table 2** Camera systems of the 2023 and 2025 mobile mapping campaigns with horizontal (h) and vertical (v) field of view (FoV), resolution and number of images. Note: for the panorama camera the single camera heads are reported.

| | Camera system | FoV [°] h | FoV [°] v | Resolution | # Images |
|---|---|---|---|---|---|
| 2023 | Stereo front | 96.0 | 64.1 | 5472 × 3084 | 8054 |
| | Stereo back left | 86.9 | 64.4 | 4864 × 3232 | 8054 |
| | Stereo back right | 86.9 | 64.4 | 4864 × 3232 | 8054 |
| | Pano camera (5×) | 73.0 | 86.0 | 1536 × 1936 | 4027 |
| 2025 | Mono front | 75.1 | 46.9 | 5472 × 3084 | 4170 |
| | Pano camera (5×) | 73.0 | 86.0 | 2918 × 3677 | 4170 |

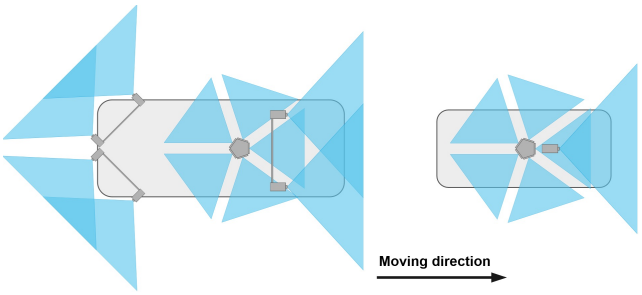


**Fig. 2** Mobile mapping systems of the 2023 (left) and 2025 (right) campaigns, drawn to scale.

In the 2023 campaign, the camera setup (Table 2 & Figure 2, left) consisted of three high-resolution stereo systems facing the front, back right, and back left, combined with a medium-resolution multi-head panorama camera, Ladybug5+. In 2025, the MMS was equipped with a high-resolution monocular camera facing the front and a high-resolution multi-head panorama camera, Ladybug6 (Table 2 & Figure 2, right). Figure 2 visualizes the camera footprints for each campaign. From 2023 to 2025, both the camera sensor setup and the size of the mapping platform changed. As a result, rear-view images from the 2025 panoramic camera contain almost no elements of the MMS itself, unlike the 2023 images. During the reference mapping campaigns, all cameras were triggered synchronously at approximately 4 m intervals. Wide roads were traversed in both directions, while narrow roads were traversed only once. Construction

activities slightly changed the road network coverage between campaigns (Figure 1).

All camera systems used were pre-calibrated, yielding distortion-free images that were, in the case of the stereo systems, also rectified. The panoramic images are only available as cube maps with stitching errors caused by the differing projection centers of the multi-head panorama cameras. Hence, we remapped them to the original camera heads, resulting in five images with a horizontal field of view (FoV) of 73° (Table 2), with the image content covering 72° per head, so that the five images cover the full panorama without overlap.

For georeferencing of the acquired imagery, both MMS were equipped with a tactical-grade IMU and a geodetic GNSS receiver comparable to the MMS described by Cavegn et al. (2018). The image poses were computed via integrated georeferencing (Eugster et al. 2012), using 32 of the ground control points (GCPs) described below, evenly distributed throughout the mapping perimeter, which yields a georeferencing accuracy in the sub-decimeter range. Evaluating visual localization at centimeter thresholds requires sub-centimeter image poses, obtained in two further steps.

First, we applied image-based georeferencing (Cavegn et al. 2018) to each campaign independently to obtain a geometrically stable photogrammetric model and compensate for residual inaccuracies in the image poses. Using Agisoft Metashape (Agisoft 2024), we refined all image poses within a bundle adjustment. Stereo systems and panorama cameras were modeled as multi-camera systems using relative orientation constraints. Each camera was assigned its initial pose from integrated georeferencing to efficiently select image pairs for image matching.

**Table 3** Co-registration of the two mobile mapping campaigns: RMSE of the coordinate differences between 2023 and 2025 at the GCPs measured at both campaigns, TPs and ITPs.

| Point type | Num points | RMSE [m] X | Y | Z | 3D |
|---|---|---|---|---|---|
| GCP | 43 | 0.001 | 0.002 | 0.001 | 0.002 |
| TP | 180 | 0.004 | 0.004 | 0.002 | 0.006 |
| ITP | 29 | 0.003 | 0.003 | 0.002 | 0.005 |

Second, for co-registration of both mobile mapping campaigns, 69 GCPs determined using RTK-GNSS and 180 clearly identifiable tie points (TPs) (Figure 1, blue points) were manually measured in approximately 10 images per campaign. Of the 69 GCPs, 66 were measurable in the 2023 and 46 in the 2025 imagery, 43 of them in both campaigns. A global bundle adjustment was then performed, incorporating all GCPs, the TPs and the initial 6-DoF image poses from integrated georeferencing. No automatic correspondences between images of different campaigns entered this adjustment, so that the two campaigns are connected exclusively via the manually measured GCPs and TPs. For the GCPs the 3D positional standard deviation was set to 5 cm, while the image poses were incorporated with a 3D position standard deviation of 10 cm and a 3D rotation standard deviation of 1°.

To validate the accuracy of co-registration, an additional set of 29 independent tie points (ITPs) (Figure 1, purple points) uniformly distributed throughout the mapping perimeter was measured in the images of both campaigns. These ITPs were excluded from the bundle adjustment to ensure independence. The 3D coordinates of the GCPs, TPs, and ITPs were then triangulated independently for each mobile mapping campaign from the fixed image poses. For each point set, we report the root mean square error (RMSE) of the coordinate differences between the 2023 and 2025 campaigns (Table 3). The two campaigns are co-registered in the sub-centimeter range, with the ITPs providing an independent check. The absolute georeferencing accuracy of each campaign, determined from the residuals of the GCPs, amounts to a 3D RMSE of 0.025 m for both the 2023 (66 GCPs) and the 2025 campaign (46 GCPs). Since these points were used in the bundle adjustment, this is an internal measure of accuracy.

### *3.1.3 Query sequences*

For query image acquisition, we identified five small-scale scenes throughout the study area (Figure 1, circled numbers). Each of these scenes has different characteristics (Figure 3). Scene 1 depicts a local access road in a suburban environment, with sidewalks on both sides of the street and a pedestrian crossing with a small island. The street traverses a residential neighborhood of detached houses and abundant greenery, including hedges and trees. In scene 2, a local access road in an area characterized by multi-story commercial, industrial, and office buildings is shown. Road construction is ongoing on the street and in its immediate surroundings; as a result, no road markings are visible and numerous temporary construction fences are in place. On both sides of the road, there are wide sidewalks and some unfinished flower beds without plants. Scene 3 represents a busy, multi-lane main road in a commercial environment, featuring car dealers and a fuel station with predominantly glass and other reflective façades, and many dynamic objects, such as cars. On both sides of the street, there is a sidewalk and only a small amount of vegetation.




**Fig. 3** Example images of the five query scenes, which differ in street type, environment and scene changes such as moving cars, construction sites and vegetation.

In scene 4, a local access road with an integrated cycle path traverses a neighborhood with multi-story urban residential and commercial buildings. Sidewalks are provided on both sides of the street, separated by flower beds. The environment is densely vegetated, with trees and hedges. Scene 5 shows a narrow residential street in a neighborhood of detached houses, with plenty of greenery, including large trees and hedges. Various cars are parked on the side of the road.

As query sensors, we employed four camera configurations with different characteristics, including resolution, field of view, and camera models. The four camera systems are: a Canon EOS R with 24 mm and 35 mm lenses, a GoPro Hero 7, and an iPhone 14 Pro (Table 4).

**Table 4** Query cameras with horizontal (h) and vertical (v) field of view (FoV), projection model (P: perspective, F: fisheye) and sensor resolution.

| Camera | FoV [°] | | Cam model | Resolution |
|---|---|---|---|---|
| | h | v | | |
| Canon 24mm | 72.3 | 51.9 | P | 6720 × 4480 |
| Canon 35mm | 53.3 | 37.0 | P | 6720 × 4480 |
| GoPro Hero 7 | 122.6 | 94.4 | F | 4000 × 3000 |
| iPhone 14 Pro | 57.1 | 71.9 | P | 1440 × 1920 |

The query image acquisition was conducted in late winter (February) 2025. The weather conditions varied. Two scenes were captured under predominantly cloudy conditions, while the remaining three scenes were captured in sunny weather. At that time, the trees were free of foliage. In each scene, a sequence of images was captured with all four cameras. The sequences differ in trigger rate, trajectory, length, and number of images because each was captured one after another. The Canon camera was triggered manually, ensuring the images were captured in roughly an upright orientation. For the GoPro, an upright configuration was used as well. Instead of manual triggering, a video was recorded at 24 frames per second (FPS) with electronic stabilization disabled. For iPhone 14 Pro sequences, PIX4DCatch (Pix4D 2025) was used and image triggering was set according to distance and angle criteria of 0.1 m and 15°, respectively. During this capture, the iPhone was deliberately pointed slightly towards the ground and, at times, moved rapidly to create a trajectory similar to walking while using an AR navigation application. Each sequence forms one or more closed loops. The GoPro and iPhone sequences were subsampled to every 10th and every 5th image, respectively, to obtain image counts comparable to the Canon sequences. The number of query images per camera and scene is listed in Table 5. Furthermore, for each scene, 7–11 clearly identifiable and distinct natural points were selected as control points (CPs) and determined using RTK-GNSS.

**Table 5** Number of query images per camera and scene.

| Camera | Scenes | | | | | Total |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| Canon 24mm | 165 | 194 | 135 | 142 | 165 | 801 |
| Canon 35mm | 154 | 158 | 116 | 168 | 181 | 777 |
| GoPro Hero 7 | 415 | 376 | 370 | 383 | 403 | 1947 |
| iPhone 14 Pro | 435 | 246 | 350 | 380 | 463 | 1874 |
| Total | 1169 | 974 | 971 | 1073 | 1212 | **5399** |

For co-registration of the acquired query images with the reference imagery, we employed a multi-stage SfM-based procedure using Agisoft Metashape (Agisoft 2024). Rather than matching query against reference images directly, which Zhang et al. (2021) identify as the main difficulty in SfM-based reference pose generation, we first create a stable photogrammetric model of the query sequences and then co-register it to the reference imagery via clearly identifiable and precisely measurable points in object space.

To create the photogrammetric model, images of all four sequences were initially aligned individually. For each, at least three CPs were measured to transform the image poses into a unified global reference frame. We then performed an additional image alignment using the image poses from the first alignment to efficiently select matching candidates across sequences. All CPs, as well

as approximately 30 additional distinctive tie points at strategically selected locations, were measured in approximately 15 images from different camera sequences. These points help to mutually stabilize the sequences from different cameras per scene.

To co-register the photogrammetric model with the reference imagery, we manually measured approximately 19–24 clearly identifiable points in both the reference and the query images. Each point was observed from as many different viewing directions as possible and their 3D coordinates were triangulated from the fixed reference image poses. For each scene, at least three points were selected as independent check points and excluded from the subsequent adjustment. Since the co-registration refers to the reference imagery, we term these check points relative check points (RCPs) and the points used for co-registration relative ground control points (RGCPs).

Finally, we performed a bundle adjustment that jointly optimizes the poses of the query images and their camera calibration parameters, with the RGCPs introduced as control points with a 3D standard deviation of 0.005 m. The CPs served only for the initial alignment and were therefore not introduced in the bundle adjustment. Query and reference images are thus connected exclusively via the manually measured points. The camera model comprises the focal length $f$, the principal point $(c_x, c_y)$, and the radial ($k_1$–$k_4$) and decentering ($p_1$, $p_2$) distortion coefficients of the Brown–Conrady model (Brown 1971), applied to a perspective projection for the Canon and iPhone cameras and to an equidistant fisheye projection for the GoPro (Agisoft 2024). The affinity and skew terms $b_1$ and $b_2$ of the extended model (Fraser 1997) were not estimated, assuming square pixels and orthogonal image axes, and because $b_1$ is highly correlated with the focal length.

**Table 6** Co-registration of the query images per scene: RMSE of the coordinate differences between points triangulated from the query and from the reference image poses. RCPs were excluded from the adjustment and serve as independent check points.

| Scenes | Points type | # | RMSE [m] X | Y | Z | 3D |
|---|---|---|---|---|---|---|
| Scene 1 | RGCP | 16 | 0.005 | 0.006 | 0.002 | 0.008 |
| | RCP | 3 | 0.002 | 0.006 | 0.003 | 0.007 |
| Scene 2 | RGCP | 18 | 0.007 | 0.002 | 0.004 | 0.008 |
| | RCP | 4 | 0.006 | 0.004 | 0.004 | 0.008 |
| Scene 3 | RGCP | 20 | 0.004 | 0.005 | 0.003 | 0.008 |
| | RCP | 4 | 0.007 | 0.004 | 0.003 | 0.009 |
| Scene 4 | RGCP | 16 | 0.005 | 0.006 | 0.002 | 0.008 |
| | RCP | 4 | 0.003 | 0.003 | 0.004 | 0.006 |
| Scene 5 | RGCP | 17 | 0.004 | 0.004 | 0.002 | 0.006 |
| | RCP | 6 | 0.004 | 0.004 | 0.002 | 0.006 |

To assess the co-registration, we triangulated the 3D coordinates of the RGCPs and RCPs from the fixed query image poses and report the RMSE of their differences to the coordinates triangulated from the fixed reference image poses (Table 6). For the RCPs, which were excluded from the adjustment, these differences provide an independent check. At 6–9 mm, they show that the photogrammetric models are co-registered to the reference imagery with sub-centimeter accuracy. This is an upper bound, since the differences also contain the uncertainty in manually measuring the natural points in the images.

The differences between the triangulated and the RTK-GNSS coordinates of the CPs provide an independent check of the georeferencing of the reference imagery, because the final adjustment is tied to it exclusively via the RGCPs. They amount to 1.4–3.0 cm per scene (Table 7) and 2.2 cm over all 48 CPs, consistent with the internal estimate of 2.5 cm (Section 3.1.2), and likewise represent an upper bound, since they additionally contain the co-registration error and the uncertainty of the RTK-GNSS measurements. For consistent reporting, we use 2.5 cm as the georeferencing accuracy of the entire dataset in the following.

**Table 7** Georeferencing accuracy of the query images per scene: 3D RMSE between the CP coordinates from RTK-GNSS and those triangulated from the fixed query image poses.

| | Scenes 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Num points | 11 | 10 | 7 | 9 | 11 |
| RMSE [m] | 0.014 | 0.024 | 0.030 | 0.020 | 0.020 |

### *3.1.4 Uncertainty of the query ground-truth poses*

Assessing the uncertainty of the individual query poses directly would require reference poses of superior accuracy, which can only be obtained with considerable effort and thus for a few images at most, not for all 5399 query images from four different cameras (Table 5). The co-registration accuracy in Table 6 does not close this gap, as it characterizes the photogrammetric models as a whole rather than the single image poses. Moreover, camera calibration parameters and image poses are correlated when estimated simultaneously (Remondino and Fraser 2006). Thus, errors in one can be compensated by the other, biasing the poses while leaving the object-space residuals small. We therefore assess how sensitive the image poses and the camera calibration are to the control point configuration. This configuration defines

**Table 8** Residual pose differences to the published ground truth after removing a common similarity transformation; 30 runs per scene with 75 %, 50 % and 25 % of the points as control. $\Delta t$: distance between the projection centers; $\Delta R$: relative rotation angle (Eq. 2). p95: per run over all images, averaged over the runs; max: over all images and runs.

| | $\Delta t$ [mm] | | | $\Delta R$ [$10^{-3}$ °] | | |
|---|---|---|---|---|---|---|
| Scene | 75 % | 50 % | 25 % | 75 % | 50 % | 25 % |
| *Quantile 95 % (p95)* | | | | | | |
| 1 | 0.6 | 1.3 | 1.5 | 0.5 | 0.8 | 1.0 |
| 2 | 0.6 | 1.0 | 1.4 | 0.9 | 1.4 | 1.8 |
| 3 | 2.5 | 2.2 | 4.8 | 1.7 | 2.1 | 3.5 |
| 4 | 0.5 | 0.5 | 0.5 | 0.4 | 0.4 | 0.6 |
| 5 | 1.3 | 2.0 | 2.7 | 0.5 | 0.8 | 1.1 |
| *Maximum value (max)* | | | | | | |
| 1 | 2.5 | 3.5 | 3.1 | 1.5 | 2.2 | 2.0 |
| 2 | 2.2 | 3.5 | 3.8 | 3.2 | 4.0 | 4.3 |
| 3 | 4.6 | 5.0 | 11.2 | 6.5 | 5.8 | 9.4 |
| 4 | 1.0 | 1.2 | 1.2 | 1.6 | 1.5 | 1.5 |
| 5 | 4.9 | 5.5 | 5.0 | 2.1 | 2.4 | 2.4 |

the datum of the poses and thus the frame in which all pose errors are evaluated. If the photogrammetric model is geometrically well conditioned by the image observations, a change of the RGCP configuration mainly changes the realization of the datum. Therefore, the residual pose differences, i.e., the differences remaining after transforming the poses into a common datum, are small, and the camera calibration remains stable. Large residual differences and varying calibration parameters would instead indicate that the poses are determined by the control point configuration rather than by the image observations.

For each scene, we pooled the RGCPs and RCPs and repeated the co-registration described in Section 3.1.3 30 times. Each run used randomly drawn subsets of 75 %, 50 % and 25 % of the points as RGCPs, while the remaining points serve as RCPs. To avoid a favorable starting point, the camera calibration was initialized in each run with nominal values: focal length from the manufacturer specifications, principal point at the image center, all other parameters set to zero. The resulting poses were transformed into the datum of the published ground truth, and the residual pose differences were computed separately for position and orientation. In addition, we assessed the variation of the calibration parameters.

The residual pose differences are at the millimeter level (Table 8). With 75 % and 50 % of the points as RGCPs, the p95 values do not exceed 2.5 mm and 0.0021°, and reducing the control from 75 % to 50 % increases them by at most 0.7 mm and 0.0005°. Even with only 25 % of the points, all runs converged and the p95 values do not exceed 4.8 mm and 0.0035°. The largest differences occur in Scene 3 (up to 11.2 mm and 0.0094° with 25 % control), where all sequences were captured from one sidewalk, resulting in the weakest network geometry of all scenes. The camera calibration remains stable as well. Across all runs, the focal length varies by at most 38 ppm (0.174 px) and the principal point by at most 55 ppm (0.370 px) in x and 57 ppm (0.257 px) in y direction.

Reducing the number of control points therefore mainly changes the realization of the datum, while the internal geometry of the photogrammetric network and the camera calibration remain largely unaffected. A compensation between calibration and image poses is therefore not evident, and the poses within the photogrammetric model are determined by the image observations rather than by the control point configuration. The resulting uncertainty of the ground-truth positions relative to the reference imagery is 6–9 mm. This estimate combines the RMSE of the co-registration determined by the RCPs (Table 6) with the p95 residual pose differences when using 75 % of the RGCPs via error propagation. The orientation uncertainty is well below 0.1°.

This analysis characterizes the precision of the ground-truth poses and their sensitivity to the realization of the datum and to the initialization of the camera calibration. Since the image observations, the manually measured points and the processing chain are identical in all runs, it cannot detect errors common to them. The resulting residual uncertainty is discussed in Section 5.4.

### 3.2 Visual Localization Pipeline

Figure 4 provides a schematic overview of our visual localization pipeline. Its central design choice is to use precisely georeferenced, high-resolution imagery acquired by vision-based MMS directly as the scene representation. This choice is motivated by the need for scalability, accuracy, and the flexibility to modify individual components of the visual localization pipeline. The pipeline is composed of established components, which we adopt with minor adaptations: prior-guided reference candidate selection (Yan et al. 2023), tiled extraction of sparse local features (Maiwald et al. 2023; Morelli et al. 2024), and pose estimation from a local SfM model reconstructed on the fly (Torii et al. 2021; Humenberger et al. 2022). Sparse local features are chosen because, unlike dense matchers, they remain manageable at high image resolutions and large scale (Section 3.2.2).

The pipeline consists of two stages. In the offline stage, global image descriptors and local features are extracted from the reference images. The global descriptors are stored in a vector database together with the image

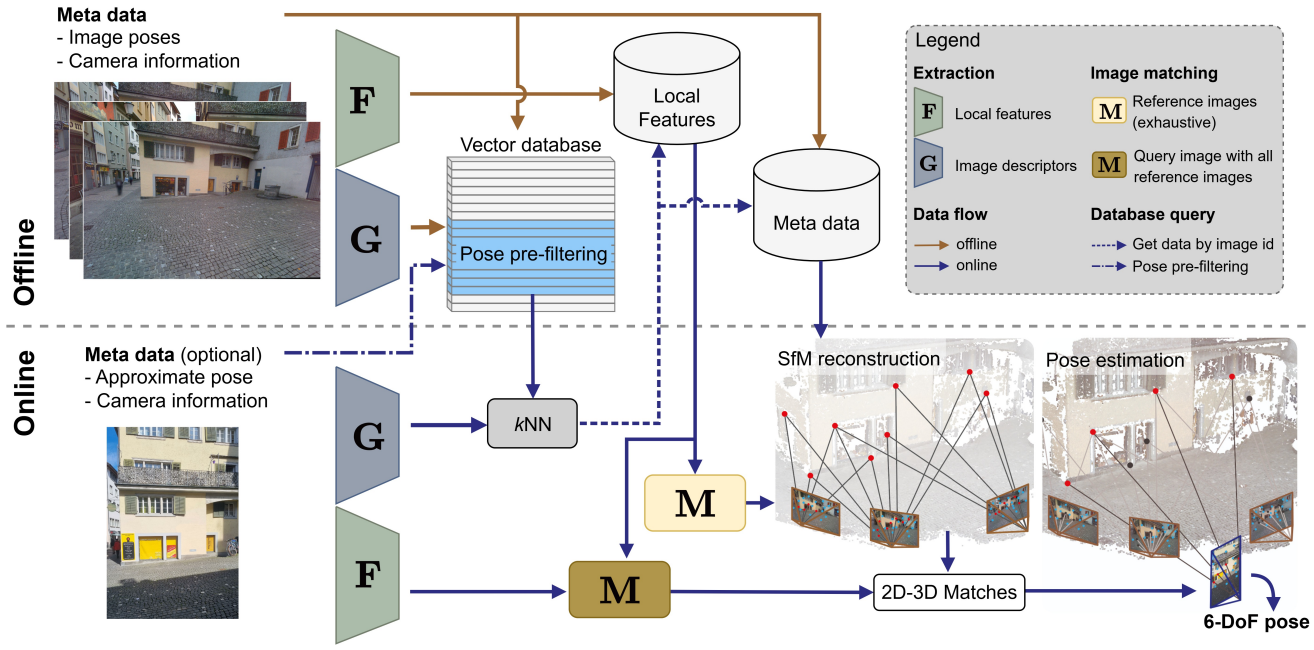


**Fig. 4** Schematic overview of the visual localization pipeline, with offline preprocessing of the reference imagery (upper half) and online localization of a query image (lower half); see Section 3.2 for details.

poses, and the local features and metadata, such as image poses and camera intrinsics, in separate databases (Figure 4, upper half). The online stage comprises three steps, which are described in the following subsections: reference candidate selection, feature extraction and matching, and local SfM model construction and pose estimation (Figure 4, lower half). The specific methods and their settings are given in Section 4.2.

### *3.2.1 Reference candidate selection*

The coarse-to-fine localization paradigm proposed by Sarlin et al. (2019) is widely adopted in visual localization methods. Although the retrieval of reference candidates has proven highly effective, these methods remain susceptible to failure when images depict different locations while exhibiting nearly identical visual content (Cai et al. 2023), a problem that grows with scale. To disambiguate the similarity search in large-scale environments and to improve efficiency, we add a pose pre-filtering step to perform IR on a spatially constrained subset of descriptors. Since many devices, such as smartphones, are nowadays equipped with a GNSS sensor, we adopt the sensor-guided retrieval procedure introduced by Yan et al. (2023) to constrain the search space. Along with each image descriptor, we store position and heading information as additional attributes in the vector database. In the online stage, an axis-aligned square search region, extending $\pm d$ around the approximate query position, where d is the positional uncertainty, is used to retrieve the candidate descriptors from the local neighborhood. This subset can be further refined by a heading filter. In the absence of explicit information about pose uncertainty, a default bounding box of 1×1 km ($d$ = 500 m) is adopted for spatial filtering. Finally, within the resulting subset, a k-nearest-neighbor (kNN) search identifies the $k$ most similar reference images. If no position prior is available, spatial filtering is omitted and the k-nearest-neighbor search covers the entire database. Since the local SfM model requires reference images acquired from diverse viewpoints (Torii et al. 2021; Humenberger et al. 2022), the selection of reference candidates has a decisive influence on pose accuracy, as Section 4.3 shows.

### *3.2.2 Feature extraction and matching*

We use sparse local features, which can be robustly re-identified under challenging conditions (DeTone et al. 2018; Tyszkiewicz et al. 2020; Zhao et al. 2023; Potje et al. 2024; Chen et al. 2025). Dense matchers outperform them under low texture, large viewpoint changes and difficult lighting (Jiang et al. 2026; Bonilla et al. 2025; Panek et al. 2026), but impose substantially higher computational and storage costs (Edstedt et al. 2024), which becomes critical in large-scale, high-resolution settings.

Current sparse local features are typically trained on images of approximately 1 MP or less, e.g., ALIKED (Zhao et al. 2023) on 800×800 and DISK (Tyszkiewicz

et al. 2020) on 1024×1024 pixels. To apply these methods to high-resolution images, we adopt a tiled feature extraction strategy, following Maiwald et al. (2023) and Morelli et al. (2024). Because features close to tile borders are often discarded for lack of context, gaps can appear near tile borders and introduce systematic effects. To mitigate this effect, tiles are arranged to overlap by a predefined number of pixels, as in Morelli et al. (2024). After feature extraction, we reassemble the tiled outputs into a single feature set and apply non-maximum suppression (NMS) with a radius of 1 pixel to remove duplicate detections in the overlapping image regions. Finally, the query and reference images are matched exhaustively to obtain 2D–2D correspondences.

#### *3.2.3 Local SfM-model reconstruction and pose estimation*

To estimate the pose of the query image, we first reconstruct a locally consistent SfM model by triangulating 3D points from geometrically verified 2D–2D correspondences between the reference images. In this step, both the camera intrinsics and the poses of the reference images are kept fixed. The 2D–2D correspondences between the query image and the reference images are subsequently examined to determine whether they are associated with reconstructed 3D points, thereby yielding a set of 2D–3D correspondences. Following the findings of Vultaggio et al. (2025), we estimate the query image's pose using a PnP solver within a RANSAC framework, followed by non-linear pose refinement. In this work, only undistorted images are used (Section 3.1.3). Therefore, the intrinsics are kept fixed during pose estimation.

## 4 Experiments

We perform four experiments, each addressing one of the sub-questions (i)–(iv) stated in Section 1. Table 9 summarizes the configuration of each experiment.

### 4.1 Evaluation protocol and metrics

For all experiments, we evaluate the performance of visual localization by the pose errors. Following common practice, we report the percentage of query images that are localized within specific error thresholds (Sattler et al. 2018). We compute the translation $\epsilon_t$ and rotation error $\epsilon_R$ according to Equations (1) and (2) and apply seven error thresholds: [0.01, 0.025, 0.05, 0.1, 0.2, 0.5, 1] meters for translation and [0.1, 0.25, 0.5, 1, 2, 5, 10] degrees for rotation. In addition, we use three error categories: accurate (0.05 m, 0.5°), moderate (0.1 m, 1°), and low (0.2 m, 2°) for concise reporting. Following Jin et al. (2021), we set $\epsilon_t$ and $\epsilon_R$ to infinity for images for which localization failed. Where appropriate, we also report the median translation and rotation errors.

$$\epsilon_t = \|t_{est} - t_{gt}\|_2 \tag{1}$$

$$\epsilon_R = \arccos\left(\frac{trace(R_{gt}^{-1} R_{est}) - 1}{2}\right) \tag{2}$$

All pose errors reported in this work are computed with respect to the ground-truth poses and therefore describe the accuracy relative to the reference imagery. The absolute accuracy additionally depends on the georeferencing of the reference imagery (Section 5.1) and has to be considered if absolute pose errors are of interest. The strictest translation threshold of 0.01 m is of the same order as the uncertainty of the ground truth: the two mobile mapping campaigns, whose images serve as queries in Sections 4.3 and 4.4, are co-registered with a 3D RMSE of 5 mm at independent tie points (Section 3.1.2), and the positions of the consumer queries have an uncertainty of 6–9 mm (Section 3.1.4). At this threshold, part of the measured error stems from the ground truth itself, so the reported percentages are a conservative estimate. The rotation thresholds are not affected, as the orientation uncertainty of the ground truth is well below 0.1°.

Images acquired in a sequence are not independent samples, as they share the static scene content and differ mainly in dynamic elements such as moving vehicles. The five scenes were moreover selected deliberately rather than drawn at random. We therefore report the results per scene and camera and base our conclusions on differences that appear consistently across the conditions varied in each experiment, such as scenes, cameras, image resolutions or retrieval methods. For each difference, we report in how many of these combinations it is observed.

### 4.2 Implementation details

The implementation of our visual localization pipeline adopts a modular design, composed of many different software modules. Unless otherwise specified, the default values for each method are used.

For reference candidate selection, we employ six IR methods: NetVLAD (Arandjelovic et al. 2016), CosPlace (Berton et al. 2022), EigenPlaces (Berton et al. 2023), AnyLoc (Keetha et al. 2024), DINOv2 SALAD (Izquierdo and Civera 2024), and MegaLoc (Berton and Masone 2025), using descriptor dimensions of 512 and

**Table 9** Configurations of the experiments: query and reference imagery, image retrieval (IR) method, position and heading pre-filtering, local features, and image resolution levels.

| Experiment | Reference imagery | Query imagery | IR method | Pre-filtering | Local features | Resolution |
|---|---|---|---|---|---|---|
| (i) Sec. 4.3 | 2023, all cameras | 2025 panorama, front head | six methods | five strategies | SuperPoint | 2048 |
| (ii) Sec. 4.4 | 2023 stereo | 2025 mono front | EigenPlaces (512) | ±15 m / ±90° | six features | 1024–full |
| (iii) Sec. 4.5 | 2023 / 2025 / both | consumer cameras | MegaLoc | ±100 m[a] / – | five features | 2048 |
| (iv) Sec. 4.6 | 2025 | consumer cameras | MegaLoc | ±100 m[a] / – | ALIKED | 1024 |

[a] Centered on one manually selected map coordinate per scene, identical for all query images of that scene.

2048 for CosPlace and EigenPlaces. For AnyLoc, we use VLAD aggregation of DINOv2 ViT-G/14 features with 32 clusters and the precomputed urban vocabulary of Keetha et al. (2024). Following Berton and Masone (2025), images are resized to 322×322 pixels for the vision transformer-based methods (AnyLoc, DINOv2 SALAD, MegaLoc), whereas for NetVLAD, CosPlace, and EigenPlaces the longer image side is scaled to 322 pixels, preserving the aspect ratio. For the $k$-nearest-neighbor search, we set $k = 10$, which offers a good trade-off between efficiency and accuracy (Humenberger et al. 2022).

We use SuperPoint (DeTone et al. 2018), DISK (Tyszkiewicz et al. 2020), ALIKED (Zhao et al. 2023), XFeat (Potje et al. 2024) and RDD (Chen et al. 2025) as local feature representations and SIFT (Lowe 2004), implemented in OpenCV (Bradski 2000), as a baseline. According to their original publications, SuperPoint, DISK and XFeat detect keypoints with pixel-level localization accuracy, while SIFT, ALIKED and RDD provide sub-pixel-accurate keypoint locations.

Within our tiled feature extraction module, we use tiles of 1024×1024 pixels, overlapping by 50 pixels, for all feature representations except SuperPoint, for which we use 1600×1600 pixels following the image resolution used by Sarlin et al. (2020). We define four image resolution levels, 1024, 2048, 4096 and full resolution, by scaling the longer image side to the chosen level while preserving the aspect ratio. The number of keypoints per image is limited to 4096 and 6144 at the resolution levels 1024 and 2048, respectively, and to 1024 keypoints per megapixel at 4096 pixels and full resolution. Apart from the number of keypoints, the default settings of each feature extraction method are used.

2D–2D image correspondences are obtained with LightGlue (Lindenberger et al. 2023), a widely used state-of-the-art baseline (Bellavia et al. 2025), except for XFeat, matched with LighterGlue, and SIFT, matched by mutual nearest neighbor (MNN). To achieve the highest accuracy and comparability across different feature representations, we disable depth confidence and point pruning in LightGlue and LighterGlue.

For 3D point triangulation and absolute pose estimation, we use COLMAP (Schönberger and Frahm 2016) with its default settings, with the following exceptions. The image poses of the reference images are kept fixed during triangulation and the camera intrinsics are not refined, as all images are undistorted (Section 3.1.3). We also enable two-view tracks to retain points observed in only two of the retrieved images, since each local model is built from few images. The query pose is estimated with COLMAP's LO-RANSAC-based estimator (Chum et al. 2003), which uses a minimal P3P solver and EPnP (Lepetit et al. 2009) for local optimization, with a reprojection-error threshold of 2 pixels, followed by non-linear pose refinement. Only for the comparison with HLoc in Section 4.6, we use HLoc's and COLMAP's default threshold of 12 pixels for both pipelines.

### 4.3 Investigation of reference candidate selection

This experiment evaluated the reference candidate selection strategy described in Section 3.2.1 by combining five pre-filtering strategies with the selected IR methods (Table 9).

As a baseline, we omitted the pre-filtering step and retrieved the 10 most similar images from the entire database. For position-based filtering, we used search regions of ±100 m and ±15 m. Each of these thresholds was combined with either no heading constraint or a heading threshold of ±90° (i.e., a 180° acceptance window). Since smartphones typically achieve GNSS accuracies of several meters, ±100 m represents an upper bound that should be met under almost all practical conditions, whereas ±15 m is a lower bound defined by the spatial density of our reference data and the aim of reconstructing a local SfM model.

The filters were applied to the poses from integrated georeferencing, i.e., to nearly exact priors. Imperfect priors are nevertheless covered, as the configuration without pre-filtering corresponds to a missing position prior and the configurations without heading constraint correspond to a missing heading. A position error within the search region is expected to yield results between those of the unfiltered and the filtered configurations,

**Table 10** Effect of reference candidate selection: percentage of poses within the accurate (acc: 0.05 m, 0.5°), moderate (mod: 0.1 m, 1°) and low (0.2 m, 2°) thresholds per pre-filtering strategy (±position / ±heading) and IR method. Queries: 2025 panorama front head; reference: 2023 campaign; SuperPoint at 2048 pixels. Best, second- and third-best results: **bold**, underlined, wavy underlined.

| IR method | Desc dim | – / – acc / mod / low | ±100 m / – acc / mod / low | ±100 m / ±90° acc / mod / low | ±15 m / – acc / mod / low | ±15 m / ±90° acc / mod / low |
|---|---|---|---|---|---|---|
| NetVLAD | 4096 | 52.1 / 66.5 / 70.2 | 60.8 / 77.1 / 81.7 | 64.3 / 82.8 / 88.0 | 71.2 / 88.7 / 92.3 | 77.2 / 94.0 / 97.0 |
| CosPlace | 512 | 79.5 / 87.4 / 89.8 | 82.8 / 91.2 / 94.0 | 84.1 / 92.4 / 95.6 | 89.3 / 95.1 / 97.3 | 89.6 / 95.9 / 98.1 |
| | 2048 | 81.0 / 87.9 / 90.4 | 84.9 / 91.8 / 94.1 | 86.4 / 92.8 / 95.3 | 89.2 / 95.3 / 97.2 | 90.4 / 96.1 / 97.9 |
| EigenPlaces | 512 | 84.0 / 91.7 / 94.0 | 86.1 / 93.9 / 96.1 | 86.5 / 94.3 / 96.7 | 90.2 / 96.2 / 97.8 | 90.3 / 96.2 / 98.0 |
| | 2048 | 84.7 / 92.2 / 94.8 | 86.3 / 94.1 / 96.8 | 87.2 / 94.8 / 97.2 | 90.3 / 96.3 / 97.9 | 90.4 / 96.4 / 98.1 |
| AnyLoc | 49152 | 78.5 / 86.2 / 88.3 | 79.6 / 87.3 / 89.7 | 84.3 / 92.5 / 95.1 | 87.3 / 93.3 / 95.6 | 90.7 / 96.4 / 97.9 |
| DINOv2 SALAD | 8448 | 85.9 / 93.3 / 95.5 | 87.6 / 94.8 / 97.0 | 87.8 / 95.1 / 97.3 | 91.2 / **96.6** / **98.3** | 91.1 / 96.5 / 98.4 |
| MegaLoc | 8448 | **90.3** / **95.9** / **97.6** | **90.3** / **96.3** / **98.3** | **90.4** / **96.0** / **98.2** | **91.5** / 96.4 / **98.3** | **91.7** / **96.7** / **98.6** |

whereas a larger error may exclude the correct reference images and cause the localization to fail. The search region should therefore be chosen conservatively with respect to the expected accuracy of the prior.

Pre-filtering improves pose accuracy in the accurate category in 31 of the 32 combinations of IR method and pre-filtering strategy, and leaves it unchanged in the remaining one (Table 10), but the gain depends heavily on the unfiltered baseline. In the accurate category, the weak NetVLAD gains 25.1 percentage points (pp), the already strong MegaLoc only 1.4 pp. Position-based filtering yields large and consistent improvements. Heading-based filtering adds only a marginal benefit for the stronger descriptors, but up to 6.0 pp for NetVLAD and 4.7 pp for AnyLoc. MegaLoc achieves the highest percentage in the accurate category for all pre-filtering strategies, and its lead is especially pronounced without pre-filtering. With the strictest pre-filtering (±15 m / ±90°), the IR methods differ by only about 2 pp (NetVLAD excepted), which shows that pre-filtering can effectively compensate for weaker descriptor expressiveness. The vision transformer-based methods MegaLoc and DINOv2 SALAD produce longer but more expressive descriptors than the convolutional neural network (CNN)-based methods NetVLAD, CosPlace, and EigenPlaces. The 512-dimensional descriptors of CosPlace and EigenPlaces perform only slightly worse than their 2048-dimensional counterparts, and for EigenPlaces almost equivalently. The clear exception is AnyLoc, which despite its transformer backbone and very long descriptor performs worse than DINOv2 SALAD and MegaLoc throughout and, without heading filter, also worse than CosPlace and EigenPlaces. This agrees with other evaluations on street-level data (Berton and Masone 2025; Izquierdo and Civera 2024; Haitz et al. 2026). AnyLoc aggregates features of a frozen DINOv2 backbone with an unsupervised vocabulary and no task-specific training (Keetha et al. 2024), whereas the other methods are trained on datasets that include street-level imagery similar to ours.

### 4.4 Effects of high-resolution images on visual localization

In general, higher image resolutions yield more accurate image observations. In our visual localization pipeline, the image observations affect the accuracy of the estimated pose of the query image in two ways, directly through the keypoints extracted from the query image and indirectly through the accuracy of the reconstructed 3D point coordinates.

To evaluate the impact of image resolution on the final image poses, we tested the six local feature representations at the four image resolution levels, using the high-resolution images of the 2025 monocular front camera as queries and the images of the three 2023 stereo systems as reference (Table 9). We excluded the panorama camera, as its substantially lower spatial resolution (Table 2) would introduce heterogeneous keypoint localization accuracies within an otherwise uniformly high-resolution setting. As the strict ±15 m / ±90° pre-filtering makes the choice of IR method nearly irrelevant (Section 4.3), we used the compact 512-dimensional EigenPlaces descriptor.

Since both query and reference images were recorded by MMS whose trajectories are constrained by lanes, strong viewpoint changes were largely avoided. This matters because many local feature representations are neither scale- nor rotation-invariant, with the lack of rotation invariance having the considerably stronger adverse effect (Tyszkiewicz et al. 2020; Zhao et al. 2023). For the same reason, we applied the resolution level to both query and reference imagery, which avoids a scale difference between them. The reported results therefore reflect the resolution of the scene representation together with that of the query imagery. The setup thus isolates the effect of image resolution, while changes

**Table 11** Effect of image resolution and local features: percentage of poses within seven error thresholds and median pose errors. Queries: 2025 mono front camera; reference: 2023 stereo cameras. Double underline: best overall per threshold; wavy underline: best resolution per feature; colored entries: best feature per resolution (**1024**, **2048**, **4096**, and **full**).

| Method | Image res | Med. pose error $\epsilon_t$ [m] | Med. pose error $\epsilon_R$ [°] | 0.01m 0.1° | 0.025m 0.25° | 0.05m 0.5° | 0.1m 1° | 0.2m 2° | 0.5m 5° | 1m 10° |
|---|---|---|---|---|---|---|---|---|---|---|
| SIFT | 1024 | 0.114 | 0.213 | 7.3 | 20.3 | 33.6 | 47.4 | 58.1 | 64.8 | 66.7 |
| | 2048 | 0.124 | 0.215 | 13.0 | 26.9 | 37.4 | 46.8 | 55.0 | 60.1 | 61.6 |
| | 4096 | 0.873 | 1.849 | 14.1 | 24.7 | 32.2 | 38.6 | 44.1 | 48.7 | 50.3 |
| | full | -[a] | -[a] | 13.6 | 22.5 | 28.8 | 34.8 | 40.4 | 45.2 | 46.8 |
| SuperPoint | 1024 | 0.022 | 0.051 | 17.4 | 56.4 | 79.4 | 91.5 | 96.5 | 98.4 | 98.7 |
| | 2048 | 0.013 | **0.031** | 35.9 | 78.1 | 92.2 | 97.2 | 98.5 | 98.9 | 99.0 |
| | 4096 | 0.011 | **0.027** | 45.9 | 82.2 | 93.7 | 97.5 | 98.2 | 98.6 | 98.7 |
| | full | 0.011 | **0.026** | 47.1 | 82.8 | 94.0 | 97.5 | 98.3 | 98.6 | 98.7 |
| DISK | 1024 | 0.021 | **0.047** | 24.4 | 55.5 | 76.0 | 89.0 | 94.3 | 97.0 | 97.6 |
| | 2048 | 0.012 | 0.033 | 41.1 | 72.3 | 85.6 | 92.2 | 94.8 | 96.1 | 96.6 |
| | 4096 | 0.015 | 0.034 | 38.3 | 62.9 | 75.5 | 81.9 | 84.8 | 86.5 | 86.9 |
| | full | 0.018 | 0.035 | 32.9 | 58.5 | 71.0 | 77.4 | 79.4 | 80.6 | 81.0 |
| ALIKED | 1024 | 0.019 | 0.049 | 21.1 | 61.8 | 81.4 | 92.5 | 96.7 | 98.3 | 98.7 |
| | 2048 | 0.012 | 0.034 | 41.4 | 80.0 | 92.8 | 97.0 | 98.3 | 98.9 | 99.0 |
| | 4096 | 0.011 | 0.030 | 46.3 | 81.6 | 93.2 | 96.8 | 97.8 | 98.2 | 98.3 |
| | full | 0.012 | 0.029 | 41.7 | 78.2 | 91.7 | 95.7 | 97.2 | 97.7 | 97.7 |
| XFeat | 1024 | 0.045 | 0.080 | 5.4 | 26.8 | 53.7 | 76.4 | 89.2 | 95.8 | 97.4 |
| | 2048 | 0.026 | 0.046 | 14.1 | 48.7 | 74.3 | 88.6 | 94.2 | 96.9 | 97.5 |
| | 4096 | 0.019 | 0.036 | 23.1 | 61.7 | 82.6 | 90.1 | 93.2 | 94.6 | 95.1 |
| | full | 0.018 | 0.035 | 28.9 | 61.9 | 79.0 | 86.6 | 89.9 | 91.2 | 91.6 |
| RDD | 1024 | **0.012** | **0.047** | **39.0** | **76.1** | **89.2** | **95.0** | **97.5** | **98.8** | **98.9** |
| | 2048 | **0.009** | 0.032 | **53.4** | **85.9** | **94.8** | **97.5** | **98.7** | **99.0** | **99.1** |
| | 4096 | **0.010** | 0.030 | **48.5** | **83.3** | **94.2** | **97.6** | **98.5** | **99.0** | **99.0** |
| | full | **0.010** | 0.029 | **48.4** | **84.7** | **95.2** | **98.2** | **98.9** | **99.1** | **99.2** |

[a] Undefined, as fewer than 50 % of the images were successfully localized.

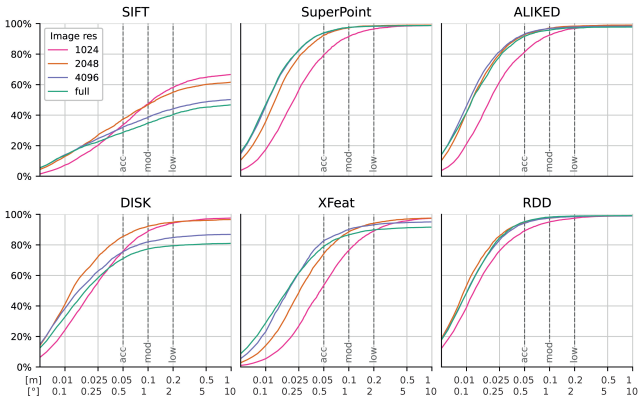


**Fig. 5** Percentage of poses with translation error $\epsilon_t$ and rotation error $\epsilon_R$ below the value on the logarithmic x-axis (0.0005–1 m and 0.005–10°), for six local features (SIFT, SuperPoint, DISK, ALIKED, XFeat and RDD) at four image resolution levels (1024, 2048, 4096 and full).

in vegetation and weather, ongoing construction work and differing camera intrinsics still reflect realistic conditions.

The effect of image resolution is most pronounced at the two strictest thresholds, 0.01 m / 0.1° and 0.025 m / 0.25° (Table 11, Figure 5). Increasing the resolution from 1024 to 2048 pixels improves all learned local features, reducing their median translation errors by 25–43 % and their median rotation errors by 30–43 %. Beyond 2048 pixels, only SuperPoint benefits from higher resolution throughout, though with diminishing returns, whereas SIFT shows no consistent optimum. DISK exhibits a resolution-dependent optimum at 2048 pixels, whereas for ALIKED, XFeat and RDD the optimum additionally depends on the error threshold, with ALIKED peaking at 4096 pixels in the high-accuracy regime, i.e., at the thresholds of 0.05 m / 0.5° and below, and at 2048 pixels at the coarser thresholds. RDD attains the highest percentage of all configurations at all seven error thresholds, by 6.3 and 3.1 pp at the two strictest thresholds and by less than 1 pp from 0.1 m / 1° onward. Note that the number of extracted keypoints grows with the resolution (Section 4.2) and that the RANSAC threshold is fixed in pixels, so the results reflect these combined settings rather than the resolution alone.

### 4.5 Visual localization performance with consumer-grade query images

We then assessed the accuracy potential using the query images acquired with consumer devices. Unlike the previous experiment, these query images include rotated and scaled image content due to strong viewpoint changes, as well as close-up images with very little visual context. We used three distinct reference configurations: the 2023 campaign, the 2025 campaign and both campaigns combined. Combining both campaigns simulates a reference map built from multiple epochs, as occurs when reference data is added rather than replaced during updates.

We omitted SIFT because of its inferior performance in Section 4.4 and used a resolution of 2048 pixels. For reference candidate selection, we used MegaLoc, the strongest IR method, with a position-based pre-filter of ±100 m. For each scene, the search region was centered on a single coordinate selected manually on a map, so that the prior was independent of the ground truth and identical for all query images of a scene.

Tables 12, 13 and 14 report the results aggregated over the five scenes for the 2023, 2025 and combined reference data, and Table 15 compares the median pose errors. Since current local feature representations exhibit environment-dependent performance due to limited training data (Bonilla et al. 2025), Figure 6 additionally shows the results per scene and camera for the 2025 reference data, which is subject to the smallest long-term changes and therefore allows the clearest comparison of the local feature representations.

Switching from the 2023 to the more recent 2025 reference data substantially improves localization accuracy. In the accurate category it yields a higher percentage for all 20 combinations of camera and feature representation (Tables 12 and 13), and the median errors decrease accordingly (Table 15).

With both reference epochs, RDD, ALIKED and SuperPoint form a leading group. Against the 2025 reference data, RDD achieves the highest percentages at 0.025 m / 0.25° and 0.05 m / 0.5° and the lowest median translation errors for all four cameras, albeit by small margins of 2.5 to 3.4 pp and 1 to 2 mm (Tables 13 and 15). DISK and XFeat perform substantially worse with both epochs. For the 2025 reference data, they lag behind the best representation in the accurate category by 15 to 26 pp and 25 to 33 pp, respectively. At the coarsest thresholds the representations converge.

Per scene and camera (Figure 6), ALIKED, RDD and SuperPoint perform comparably and, in the accurate category, even the weakest of the three outperforms XFeat and matches or exceeds DISK in all 20 combinations of scene and camera. None of the three leads consistently, as RDD is best in 9 of the 20 combinations, ALIKED in 6 and SuperPoint in 5. The most pronounced drops for XFeat occur in scene 3, where moving or parked vehicles and reflective glass façades produce features that cannot be matched reliably, and in scene 4, where bare branches yield fine-scale structures that are difficult to re-detect and repeated windows produce ambiguous correspondences, to which the comparatively compact descriptors of XFeat may be more susceptible.

Comparisons between the cameras must be interpreted with caution, as the sequences differ not only in the field of view, but also in trajectory, triggering, projection model and sensor resolution, and at 2048 pixels a wider field of view comes with a coarser angular resolution per pixel. With this caveat, the GoPro images (horizontal FoV of 122.6°) yield the lowest median translation errors, whereas the Canon 35 mm images (horizontal FoV: 53.3°) achieve the lowest median rotation errors, about half of those of the Canon 24 mm and the GoPro images (Table 15), consistent with their finer angular resolution.

With the combined reference data (Table 14), the median translation and rotation errors lie between those of the two single epochs for 19 and 18 of the 20 combinations, respectively (Table 15). At coarser thresholds,

**Table 12** Localizing consumer cameras against the 2023 reference campaign: percentage of poses within seven error thresholds per camera and local feature (S: SuperPoint, D: DISK, A: ALIKED, X: XFeat, R: RDD). MegaLoc with ±100 m position pre-filtering; 2048 pixels. Best per camera in bold.

| Cam | Feat | 0.01m 0.1° | 0.025m 0.25° | 0.05m 0.5° | 0.1m 1° | 0.2m 2° | 0.5m 5° | 1m 10° |
|---|---|---|---|---|---|---|---|---|
| Canon 24mm | S | **4.7** | **28.1** | 58.8 | 85.8 | 95.3 | 97.5 | 97.8 |
| | D | 2.0 | 12.6 | 33.0 | 55.6 | 69.0 | 74.4 | 75.7 |
| | A | 2.1 | 21.8 | 53.4 | 80.8 | 91.8 | 94.3 | 94.5 |
| | X | 1.2 | 7.1 | 22.2 | 43.4 | 54.8 | 61.7 | 64.7 |
| | R | 3.1 | 27.0 | **59.3** | **86.0** | **95.9** | **98.0** | **98.3** |
| Canon 35mm | S | 10.4 | 32.4 | 59.5 | **80.6** | 90.1 | 94.0 | 95.1 |
| | D | 3.6 | 14.5 | 27.0 | 44.9 | 51.0 | 56.8 | 58.0 |
| | A | 8.8 | 28.6 | 55.1 | 77.2 | 89.1 | 94.1 | 95.2 |
| | X | 1.9 | 9.5 | 22.4 | 35.6 | 46.2 | 52.3 | 55.5 |
| | R | **10.9** | **37.5** | **60.7** | 79.9 | **92.8** | **97.0** | **98.3** |
| GoPro | S | 6.2 | 37.3 | 71.8 | 91.2 | 96.9 | 98.5 | 99.2 |
| | D | 3.5 | 20.2 | 47.5 | 69.7 | 79.4 | 85.5 | 87.1 |
| | A | 3.9 | 30.0 | 66.7 | 88.6 | 95.8 | 97.3 | 98.1 |
| | X | 1.8 | 13.7 | 35.6 | 62.1 | 75.3 | 83.1 | 85.5 |
| | R | **6.4** | **40.2** | **72.3** | **92.9** | **98.4** | **99.5** | **99.6** |
| iPhone | S | 5.0 | **26.4** | **50.1** | **72.4** | **85.4** | **90.0** | **92.1** |
| | D | 1.8 | 9.3 | 19.3 | 33.2 | 44.4 | 50.2 | 52.9 |
| | A | **5.1** | 24.5 | 48.1 | 69.4 | 81.2 | 86.3 | 87.9 |
| | X | 0.8 | 6.0 | 18.1 | 36.4 | 54.9 | 68.0 | 72.5 |
| | R | 4.9 | 24.7 | 45.0 | 67.3 | 81.1 | 87.2 | 90.2 |

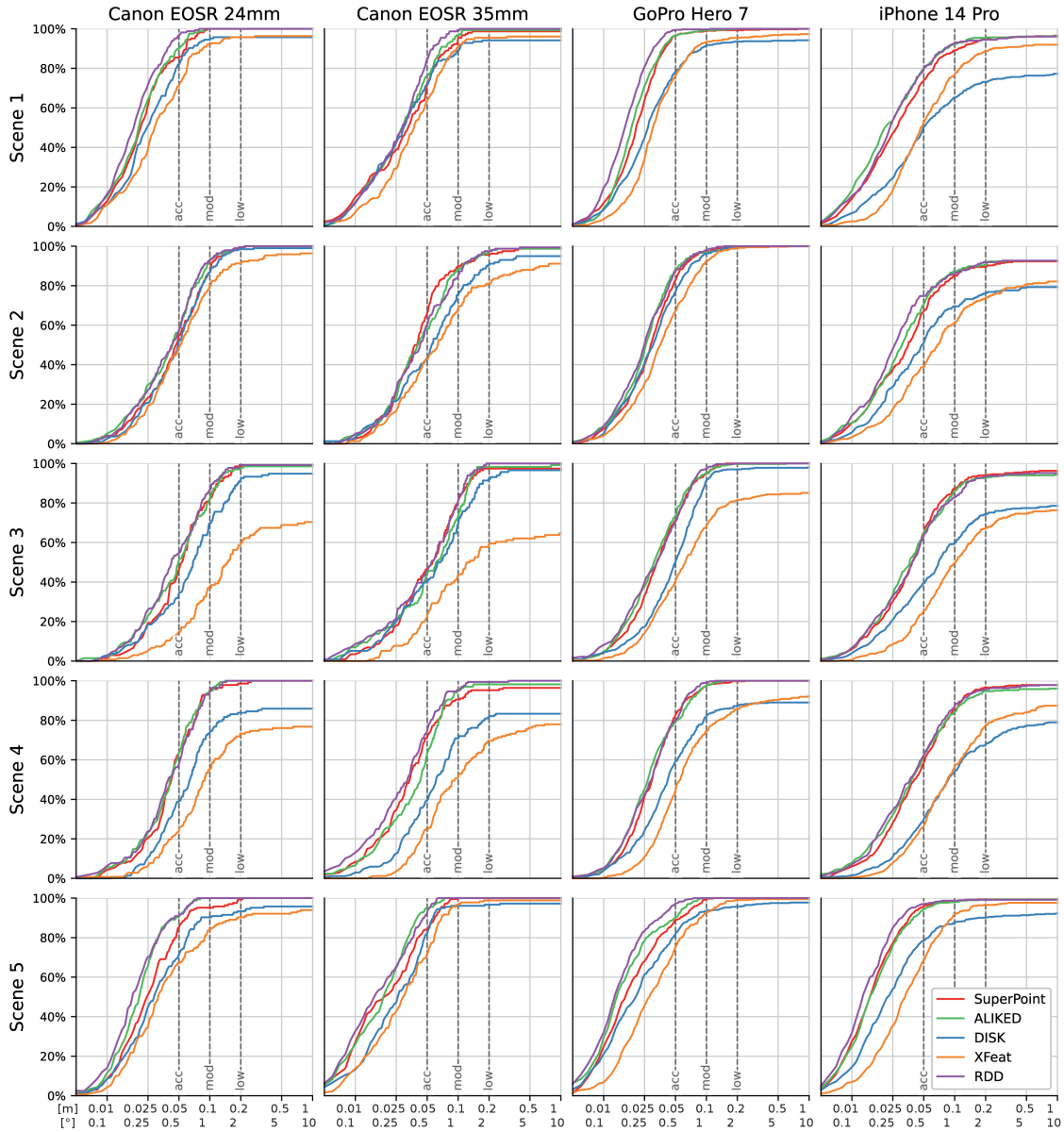


**Fig. 6** Localization performance per scene and camera against the 2025 reference imagery: percentage of poses with $\epsilon_t$ and $\epsilon_R$ below the value on the logarithmic x-axis (0.0005–1 m and 0.005–10°). MegaLoc with ±100 m position pre-filtering; image resolution 2048 pixels.

however, the combined data frequently perform worse than the 2023 data alone, for 13 of the 20 combinations at 1 m / 10° (e.g., 94.8 % vs. 99.2 % for the GoPro with SuperPoint), indicating more localization failures. Possible reasons for this behavior and its implications for updating the reference data are discussed in Section 5.2.

### 4.6 Local vs. global SfM-based visual localization

According to Panek et al. (2026), global SfM-based visual localization methods are among the most accurate. To quantify the accuracy cost of building a local SfM model on the fly from georeferenced images, we compared our pipeline with HLoc (Sarlin et al. 2019), which we consider a representative method of this category.

We localized the query images of all four cameras against the 2025 reference data, using identical settings for both pipelines. These are: MegaLoc with $k = 10$, ALIKED features at 1024 pixels, and HLoc's default RANSAC threshold of 12 pixels instead of the 2 pixels used in the other experiments. For each scene, we selected the reference images within ±100 m of the same manually selected coordinate as in Section 4.5 and triangulated a globally consistent SfM model from their exhaustive matches, as required by HLoc. For each query

**Table 13** Localizing consumer cameras against the 2025 reference campaign: percentage of poses within seven error thresholds per camera and local feature (S: SuperPoint, D: DISK, A: ALIKED, X: XFeat, R: RDD). MegaLoc with ±100 m position pre-filtering; 2048 pixels. Best per camera in bold.

| Cam | Feat | 0.01m 0.1° | 0.025m 0.25° | 0.05m 0.5° | 0.1m 1° | 0.2m 2° | 0.5m 5° | 1m 10° |
|---|---|---|---|---|---|---|---|---|
| Canon 24mm | S | 6.2 | 35.6 | 67.5 | 92.0 | 99.1 | 99.8 | 99.8 |
| | D | 5.1 | 29.7 | 56.8 | 83.9 | 93.1 | 94.6 | 94.6 |
| | A | 8.1 | 42.1 | 71.0 | 94.4 | 99.4 | 99.8 | 99.8 |
| | X | 3.5 | 21.7 | 46.9 | 71.5 | 83.5 | 87.1 | 88.0 |
| | R | **8.7** | **44.6** | **72.2** | **95.0** | **99.8** | **99.9** | **99.9** |
| Canon 35mm | S | 12.1 | 36.7 | 69.9 | 92.1 | 97.4 | 98.3 | 98.3 |
| | D | 7.6 | 29.7 | 56.5 | 81.1 | 91.1 | 93.1 | 93.1 |
| | A | 12.0 | 38.1 | 69.5 | 92.0 | 98.6 | 99.0 | 99.1 |
| | X | 4.9 | 21.2 | 47.0 | 71.8 | 82.4 | 85.7 | 87.1 |
| | R | **15.2** | **41.4** | **73.7** | **92.9** | **99.4** | **99.9** | **99.9** |
| GoPro | S | 10.9 | 49.8 | 84.7 | 97.6 | 99.7 | 99.9 | **100.0** |
| | D | 8.7 | 38.3 | 69.7 | 90.9 | 94.6 | 95.6 | 95.8 |
| | A | 11.6 | 56.7 | 85.7 | 97.7 | 99.8 | 99.9 | **100.0** |
| | X | 3.2 | 27.5 | 61.4 | 84.3 | 92.2 | 94.3 | 95.0 |
| | R | **15.3** | **60.1** | **87.4** | **98.9** | **100.0** | **100.0** | **100.0** |
| iPhone | S | 13.8 | 46.4 | 73.6 | 89.6 | **95.4** | **96.7** | **96.9** |
| | D | 7.0 | 28.9 | 51.2 | 68.1 | 76.9 | 80.8 | 81.9 |
| | A | 14.6 | 49.5 | 76.0 | 90.2 | 95.1 | 96.0 | 96.2 |
| | X | 2.9 | 18.5 | 44.2 | 69.2 | 82.4 | 86.9 | 88.2 |
| | R | **16.0** | **52.9** | **76.7** | **90.3** | **95.4** | 96.4 | 96.6 |

**Table 14** Localizing consumer cameras against the combined 2023 and 2025 reference campaigns: percentage of poses within seven error thresholds per camera and local feature (S: SuperPoint, D: DISK, A: ALIKED, X: XFeat, R: RDD). MegaLoc with ±100 m position pre-filtering; image resolution 2048 pixels. Best per camera in bold.

| Cam | Feat | 0.01m 0.1° | 0.025m 0.25° | 0.05m 0.5° | 0.1m 1° | 0.2m 2° | 0.5m 5° | 1m 10° |
|---|---|---|---|---|---|---|---|---|
| Canon 24mm | S | 5.0 | 29.1 | 62.8 | **88.0** | **94.3** | **96.0** | **97.3** |
| | D | 4.9 | 22.7 | 48.7 | 74.7 | 83.5 | 85.5 | 87.8 |
| | A | 5.9 | 33.5 | 61.8 | 85.6 | 91.3 | 92.8 | 94.4 |
| | X | 2.2 | 16.5 | 39.8 | 60.8 | 72.3 | 77.3 | 80.8 |
| | R | **6.9** | **38.0** | **64.9** | 87.3 | 92.4 | 93.5 | 94.8 |
| Canon 35mm | S | 10.4 | 29.9 | 63.3 | 88.2 | 93.7 | 94.5 | 96.1 |
| | D | 5.3 | 24.5 | 51.2 | 73.6 | 84.3 | 86.0 | 87.1 |
| | A | 9.8 | 31.7 | 64.2 | 88.2 | 92.3 | 93.3 | 94.5 |
| | X | 4.8 | 19.3 | 42.3 | 65.6 | 75.5 | 79.8 | 82.6 |
| | R | **12.1** | **37.1** | **68.3** | **89.8** | **94.7** | **95.6** | **96.4** |
| GoPro | S | 6.3 | 36.1 | 68.0 | 86.0 | **91.8** | **93.2** | **94.8** |
| | D | 4.8 | 26.2 | 51.4 | 68.8 | 76.7 | 81.2 | 83.9 |
| | A | 5.6 | 35.8 | 67.7 | 83.6 | 88.2 | 89.6 | 90.7 |
| | X | 2.1 | 16.4 | 39.7 | 63.6 | 75.6 | 80.8 | 83.3 |
| | R | **8.2** | **42.9** | **71.6** | **86.5** | 90.1 | 91.4 | 92.5 |
| iPhone | S | 8.9 | 32.7 | 55.3 | **71.5** | **81.1** | **86.1** | **88.8** |
| | D | 5.1 | 21.0 | 34.0 | 46.7 | 54.6 | 60.6 | 64.6 |
| | A | 11.7 | 37.1 | **57.8** | 71.2 | 78.5 | 83.2 | 85.7 |
| | X | 2.2 | 11.9 | 29.2 | 47.4 | 60.6 | 69.5 | 74.2 |
| | R | **12.6** | **37.3** | 56.5 | 68.7 | 76.9 | 82.7 | 85.5 |

image, MegaLoc retrieved the 10 most similar images from this subset, so that both pipelines operated on the same retrieved set. HLoc derives the 2D–3D correspondences from the precomputed 3D points, whereas our pipeline triangulates the local model on the fly from the 10 retrieved images.

Consistent with Panek et al. (2026), HLoc outperforms our pipeline in the high-accuracy regime, but the margin depends strongly on the query camera (Table 16). At 0.05 m / 0.5° it leads by 15.8 pp for both Canon lenses and by only 1.1 pp for the GoPro and 1.2 pp for the iPhone. At 0.025 m / 0.25° the spread is wider, from 23.9 pp for the Canon 35 mm to 3.4 pp for the GoPro, whereas for the iPhone our pipeline is 1.1 pp ahead. From 0.1 m / 1° onward the two pipelines are largely on par, with differences of at most 5.6 pp and a small advantage for our pipeline with the GoPro images. Per scene and camera, HLoc leads in the accurate category in 15 of the 20 combinations, in all ten for the two Canon lenses but in only five of ten for the GoPro and the iPhone. A precomputed, globally consistent model thus yields more accurate 3D points. This improves the estimated pose only where the 3D point coordinates are the dominant source of pose error. It has little effect where the geometry of the retrieved images or the quality of the query imagery dominates instead.

## 5 Discussion

### 5.1 Accuracy potential of sparse feature-based visual localization

For daytime imagery and pre-calibrated cameras, the best-performing local features achieve median pose errors of 1–5 cm in translation and 0.03–0.1° in rotation (Tables 11 and 15). Reaching the lower end of that range requires image resolutions of at least 2048 pixels and a suitable local feature representation, while the corresponding absolute accuracy additionally depends on the georeferencing of the scene representation. We first discuss the geometric factors influencing pose accuracy, then the local feature representations and what their differences imply for the evaluation of visual localization, and finally the scene representation and the relation between relative and absolute accuracy.

In PnP-based pose estimation, the accuracy of the resulting pose depends mainly on the angular accuracy with which the 3D points are observed, their distribution in image and object space, and the accuracy of their coordinates. Increasing the angular resolution of the image observations by using high-resolution images substantially improved the accuracy of the final poses (Table 11 and Figure 5). Doubling the image resolution

**Table 15** Median translation ($\epsilon_t$) and rotation ($\epsilon_R$) errors per reference imagery, camera and local feature (S: SuperPoint, D: DISK, A: ALIKED, X: XFeat, R: RDD). MegaLoc with ±100 m position pre-filtering; image resolution 2048 pixels. Best entry per camera in bold.

| Cam | Feat | 2023 | | 2025 | | 2023 & 2025 | |
|---|---|---|---|---|---|---|---|
| | | $\epsilon_t$ [m] | $\epsilon_R$ [°] | $\epsilon_t$ [m] | $\epsilon_R$ [°] | $\epsilon_t$ [m] | $\epsilon_R$ [°] |
| Canon 24mm | S | 0.042 | **0.092** | 0.035 | 0.073 | 0.039 | 0.085 |
| | D | 0.081 | 0.143 | 0.043 | 0.078 | 0.052 | 0.097 |
| | A | 0.046 | 0.099 | 0.030 | **0.069** | 0.037 | **0.082** |
| | X | 0.147 | 0.249 | 0.054 | 0.104 | 0.071 | 0.124 |
| | R | **0.041** | 0.096 | **0.029** | 0.070 | **0.034** | **0.082** |
| Canon 35mm | S | 0.040 | 0.055 | 0.034 | 0.034 | 0.038 | 0.038 |
| | D | 0.179 | 0.142 | 0.042 | 0.040 | 0.049 | 0.045 |
| | A | 0.044 | **0.053** | 0.032 | **0.032** | 0.038 | **0.036** |
| | X | 0.280 | 0.362 | 0.054 | 0.058 | 0.061 | 0.061 |
| | R | **0.037** | 0.054 | **0.031** | **0.032** | **0.035** | 0.037 |
| GoPro | S | 0.032 | 0.086 | 0.025 | 0.063 | 0.033 | 0.087 |
| | D | 0.053 | 0.106 | 0.033 | 0.067 | 0.047 | 0.107 |
| | A | 0.036 | 0.087 | 0.022 | **0.060** | 0.033 | **0.082** |
| | X | 0.069 | 0.144 | 0.039 | 0.082 | 0.065 | 0.136 |
| | R | **0.031** | **0.085** | **0.020** | 0.061 | **0.029** | 0.083 |
| iPhone | S | **0.050** | **0.094** | 0.028 | 0.056 | 0.042 | 0.084 |
| | D | 0.472 | 0.533 | 0.048 | 0.068 | 0.127 | 0.160 |
| | A | 0.052 | 0.098 | 0.025 | **0.053** | **0.037** | **0.078** |
| | X | 0.165 | 0.243 | 0.057 | 0.084 | 0.114 | 0.163 |
| | R | 0.058 | 0.104 | **0.023** | **0.053** | 0.038 | 0.083 |

**Table 16** Local SfM (ours) versus global SfM (HLoc; Sarlin et al. 2019): percentage of poses localized against the 2025 reference campaign within seven error thresholds per camera, aggregated over the five scenes. Both methods use MegaLoc ($k = 10$), ALIKED at 1024 pixels and a RANSAC threshold of 12 pixels (HLoc default).

| Camera | 0.01m 0.1° | 0.025m 0.25° | 0.05m 0.5° | 0.1m 1° | 0.2m 2° | 0.5m 5° | 1m 10° |
|---|---|---|---|---|---|---|---|
| *HLoc (Sarlin et al. 2019)* | | | | | | | |
| Canon 24 | 6.8 | 43.5 | 78.5 | 96.7 | 99.9 | 100.0 | 100.0 |
| Canon 35 | 14.1 | 52.6 | 78.5 | 94.6 | 98.7 | 99.5 | 99.6 |
| GoPro | 9.6 | 47.6 | 81.6 | 95.9 | 99.5 | 100.0 | 100.0 |
| iPhone | 6.7 | 32.4 | 61.5 | 87.2 | 95.1 | 97.0 | 97.4 |
| *Ours* | | | | | | | |
| Canon 24 | 4.7 | 32.8 | 62.7 | 91.1 | 97.6 | 99.5 | 99.5 |
| Canon 35 | 6.6 | 28.7 | 62.7 | 91.4 | 97.8 | 99.1 | 99.2 |
| GoPro | 7.6 | 44.2 | 80.5 | 96.8 | 99.8 | 99.9 | 99.9 |
| iPhone | 6.1 | 33.5 | 60.3 | 84.9 | 94.2 | 96.2 | 96.6 |

from 1024 to 2048 pixels reduces the median pose error by roughly 25 to 45 % for most tested local feature representations, and further increases can yield additional, though smaller, improvements depending on the representation. The localization accuracy of the keypoints also matters, as sub-pixel-accurate keypoints yield more accurate poses (Lindenberger et al. 2021; Kim et al. 2025; Vultaggio et al. 2024), especially at low image resolutions. Among the learned representations, ALIKED and RDD provide sub-pixel keypoint locations through a differentiable keypoint detection module (Zhao et al. 2022). Both lead in the high-accuracy regime. At the strictest threshold, RDD is ahead of all other representations at every tested resolution, while ALIKED matches or exceeds SuperPoint up to 4096 pixels (Table 11). SIFT also provides sub-pixel keypoint locations, but its performance is limited by the sensitivity of its handcrafted descriptor to the appearance changes between the campaigns rather than by keypoint precision.

Nevertheless, keypoint precision alone does not explain the performance differences between local feature representations. Among the representations with pixel-level keypoints, DISK matches or exceeds ALIKED at the strictest threshold at 1024 and 2048 pixels, and SuperPoint does so at full resolution. Local features with sub-pixel keypoint localization therefore tend to lead at the strictest thresholds, although other factors also influence the resulting pose accuracy. Following Zhang et al. (2021), we hypothesize that the quantity and distribution of keypoints across the image plane are equally important, so that many well-distributed but less accurately localized correspondences can outperform a few highly accurate ones, particularly in geometrically unfavorable configurations, although our experiments do not isolate this effect. Wide field-of-view cameras can provide a favorable distribution of image observations, consistent with the lowest median translation errors of the GoPro images (Section 4.5), although isolating the effect of the field of view would require a controlled comparison with identical trajectories.

Learned feature representations typically degrade on out-of-domain data (Bonilla et al. 2025). The resolution-dependent optimum observed for most representations (Table 11) may therefore be attributed to the lack of high-resolution training data. SuperPoint (DeTone et al. 2018), which benefited most consistently from higher resolutions, is trained with synthetic homographic warps rather than SfM-derived correspondences. This may explain why it generalizes to high-resolution images and extracts repeatable, well-localized keypoints even on out-of-domain data. Recent methods such as RDD additionally leverage the global image context to produce more robust descriptors. Tiled feature extraction limits this context to the tile, which may explain why RDD does not improve beyond 2048 pixels at the two strictest thresholds, although it continues to benefit from higher resolution at the coarser ones.

With the consumer-grade query images, RDD is the strongest representation, but no representation is best across all cameras, metrics and scenes (Section 4.5). DISK and XFeat, which perform comparably to the other representations on standard benchmarks such as Aachen (Sattler et al. 2018), underperform with both

reference epochs. This is consistent with Bonilla et al. (2025), who argue that the image matching method should be selected carefully for the specific scenario. A similar accuracy potential was demonstrated by Lindenberger et al. (2021), Zhai et al. (2025), and Ress et al. (2025), but on datasets with few challenging conditions (ETH3D (Schops et al. 2017), 7-Scenes (Shotton et al. 2013), and a proprietary indoor dataset), which emphasizes the importance of robust image matching under challenging conditions.

The performance differences between the local feature representations also show that the datasets currently used to evaluate visual localization give only a limited indication of its accuracy potential. Their ground-truth poses are typically accurate to the decimeter level (Table 1), so methods cannot be evaluated reliably at fine error thresholds. On the day subset of Aachen Day–Night (Sattler et al. 2018), Chen et al. (2025) report a spread of only 2.7 pp between SuperPoint, DISK, ALIKED, XFeat and RDD at the strictest threshold of 0.25 m and 2°. In the accurate category of our experiments, by contrast, XFeat lags behind the best representation by 36 pp at 1024 pixels (Table 11) and, for the consumer cameras, by 32 to 38 pp and 25 to 33 pp against the 2023 and 2025 reference data (Tables 12 and 13). Such differences only become visible with ground-truth poses whose uncertainty is well below the evaluated error thresholds (Section 3.1.4).

Precomputed, globally consistent SfM models typically provide the most accurate 3D points but require extensive preprocessing, whereas meshes and city models abstract geometric and visual information. Our pipeline avoids both by triangulating the 3D points directly from the georeferenced high-resolution imagery, at a cost of up to 15.8 pp fewer poses within 0.05 m and 0.5° compared with a globally consistent SfM model (Table 16). We attribute this to the weaker triangulation geometry of the local model, and the camera dependence reported in Section 4.6 supports this interpretation. The penalty is largest for the two Canon lenses and absent for the GoPro, whose wide field of view already provides a favorable distribution of image observations among the few retrieved images, whereas for the iPhone the pose error is dominated by the query imagery rather than by the accuracy of the 3D points.

The errors reported above exclude the georeferencing error of the reference imagery. Combined with the georeferencing RMSE of 0.025 m (Section 3.1.3), the median relative errors of 1–5 cm indicate absolute errors of approximately 3–6 cm, dominated by the georeferencing for the most accurate poses. As this combines medians with an RMSE, the comparison with survey-grade GNSS positioning is indicative rather than strict. Approaching survey-grade accuracy requires both accurate relative localization and an accurately georeferenced scene representation, and is ultimately bounded by the latter.

### 5.2 Reference epoch and map maintenance

For all cameras and feature representations, localization against the 2025 reference imagery is more accurate than against the 2023 reference imagery, and combining both epochs yields median errors in between (Tables 12–15). This ordering cannot be attributed to a single factor, as the two campaigns differ in several respects at once. The 2025 campaign was acquired approximately one month after the query images, whereas the 2023 campaign preceded them by about 16 months, during which ongoing construction work changed parts of the street environment. The vegetation (trees free of foliage) of the 2025 campaign matches that of the query images, and its illumination (overcast, diffuse light) that of the two query scenes captured under cloudy conditions, whereas the 2023 campaign was acquired under predominantly sunny conditions with shadow patterns and foliage. In addition, the sensor setup changed between the campaigns, including a panorama camera with a substantially higher resolution in 2025 (Table 2). Our experiments therefore show the combined effect of these factors but cannot separate them.

Adding reference data does not necessarily improve localization, as combining both epochs performs worse than using the 2025 imagery alone. We see two plausible explanations, which our experiments do not separate. Both lie in the reference candidate selection, which retrieves a fixed number of the most similar reference images ($k = 10$). First, with both epochs in the database, some of the k retrieved images are from the 2023 campaign, which differ more in appearance from the query images and are likely to yield fewer correct correspondences, also among the reference images. Second, images of both epochs acquired at nearly the same positions may be retrieved together, so that the retrieved set covers fewer distinct viewpoints and the triangulation geometry of the local model becomes weaker. Geometric inconsistencies between the epochs, in contrast, are unlikely to be the cause, as both campaigns are co-registered with a 3D RMSE of 0.005 m at the independent tie points (Table 3). If confirmed, this would suggest selecting reference candidates not by visual similarity alone, but also by their geometric configuration, e.g., by requiring a minimum pose difference between the selected images.

For the maintenance of image-based scene representations, these results suggest that recent reference imagery should be preferred over accumulated imagery, e.g., by

replacing outdated images, by restricting the retrieval to the most recent epoch available in a region, or by map summarization and change-aware selection. Without such measures, the scene representation also grows without bounds. Updating and bounding long-term localization maps has been studied in robotics (Burki et al. 2018), and adapting such map management techniques is a necessary step toward consistently high-accuracy visual localization. Due to the lack of accurately co-registered multi-epoch reference data, this has not yet been investigated extensively for high-accuracy visual localization. The two co-registered reference epochs of our dataset provide a basis for such investigations.

### 5.3 Applications of accurate visual localization

In the following, we distinguish between applications that are directly supported by our experiments and prospective applications that our results motivate but do not test. Regarding applications that require accurate pose information with respect to a reference frame, the georeferencing of the scene representation is critical, as it determines the achievable absolute pose accuracy (Section 5.1). In addition, an up-to-date scene representation is needed to maintain consistently high pose accuracy (Section 5.2). Reference mapping therefore has to be repeated at regular intervals, and each new epoch requires accurate georeferencing and precise co-registration with the existing epochs. Our experiments directly demonstrate the localization of mobile mapping imagery against an earlier campaign. In the best case, about 86 % of the 2025 images were localized with errors below 0.025 m and 0.25° relative to the 2023 campaign (Table 11). Introducing such poses as observations in integrated georeferencing (Eugster et al. 2012) could reduce the need for manually measured GCPs in subsequent epochs, limiting the manual georeferencing effort largely to the initial reference mapping. More generally, visual localization can be seen as a method for the automated generation of control information. This applies to both co-registration and georeferencing, provided that a precisely georeferenced scene representation is available.

A prospective application is the acquisition of 3D geospatial data with consumer devices. Smartphones equipped with LiDAR sensors are emerging as a cost-effective and user-friendly alternative to traditional surveying equipment, but accurate georeferencing currently requires additional GNSS hardware (Rehak et al. 2025) or GCPs determined with traditional surveying methods. Visual localization could provide this georeferencing directly from the captured imagery. Our experiments support this only in part. The iPhone images were localized with median errors of 0.023 m and 0.053° (Table 15), but they were pre-calibrated (Section 5.4).

Neither use case necessarily requires real-time processing, as a few localized frames may be sufficient for georeferencing. The runtime of approximately 3.3 s per image is therefore not critical (Appendix A.2). The memory needed for the preprocessed global image descriptors and local features is more relevant, particularly when scaling to entire cities or states with millions of images (Appendix A.1).

### 5.4 Limitations and future work

While our experiments demonstrate the accuracy potential of visual localization, this potential is established under specific conditions that bound the scope of our findings. First, the query images were acquired in five scenes of a single suburban municipality, during daytime in late winter under cloudy and sunny conditions. Our dataset includes only mild weather-induced appearance changes and no imagery acquired under low-light conditions, such as at dusk or during nighttime (Table 1). The transfer of the accuracy potential to other environments, such as dense urban areas, and to a wider range of seasonal and illumination conditions therefore remains to be shown. Localization against the 2023 reference campaign nevertheless provides a first indication, as it combines an interval of about 16 months with a change of season and illumination and with intervening construction work. Under these conditions the best median translation error per camera rises from between 2.0 and 3.1 cm to between 3.1 and 5.0 cm (Table 15), a degradation by factors of 1.2 to 2.2. The reference database likewise covers a single 10 km street network, so the scalability of the pipeline to city- or region-wide databases was not evaluated beyond the memory and runtime figures reported in Appendix A. Second, all query images were pre-calibrated and undistorted using the camera calibration estimated during ground-truth generation, and the intrinsics were kept fixed during pose estimation. This is typically given for imagery captured by professional vision-based MMS, but not for consumer-grade devices, whose calibration is usually unknown or less accurate. The effect of inaccurate or unknown intrinsics was not evaluated. Third, the scene representation consists of high-resolution imagery acquired with professional mobile mapping systems and georeferenced with an accuracy of a few centimeters. Consequently, the reported results represent the accuracy potential under favorable conditions and may degrade with less accurate scene representations, strong lens distortions or more challenging acquisition conditions. The contributions of query and reference resolution were not separated,

as varying them independently would introduce a scale difference between the two and confound resolution with scale.

Fourth, ground truth and localization pipeline follow the same photogrammetric principle. Brachmann et al. (2021) and Zhang et al. (2021) found that localization methods resembling the ground-truth procedure tend to be favored. In our case, the two differ in software, features, matching and 3D points. No automatic correspondences between query and reference images enter the ground truth, which is connected to the reference imagery exclusively through manually measured points (Section 3.1.3). The same applies to the mobile mapping queries in Sections 4.3 and 4.4, as the two campaigns are connected exclusively via manually measured GCPs and TPs (Section 3.1.2). The two share only the reference image poses and the query camera calibration. The reference poses define the reference frame and are therefore inherent to the relative accuracy reported here. The calibration is estimated in the ground-truth adjustment and used to undistort the query images, so a biased calibration would shift ground-truth and estimated poses in a similar way. The sensitivity analysis in Section 3.1.4, however, shows no compensation between calibration and image poses and limits such correlated errors to the millimeter level.

Fifth, the evaluation is limited to sparse local features. Dense matchers (Edstedt et al. 2024), approaches built on dense correspondences such as ImLoc (Jiang et al. 2026), feature refinement (Lindenberger et al. 2021) and feed-forward reconstruction models (Leroy et al. 2025) were not evaluated. Most of these methods operate at substantially reduced image resolutions, and their computational and storage costs increase considerably at the resolutions considered here. Whether they can exploit high-resolution imagery to reach comparable or higher accuracy remains to be investigated.

In future research, we intend to address several of these limitations. First, we plan to extend the dataset with additional areas along the road network and more challenging conditions, such as diverse weather, seasonal changes and, in particular, imagery acquired during dusk and nighttime. Second, we aim to investigate the use of uncalibrated imagery from consumer-grade devices for georeferencing, with a focus on robust localization under multiple, simultaneously occurring adverse conditions.

## 6 Conclusion

In this paper, we investigated the accuracy potential of sparse feature-based visual localization using high-end street-level imagery. Due to the lack of a suitable dataset for this purpose, we introduced the FHNW Muttenz dataset, which is publicly available to foster future research. It comprises high-resolution imagery from two mobile mapping campaigns of a 10 km street network and query sequences acquired with four different cameras in five scenes. All images are co-registered to the reference imagery, and the ground-truth poses of the query images have an uncertainty of 6–9 mm in position and well below 0.1° in orientation. To evaluate the accuracy potential on this dataset, we used a modular visual localization pipeline built from established components. It relies on the simple scene representation of georeferenced imagery and combines a prior-guided reference candidate selection strategy with SfM for building a local model on the fly and subsequently estimating the query image's pose. For daytime imagery and precalibrated cameras, our experiments yield median pose errors of 1–5 cm in translation and 0.03–0.1° in rotation relative to the reference imagery. Furthermore, the experiments revealed performance differences between local feature representations that are not resolved at the coarse error thresholds of existing benchmarks and only become visible with highly accurate ground-truth poses. The experiments further show that localization accuracy depends on the epoch of the reference imagery, and that accumulating imagery from several epochs does not improve it. Maintaining an image-based scene representation therefore requires replacing or selecting reference imagery rather than merely adding to it. The absolute accuracy is additionally bounded by the georeferencing of the scene representation, which is about 2.5 cm in our case. Given an accurately georeferenced scene representation, visual localization can therefore approach the accuracy of survey-grade GNSS positioning, with the advantage of yielding 6-DoF poses in any environment mapped in advance. This motivates its use as a complementary positioning modality to GNSS and for the automated georeferencing of imagery, which traditionally involves a substantial amount of manual effort. Extending these findings to uncalibrated consumer devices, low-light conditions and other environments remains future work.

## Declarations

*Acknowledgements* Not applicable

*Funding* This research was supported by the Forschungsfonds Kanton Aargau and our industrial partner iNovitas AG through Grants 20200331_04_PReCIS and 20240930_16_GRAVIS3D, as well as by Innosuisse via the Innovation Booster Artificial Intelligence program under Grant 03.10.2024-07.

*Conflicts of interest/Competing interests* The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

*Availability of data and material* The data used in this work is publicly available at: `https://fhnw-muttenz-vl-dataset.github.io/`.

*Code availability* The visual localization pipeline is composed of publicly available components. The methods used and all settings deviating from their defaults are given in Section 4.2, so that the experiments can be reproduced from the published implementations and the released dataset.

*Authors' contributions* **J.M.:** Conceptualization, Methodology, Validation, Investigation, Data curation, Writing – original draft, Visualization. **S.N.:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. **P.T.:** Writing – review & editing, Resources, Data curation. **N.H.:** Methodology, Writing – review & editing, Supervision.

## A Memory requirements and runtime

### A.1 Memory consumption

Our visual localization pipeline requires precomputed global descriptors stored in a vector database. For runtime and efficiency reasons, we also use precomputed local features. The memory requirements of both the global image descriptors and the local features may therefore be of interest when selecting suitable methods. This is particularly true when the environment spans entire cities or states. Table 17 reports the memory footprint of a single global image descriptor for each IR method. Analogously, Table 18 presents the memory needed to store 1024 sparse local features per representation. All values reflect uncompressed memory usage to ensure fair and consistent comparison.

**Table 17** Memory requirement of one global descriptor per IR method.

| Method | Desc dim | dtype | Data [KB] |
|---|---|---|---|
| NetVLAD | 4096 | float32 | 16 |
| CosPlace | 512 \| 2048 | float32 | 2 \| 8 |
| EigenPlaces | 512 \| 2048 | float32 | 2 \| 8 |
| AnyLoc | 49 152 | float32 | 192 |
| DINOv2 SALAD | 8448 | float32 | 33 |
| MegaLoc | 8448 | float32 | 33 |

**Table 18** Memory requirement to store 1024 sparse local features per representation.

| Method | Descriptor length | Descriptor dtype | Keypoint dtype | Data per 1024 features [KB] |
|---|---|---|---|---|
| SuperPoint | 256 | float32 | int16 | 1028 |
| DISK | 128 | float32 | int16 | 516 |
| ALIKED | 128 | float32 | float32 | 520 |
| XFeat | 64 | float32 | int16 | 260 |
| RDD | 256 | float32 | float32 | 1032 |

### A.2 Runtime

The runtime evaluation was conducted on a workstation equipped with an AMD Ryzen Threadripper 7970X 32-Core @ 4.0 GHz CPU and an NVIDIA RTX 4090 GPU. The runtime of our visual localization pipeline strongly depends on the chosen settings, including the local feature representation, image resolution, and the number of features extracted. We provide the runtime of each major pipeline step and its outputs with respect to time-relevant settings.

We randomly selected 1000 images from the front-facing monocular camera of the 2025 campaign as a test set. These images remained unchanged across all runtime evaluations. We applied our pipeline with settings identical to those used in the experiment investigating the role of high-resolution imagery (Section 4.4) and measured the time.

Tables 19, 20, and 21 report the average runtime for extracting global image descriptors, sparse local features, and image matching, respectively. Table 22 reports the average runtime for our local SfM-based pose estimation procedure. In addition to the total runtime (Total), we also report sub-part runtimes, such as geometric verification (G), point triangulation from known poses and camera intrinsics (T), and pose estimation via PnP+RANSAC followed by non-linear pose refinement (PnP) (Table 22).

**Table 19** Average runtime [ms] for extracting one global descriptor per IR method.

| Method | Time [ms] |
|---|---|
| NetVLAD | 51.4 |
| CosPlace | 50.6 |
| EigenPlaces | 50.3 |
| AnyLoc | 101.6 |
| DINOv2 SALAD | 56.7 |
| MegaLoc | 56.2 |

**Table 20** Average runtime [ms] of the tiled feature extraction for one image per resolution and local feature (S: SuperPoint, D: DISK, A: ALIKED, X: XFeat, R: RDD).

| Image res | S | D | A | X | R |
|---|---|---|---|---|---|
| 1024 | 75.6 | 97.9 | 85.4 | 80.5 | 196.5 |
| 2048 | 122.4 | 181.8 | 137.8 | 117.8 | 644.8 |
| 4096 | 243.3 | 451.0 | 269.0 | 202.1 | 2347.0 |
| full | 318.7 | 698.6 | 373.6 | 249.6 | 4114.5 |

**Table 21** Average matching runtime [ms] per number of features for one image pair (P) and for exhaustive matching of 11 images, i.e., 55 pairs (EM). Matchers: LightGlue (Lindenberger et al. 2023); LighterGlue (Potje et al. 2024) for XFeat.

| # Features | LightGlue [ms] | | LighterGlue [ms] | |
|---|---|---|---|---|
| | P | EM | P | EM |
| 1024 | 8.7 | 476.3 | 5.7 | 315.7 |
| 2048 | 10.7 | 586.9 | 6.9 | 379.5 |
| 4096 | 19.5 | 1071.4 | 7.8 | 429.6 |
| 6144 | 34.5 | 1899.7 | 11.1 | 612.2 |
| 8192 | 45.8 | 2520.1 | 18.1 | 996.6 |
| 10240 | 70.6 | 3885.2 | 23.9 | 1314.5 |
| 12288 | 89.0 | 4893.4 | 29.1 | 1603.3 |
| 14336 | 120.2 | 6613.8 | 36.3 | 1997.6 |

**Table 22** Average runtime [ms] of the local SfM-based pose estimation with 10 reference images per number of features. Geometric verification (G), triangulation (T), PnP-based pose estimation (PnP) and total. Precomputed SuperPoint features, matched with LightGlue.

| # Features | G | T | PnP | Total |
|---|---|---|---|---|
| 1024 | 143.4 | 140.8 | 55.3 | 388.3 |
| 2048 | 170.4 | 269.6 | 82.6 | 576.5 |
| 4096 | 261.2 | 517.3 | 130.0 | 974.0 |
| 6144 | 315.0 | 639.1 | 149.8 | 1178.3 |
| 8192 | 362.6 | 756.1 | 164.5 | 1365.3 |
| 10 240 | 412.5 | 942.9 | 188.9 | 1634.7 |
| 12 288 | 455.2 | 1111.6 | 207.9 | 1872.6 |
| 14 336 | 486.3 | 1242.9 | 220.7 | 2056.7 |

To obtain the approximate runtime of the entire pipeline, the runtimes in Tables 19, 20, 21, and 22 for the desired methods must be summed. For example, using MegaLoc as the IR method and ALIKED as the local feature representation with 6144 features at an image resolution of 2048, our pipeline takes approximately 3.3 seconds to localize a query image. Note that loading the query image and the preprocessed local features as well as querying the vector database are neglected in the runtime ablation study. Querying the vector database typically takes 10–20 ms and is therefore negligible. Loading times are also excluded, as the preprocessed features were stored on external drives in our setup, and the resulting I/O cost is not representative of an optimized deployment.

## References


Agisoft (2024) Metashape Professional, version 2.2.1. Agisoft LLC, St. Petersburg, Russia, URL `https://www.agisoft.com`

Ali-Bey A, Chaib-Draa B, Giguère P (2022) GSV-Cities: Toward appropriate supervised visual place recognition. Neurocomputing 513:194–203, DOI 10.1016/j.neucom.2022.09.127

Ali-Bey A, Chaib-Draa B, Giguère P (2023) MixVPR: Feature Mixing for Visual Place Recognition. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, pp 2997–3006, DOI 10.1109/WACV56688.2023.00301

Arandjelovic R, Gronat P, Torii A, Pajdla T, Sivic J (2016) NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, pp 5297–5307, DOI 10.1109/CVPR.2016.572

Arth C, Wagner D, Klopschitz M, Irschara A, Schmalstieg D (2009) Wide area localization on mobile phones. In: 2009 8th IEEE International Symposium on Mixed and Augmented Reality, IEEE, Orlando, FL, USA, pp 73–82, DOI 10.1109/ISMAR.2009.5336494

Badino H, Huber D, Kanade T (2011) Visual topometric localization. In: 2011 IEEE Intelligent Vehicles Symposium (IV), IEEE, Baden-Baden, Germany, pp 794–799, DOI 10.1109/IVS.2011.5940504

Bellavia F, Matas J, Mishkin D, Morelli L, Remondino F, Tabb A, Trulls E, Yi KM, Dane S, Howard A, Cruz M (2025) Image matching challenge 2025. `https://kaggle.com/competitions/image-matching-challenge-2025`, Kaggle

Berton G, Masone C (2025) MegaLoc: One Retrieval to Place Them All. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Nashville, TN, USA, pp 2852–2858, DOI 10.1109/CVPRW67362.2025.00269

Berton G, Masone C, Caputo B (2022) Rethinking Visual Geo-localization for Large-Scale Applications. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA, pp 4868–4878, DOI 10.1109/CVPR52688.2022.00483

Berton G, Trivigno G, Caputo B, Masone C (2023) EigenPlaces: Training Viewpoint Robust Models for Visual Place Recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 11046–11056, DOI 10.1109/ICCV51070.2023.01017

Bieringer A, Wysocki O, Tuttas S, Hoegner L, Holst C (2024) Analyzing the impact of semantic LoD3 building models on image-based vehicle localization. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences X-4/W5-2024:55–62, DOI 10.5194/isprs-annals-X-4-W5-2024-55-2024

Blum H, Mercurio A, O'Reilly J, Engelbracht T, Dusmanu M, Pollefeys M, Bauer Z (2025) CroCoDL: Cross-device Collaborative Dataset for Localization. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, pp 27424–27434, DOI 10.1109/CVPR52734.2025.02554

Bonilla S, Di Vece C, Daher R, Ju X, Stoyanov D, Vasconcelos F, Bano S (2025) Mismatched: Evaluating the Limits of Image Matching Approaches and Benchmarks. In: Del Bue A, Canton C, Pont-Tuset J, Tommasi T (eds) Computer Vision – ECCV 2024 Workshops, vol 15645, Springer Nature Switzerland, Cham, pp 120–137, DOI 10.1007/978-3-031-91989-3_8

Brachmann E, Rother C (2018) Learning Less is More - 6D Camera Localization via 3D Surface Regression. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, pp 4654–4662, DOI 10.1109/CVPR.2018.00489

Brachmann E, Humenberger M, Rother C, Sattler T (2021) On the Limits of Pseudo Ground Truth in Visual Camera Re-localisation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp 6198–6208, DOI 10.1109/ICCV48922.2021.00616

Bradski G (2000) The OpenCV Library. Dr Dobb's Journal of Software Tools

Brown DC (1971) Close-range camera calibration. Photogrammetric Engineering 37(8):855–866

Burki M, Dymczyk M, Gilitschenski I, Cadena C, Siegwart R, Nieto J (2018) Map Management for Efficient Long-Term Visual Localization in Outdoor Environments. In: 2018 IEEE Intelligent Vehicles Symposium (IV), IEEE, Changshu, pp 682–688, DOI 10.1109/IVS.2018.8500432

Cai R, Tung J, Wang Q, Averbuch-Elor H, Hariharan B, Snavely N (2023) Doppelgangers: Learning to Disambiguate Images of Similar Structures. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Paris, France, pp 34–44, DOI 10.1109/ICCV51070.2023.00010

Carlevaris-Bianco N, Ushani AK, Eustice RM (2016) University of Michigan North Campus long-term vision and lidar dataset. The International Journal of Robotics Research 35(9):1023–1035, DOI 10.1177/0278364915614638

Cavegn S, Blaser S, Nebiker S, Haala N (2018) Robust And Accurate Image-Based Georeferencing Exploiting Relative Orientation Constraints. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences IV-2:57–64, DOI 10.5194/isprs-annals-IV-2-57-2018

Chen DM, Baatz G, Koser K, Tsai SS, Vedantham R, Pylvanainen T, Roimela K, Chen X, Bach J, Pollefeys M, Girod B, Grzeszczuk R (2011) City-scale landmark identification on mobile devices. In: CVPR 2011, IEEE, Colorado Springs, CO, USA, pp 737–744, DOI 10.1109/CVPR.2011.5995610

Chen G, Fu T, Chen H, Teng W, Xiao H, Zhao Y (2025) RDD: Robust Feature Detector and Descriptor using Deformable Transformer. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, pp 6394–6403, DOI 10.1109/CVPR52734.2025.00599

Chen S, Bhalgat Y, Li X, Bian JW, Li K, Wang Z, Prisacariu VA (2024a) Neural refinement for absolute pose regression with feature synthesis. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 20987–20996, DOI 10.1109/CVPR52733.2024.01983

Chen S, Cavallari T, Prisacariu VA, Brachmann E (2024b) Map-relative pose regression for visual re-localization. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 20665–20674, DOI 10.1109/CVPR52733.2024.01953

Chum O, Matas J, Kittler J (2003) Locally Optimized RANSAC. In: Goos G, Hartmanis J, Van Leeuwen J, Michaelis B, Krell G (eds) Pattern Recognition, vol 2781, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 236–243, DOI 10.1007/978-3-540-45243-0_31

DeTone D, Malisiewicz T, Rabinovich A (2018) Superpoint: Self-supervised interest point detection and description. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 337–33712, DOI 10.1109/CVPRW.2018.00060

Dong S, Liu S, Guo H, Chen B, Pollefeys M (2023) Lazy Visual Localization via Motion Averaging. arXiv preprint DOI 10.48550/arXiv.2307.09981

Edstedt J, Sun Q, Bökman G, Wadenbäck M, Felsberg M (2024) RoMa: Robust Dense Feature Matching. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp 19790–19800, DOI 10.1109/CVPR52733.2024.01871

Eugster H, Huber F, Nebiker S, Gisi A (2012) Integrated georeferencing of stereo image sequences captured with a stereovision mobile mapping system - approaches and practical results. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXIX-B1:309–314, DOI 10.5194/isprsarchives-XXXIX-B1-309-2012

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

Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6):381–395, DOI 10.1145/358669.358692

Fraser CS (1997) Digital camera self-calibration. ISPRS Journal of Photogrammetry and Remote Sensing 52(4):149–159, DOI 10.1016/S0924-2716(97)00005-1

Gaisbauer S, Gyawali P, Zhang Q, Wysocki O, Jutzi B (2025) To Glue or Not to Glue? Classical vs Learned Image Matching for Mobile Mapping Cameras to Textured Semantic 3D Building Models. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences X-1/W2-2025:35–42, DOI 10.5194/isprs-annals-X-1-W2-2025-35-2025

Haitz D, Shetty AS, Weinmann M, Ulrich M (2026) Evaluation of Visual Place Recognition Methods for Image Pair Retrieval in 3D Vision and Robotics. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences XI-2-2026:647–656, DOI 10.5194/isprs-annals-XI-2-2026-647-2026

Haralick BM, Lee CN, Ottenberg K, Nölle M (1994) Review and analysis of solutions of the three point perspective pose estimation problem. International Journal of Computer Vision 13(3):331–356, DOI 10.1007/BF02028352

Huang H, Liu C, Zhu Y, Cheng H, Braud T, Yeung SK (2024) 360Loc: A Dataset and Benchmark for Omnidirectional Visual Localization with Cross-Device Queries. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp 22314–22324, DOI 10.1109/CVPR52733.2024.02106

Humenberger M, Cabon Y, Pion N, Weinzaepfel P, Lee D, Guérin N, Sattler T, Csurka G (2022) Investigating the Role of Image Retrieval for Visual Localization: An Exhaustive Benchmark. International Journal of Computer Vision 130(7):1811–1836, DOI 10.1007/s11263-022-01615-7

Irschara A, Zach C, Frahm JM, Bischof H (2009) From structure-from-motion point clouds to fast location recognition. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 2599–2606, DOI 10.1109/CVPR.2009.5206587

Izquierdo S, Civera J (2024) Optimal Transport Aggregation for Visual Place Recognition. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp 17658–17668, DOI 10.1109/CVPR52733.2024.01672

Jafarzadeh A, Antequera ML, Gargallo P, Kuang Y, Toft C, Kahl F, Sattler T (2021) CrowdDriven: A New Challenging Dataset for Outdoor Visual Localization. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp 9825–9835, DOI 10.1109/ICCV48922.2021.00970

Jegou H, Douze M, Schmid C, Perez P (2010) Aggregating local descriptors into a compact image representation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, San Francisco, CA, USA, pp 3304–3311, DOI 10.1109/CVPR.2010.5540039

Jiang X, Wang F, Galliani S, Vogel C, Pollefeys M (2026) ImLoc: Revisiting Visual Localization with Image-based

Representation. arXiv preprint DOI 10.48550/arXiv.2601.04185

Jin Y, Mishkin D, Mishchuk A, Matas J, Fua P, Yi KM, Trulls E (2021) Image Matching Across Wide Baselines: From Paper to Practice. International Journal of Computer Vision 129(2):517–547, DOI 10.1007/s11263-020-01385-0

Keetha N, Mishra A, Karhade J, Jatavallabhula KM, Scherer S, Krishna M, Garg S (2024) AnyLoc: Towards Universal Visual Place Recognition. IEEE Robotics and Automation Letters 9(2):1286–1293, DOI 10.1109/LRA.2023.3343602

Kendall A, Grimes M, Cipolla R (2015) PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, Santiago, Chile, pp 2938–2946, DOI 10.1109/ICCV.2015.336

Kim S, Pollefeys M, Barath D (2025) Learning to Make Keypoints Sub-pixel Accurate. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G (eds) Computer Vision – ECCV 2024, vol 15136, Springer Nature Switzerland, Cham, pp 413–431, DOI 10.1007/978-3-031-73229-4_24

Laskar Z, Melekhov I, Kalia S, Kannala J (2017) Camera Relocalization by Computing Pairwise Relative Poses Using Convolutional Neural Network. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), IEEE, Venice, Italy, pp 920–929, DOI 10.1109/ICCVW.2017.113

Lee D, Ryu S, Yeon S, Lee Y, Kim D, Han C, Cabon Y, Weinzaepfel P, Guerin N, Csurka G, Humenberger M (2021) Large-scale Localization Datasets in Crowded Indoor Spaces. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, pp 3226–3235, DOI 10.1109/CVPR46437.2021.00324

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

Leroy V, Cabon Y, Revaud J (2025) Grounding Image Matching in 3D with MASt3R. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G (eds) Computer Vision – ECCV 2024, vol 15130, Springer Nature Switzerland, Cham, pp 71–91, DOI 10.1007/978-3-031-73220-1_5

Lindenberger P, Sarlin PE, Larsson V, Pollefeys M (2021) Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp 5967–5977, DOI 10.1109/ICCV48922.2021.00593

Lindenberger P, Sarlin PE, Pollefeys M (2023) LightGlue: Local Feature Matching at Light Speed. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Paris, France, pp 17581–17592, DOI 10.1109/ICCV51070.2023.01616

Liu J, Nie Q, Liu Y, Wang C (2023) NeRF-Loc: Visual Localization with Conditional Neural Radiance Field. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, London, United Kingdom, pp 9385–9392, DOI 10.1109/ICRA48891.2023.10161420

Loeper Y, Gerke M, Alamouri A, Kern A, Bajauri MS, 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, DOI 10.5194/isprs-archives-XLVIII-2-W8-2024-311-2024

Lowe DG (2004) Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2):91–110, DOI 10.1023/B:VISI.0000029664.99615.94

Maddern W, Pascoe G, Linegar C, Newman P (2017) 1 year, 1000 km: The Oxford RobotCar dataset. The International Journal of Robotics Research 36(1):3–15, DOI 10.1177/0278364916679498

Maiwald F, Feurer D, Eltner A (2023) Solving photogrammetric cold cases using AI-based image matching: New potential for monitoring the past with historical aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 206:184–200, DOI 10.1016/j.isprsjprs.2023.11.008

Majdik AL, Verda D, Albers-Schoenberg Y, Scaramuzza D (2014) Micro air vehicle localization and position tracking from textured 3d cadastral models. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp 920–927, DOI 10.1109/ICRA.2014.6906964

Matteo B, Tsesmelis T, James S, Poiesi F, Del Bue A (2025) 6DGS: 6D Pose Estimation from a Single Image and a 3D Gaussian Splatting Model. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G (eds) Computer Vision – ECCV 2024, vol 15110, Springer Nature Switzerland, Cham, pp 420–436, DOI 10.1007/978-3-031-72943-0_24

Morelli L, Ioli F, Maiwald F, Mazzacca G, Menna F, Remondino F (2024) Deep-Image-Matching: A toolbox for multiview image matching of complex scenarios. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-2/W4-2024:309–316, DOI 10.5194/isprs-archives-xlviii-2-w4-2024-309-2024

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A, Assran M, Ballas N, Galuba W, Howes R, Huang PY, Li SW, Misra I, Rabbat M, Sharma V, Synnaeve G, Xu H, Jegou H, Mairal J, Labatut P, Joulin A, Bojanowski P (2024) DINOv2: Learning Robust Visual Features without Supervision. arXiv preprint DOI 10.48550/arXiv.2304.07193

Panek V, Kukelova Z, Sattler T (2022) MeshLoc: Mesh-Based Visual Localization. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T (eds) Computer Vision – ECCV 2022, vol 13682, Springer Nature Switzerland, Cham, pp 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. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Vancouver, BC, Canada, pp 13175–13186, DOI 10.1109/CVPR52729.2023.01266

Panek V, Zhou Q, Ding Y, Agostinho S, Kukelova Z, Sattler T, Leal-Taixé L (2026) A Guide to Structureless Visual Localization. International Journal of Computer Vision 134(6):263, DOI 10.1007/s11263-026-02780-9

Pix4D (2025) PIX4DCatch: 3D scanner, version 2.8.0 (build 82.1450). Pix4D S.A., Prilly, Switzerland, URL `https://www.pix4d.com/product/pix4dcatch/`

Potje G, Cadar F, Araujo A, Martins R, Nascimento ER (2024) XFeat: Accelerated Features for Lightweight Image Matching. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp 2682–2691, DOI 10.1109/CVPR52733.2024.00259

Puente I, González-Jorge H, Martínez-Sánchez J, Arias P (2013) Review of mobile mapping and surveying technologies. Measurement 46(7):2127–2145, DOI 10.1016/j.measurement.2013.03.006

Rehak M, Cucci D, Magnin JB, Strecha C (2025) Accurate Mapping of Subterranean Structures with Mobile Phones. The International Archives of the Photogrammetry, Remote Sensing and Spatial Informa-

tion Sciences XLVIII-G-2025:1269–1275, DOI 10.5194/isprs-archives-XLVIII-G-2025-1269-2025

Remondino F, Fraser C (2006) Digital camera calibration methods: Considerations and comparisons. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 36(5):266–272, DOI 10.3929/ETHZ-B-000158067

Ress V, Meyer J, Zhang W, Skuddis D, Soergel U, Haala N (2025) 3D Gaussian Splatting aided Localization for Large and Complex Indoor-Environments. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-G-2025:1283–1290, DOI 10.5194/isprs-archives-XLVIII-G-2025-1283-2025

Revaud J, Cabon Y, Brégier R, Lee J, Weinzaepfel P (2024) SACReg: Scene-agnostic coordinate regression for visual localization. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 688–698, DOI 10.1109/CVPRW63382.2024.00073

Sarlin PE, Cadena C, Siegwart R, Dymczyk M (2019) From coarse to fine: Robust hierarchical localization at large scale. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June:12708–12717, DOI 10.1109/CVPR.2019.01300

Sarlin PE, DeTone D, Malisiewicz T, Rabinovich A (2020) SuperGlue: Learning Feature Matching With Graph Neural Networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp 4937–4946, DOI 10.1109/CVPR42600.2020.00499

Sarlin PE, Unagar A, Larsson M, Germain H, Toft C, Larsson V, Pollefeys M, Lepetit V, Hammarstrand L, Kahl F, Sattler T (2021) Back to the Feature: Learning Robust Camera Localization from Pixels to Pose. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, pp 3246–3256, DOI 10.1109/CVPR46437.2021.00326

Sarlin PE, Dusmanu M, Schonberger JL, Speciale P, Gruber L, Larsson V, Miksik O, Pollefeys M (2022) LaMAR: Benchmarking localization and mapping for augmented reality. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII, Springer-Verlag, Berlin, Heidelberg, p 686–704, DOI 10.1007/978-3-031-20071-7_40

Sattler T, Weyand T, Leibe B, Kobbelt L (2012) Image retrieval for image-based localization revisited. In: BMVC, vol 1, p 4

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. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8601–8610, DOI 10.1109/CVPR.2018.00897

Sattler T, Zhou Q, Pollefeys M, Leal-Taixe L (2019) Understanding the Limitations of CNN-Based Absolute Camera Pose Regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, pp 3297–3307, DOI 10.1109/CVPR.2019.00342

Schops T, Schonberger JL, Galliani S, Sattler T, Schindler K, Pollefeys M, Geiger A (2017) A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, pp 2538–2547, DOI 10.1109/CVPR.2017.272

Schönberger JL, Frahm JM (2016) Structure-from-Motion Revisited. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, pp 4104–4113, DOI 10.1109/CVPR.2016.445

Shavit Y, Ferens R, Keller Y (2021) Learning multi-scene absolute pose regression with transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp 2713–2722, DOI 10.1109/ICCV48922.2021.00273

Shotton J, Glocker B, Zach C, Izadi S, Criminisi A, Fitzgibbon A (2013) Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Portland, OR, USA, pp 2930–2937, DOI 10.1109/CVPR.2013.377

Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos. In: Proceedings Ninth IEEE International Conference on Computer Vision, IEEE, Nice, France, pp 1470–1477 vol.2, DOI 10.1109/ICCV.2003.1238663

Sun X, Xie Y, Luo P, Wang L (2017) A Dataset for Benchmarking Image-Based Localization. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, pp 5641–5649, DOI 10.1109/CVPR.2017.598

Taira H, Okutomi M, Sattler T, Cimpoi M, Pollefeys M, Sivic J, Pajdla T, Torii A (2021) InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(4):1293–1307, DOI 10.1109/TPAMI.2019.2952114

Torii A, Arandjelovic R, Sivic J, Okutomi M, Pajdla T (2015) 24/7 place recognition by view synthesis. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, pp 1808–1817, DOI 10.1109/CVPR.2015.7298790

Torii A, Taira H, Sivic J, Pollefeys M, Okutomi M, Pajdla T, Sattler T (2021) Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? IEEE Transactions on Pattern Analysis and Machine Intelligence 43(3):814–829, DOI 10.1109/TPAMI.2019.2941876

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

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, DOI 10.5194/isprs-archives-XLVIII-2-W8-2024-447-2024

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, DOI 10.5194/isprs-archives-XLVIII-1-W4-2025-131-2025

Wang J, Chen M, Karaev N, Vedaldi A, Rupprecht C, Novotny D (2025) VGGT: Visual Geometry Grounded Transformer. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, pp 5294–5306, DOI 10.1109/CVPR52734.2025.00499

Wang S, Leroy V, Cabon Y, Chidlovskii B, Revaud J (2024) DUSt3R: Geometric 3D Vision Made Easy. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp 20697–20709, DOI 10.1109/CVPR52733.2024.01956

Wang Z, Bian W, Li X, Tao Y, Wang J, Fallon M, Prisacariu V (2026) Seeing in the dark: Benchmarking egocentric 3d vision with the oxford day-and-night dataset. Advances

in Neural Information Processing Systems 38

Wysocki O, Schwab B, Biswanath MK, Greza M, Zhang Q, Zhu J, Froech T, Heeramaglore M, Hijazi I, Kanna K, Pechinger M, Chen Z, Sun Y, Segura AR, Xu Z, AbdelGafar O, Mehranfar M, Yeshwanth C, Liu YC, Yazdi H, Wang J, Auer S, Anders K, Bogenberger K, Borrmann A, Dai A, Hoegner L, Holst C, Kolbe TH, Ludwig F, Nießner M, Petzold F, Zhu XX, Jutzi B (2026) TUM2TWIN: Introducing the large-scale multimodal urban digital twin benchmark dataset. ISPRS Journal of Photogrammetry and Remote Sensing 232:810–830, DOI 10.1016/j.isprsjprs.2025.12.013

Yan S, Liu Y, Wang L, Shen Z, Peng Z, Liu H, Zhang M, Zhang G, Zhou X (2023) Long-Term Visual Localization with Mobile Sensors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Vancouver, BC, Canada, pp 17245–17255, DOI 10.1109/CVPR52729.2023.01654

Yen-Chen L, Florence P, Barron JT, Rodriguez A, Isola P, Lin TY (2021) iNeRF: Inverting neural radiance fields for pose estimation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 1323–1330, DOI 10.1109/IROS51168.2021.9636708

Yu J, Zhu C, Zhang J, Huang Q, Tao D (2020) Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition. IEEE Transactions on Neural Networks and Learning Systems 31(2):661–674, DOI 10.1109/TNNLS.2019.2908982

Zhai H, Zhang X, Zhao B, Li H, He Y, Cui Z, Bao H, Zhang G (2025) SplatLoc: 3D Gaussian Splatting-based Visual Localization for Augmented Reality. IEEE Transactions on Visualization and Computer Graphics 31(5):3591–3601, DOI 10.1109/TVCG.2025.3549563

Zhang L, Tao Y, Lin J, Zhang F, Fallon M (2026) Visual localization in 3D maps: comparing point cloud, mesh, and NeRF representations. Autonomous Robots 50(1):14, DOI 10.1007/s10514-025-10232-5

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, DOI 10.1007/s11263-020-01399-8

Zhao X, Wu X, Miao J, Chen W, Chen PCY, Li Z (2022) ALIKE: Accurate and lightweight keypoint detection and descriptor extraction. IEEE Transactions on Multimedia 25:3101–3112, DOI 10.1109/TMM.2022.3155927

Zhao X, Wu X, Chen W, Chen PCY, Xu Q, Li Z (2023) ALIKED: A Lighter Keypoint and Descriptor Extraction Network via Deformable Transformation. IEEE Transactions on Instrumentation and Measurement 72:1–16, DOI 10.1109/TIM.2023.3271000

Zhou Q, Sattler T, Pollefeys M, Leal-Taixe L (2020) To Learn or Not to Learn: Visual Localization from Essential Matrices. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, Paris, France, pp 3319–3326, DOI 10.1109/ICRA40945.2020.9196607

Zhou Q, Maximov M, Litany O, Leal-Taixé L (2024) The NeRFect match: Exploring nerf features for visual localization. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXIV, Springer-Verlag, Berlin, Heidelberg, p 108–127, DOI 10.1007/978-3-031-72691-0_7