

Long-term Visual Localization with Mobile Sensors

Shen Yan^{1,3} Yu Liu¹ Long Wang² Zehong Shen³ Zhen Peng²
 Haomin Liu² Maojun Zhang¹ Guofeng Zhang³ Xiaowei Zhou^{3†}

¹National University of Defense Technology ²SenseTime Research ³Zhejiang University

Abstract

Despite the remarkable advances in image matching and pose estimation, image-based localization of a camera in a temporally-varying outdoor environment is still a challenging problem due to huge appearance disparity between query and reference images caused by illumination, seasonal and structural changes. In this work, we propose to leverage additional sensors on a mobile phone, mainly GPS, compass, and gravity sensor, to solve this challenging problem. We show that these mobile sensors provide decent initial poses and effective constraints to reduce the searching space in image matching and final pose estimation. With the initial pose, we are also able to devise a direct 2D-3D matching network to efficiently establish 2D-3D correspondences instead of tedious 2D-2D matching in existing systems. As no public dataset exists for the studied problem, we collect a new dataset that provides a variety of mobile sensor data and significant scene appearance variations, and develop a system to acquire ground-truth poses for query images. We benchmark our method as well as several state-of-the-art baselines and demonstrate the effectiveness of the proposed approach. Our code and dataset are available on the project page: <https://zju3dv.github.io/sensloc/>.

1. Introduction

Visual localization aims at estimating the camera translation and orientation for a given image relative to a known scene. Solving this problem is crucial for many applications such as autonomous driving [11], robot navigation [37] and augmented and virtual reality [7, 41].

State-of-the-art approaches to visual localization typically involve matching 3D points in a pre-built map and 2D pixels in a query image [4, 22, 42, 50–52, 67, 69]. An intermediate image retrieval step [2, 21, 23, 46] is often applied to determine which parts of the scene are likely visible

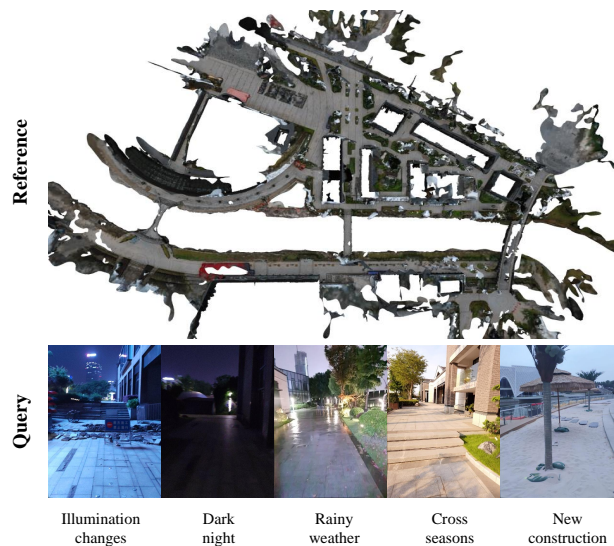


Figure 1. **Visual localization under extremely challenging conditions.** The proposed benchmark dataset *SensLoc* exhibits long-term appearance changes due to illumination, weather, season, day-night alternation, and new constructions.

in the query image, in order to handle large-scale scenes. The resulting camera poses are estimated using a standard Perspective-n-Point (PnP) solver [24, 32] inside a robust RANSAC [3, 12, 13, 18] loop. However, in real-world outdoor scenarios, obtaining such correspondences and further recovering the 6-DoF pose are difficult, since the outdoor scenes can experience large appearance variations caused by illumination (e.g., day and night), seasonal (e.g., summer and winter) and structure changes. Visual localization under such challenging conditions remains an unsolved problem, as reported by recent benchmarks [48, 53, 76].

Fortunately, nowadays, with the popularity of smart devices that come equipped with various sensors such as Inertial Measurement Unit (IMU), gravity, compass, GPS, or radio signals (like WiFi and Bluetooth), new possibilities arise for mobile phone pose estimation exploiting these additional multi-modality sensors. Nevertheless, previous works

The authors from Zhejiang University are affiliated with the State Key Lab of CAD&CG and ZJU-SenseTime Joint Lab of 3D Vision.
[†]Corresponding author: Xiaowei Zhou.

only take independent sensors into consideration. For example, some methods utilize GPS as a prior to bound the Visual-Inertial Odometry (VIO) drift [31, 45, 56, 75] or simplify the image retrieval process [35, 72, 73], while others focus on employing the gravity direction as a reliable prior to benefit the PnP solver [1, 20, 33, 52, 63–65].

In this paper, we introduce a novel framework, named SensLoc, that localizes an image by tightly coupling visual perception with complementary mobile sensor information for robust localization under extreme conditions. In the first stage, our approach leverages GPS and compass to constrain the search space for image retrieval, which not only reduces the risk of misrepresentation of global features but also speeds up the retrieval procedure with fewer database candidates. In the second stage, inspired by the recent CAD-free object pose estimation method, OnePose++ [26], we design a transformer-based network to directly match 3D points of the retrieved sub-map to dense 2D pixels of the query photo in a coarse-to-fine manner. Compared to the modern visual localization pipeline, which establishes 2D-3D correspondences by repeatedly matching 2D-2D local features between query and retrieve images, our solution shows a significant acceleration and a better performance especially under challenging appearance changes. In the last stage, we implement a simple yet effective gravity validation algorithm and integrate it into the RANSAC loop to filter the wrong pose hypotheses, leveraging the precise roll and pitch angles from mobile gravity sensors. The gravity validation leads to an improvement of RANSAC in terms of efficiency and accuracy, as false hypotheses can be removed in advance.

To the best of our knowledge, there is no public dataset for multi-sensor localization under strong visual changes. To facilitate the research of this area, we created a new benchmark dataset *SensLoc*, as shown in Fig. 1. Specifically, we first used a consumer-grade panoramic camera (Insta360) and a handheld Real-time Kinematic (RTK) recorder, to capture and reconstruct a large-scale reference map. Half a year later, we collected query sequences with large scene appearance changes through a mobile phone bounded with a RTK and recorded all available built-in sensor data. As direct registration between the query sequences and the reference map is difficult, we rebuilt an auxiliary map at the same time as acquiring the query sequences using Insta360 and RTK and aligned the auxiliary map with the reference map through ICP. Thus, we only needed to register the query images with the auxiliary map, which was easier as they were captured at the same time. To achieve this, we developed a pseudo-ground-truth (GT) generation algorithm to accurately register each query sequence against the auxiliary map by incorporating feature matching, visual-inertial-odometry, RTK positions and gravity directions. The GT generation algorithm does not ask for any manual intervention or extra setup in the environment, enabling scalable pose labeling.

We evaluate several state-of-the-art image retrieval and localization methods on our proposed dataset. We show that the performance of the existing methods can be drastically improved by considering sensor priors available in mobile devices, such as GPS, compass, and gravity direction. The experiments also demonstrate that our method outperforms the state-of-the-art approach HLoc [50, 51] by a large margin in challenging night-time environments, while taking only 66ms to find 2D-3D correspondences on a GPU and 8ms for PnP RANSAC.

In summary, our main contributions include:

- A novel outdoor visual localization framework with multi-sensor prior for robust and accurate localization under extreme visual changes.
- A new dataset for multi-sensor visual localization with seasonal and illumination variations.
- Benchmarking existing methods and demonstrating the effectiveness of the proposed approach.

2. Related work

Visual Localization. Conventionally, most existing methods [40, 50, 51, 60] tackle pose estimation by establishing correspondences between the query image and sparse SfM reconstruction of the scene, typically with hand-crafted local features like SIFT [40]. The pipeline scales up to large scenes using image retrieval [2, 21, 23, 46, 59, 74], which restricts 2D-3D matching into visible parts of the query image. Recently, many traditional hand-crafted steps have been substituted by learning-based counterparts [16, 17, 51, 60, 68]. HLoc [50, 51] summarizes and provides a complete and leading toolbox for structure-based visual localization. However, during the retrieval phase, HLoc generally fails when the scene exhibits large visual variations, due to the limited information contained in global features. During the 2D-3D matching phase, HLoc is slow because it depends on multiple 2D-2D image matchings as the proxy, and it may still output a lot of outliers over notable appearance variations (weather, seasonal and illumination). For the retrieval and matching problem, we believe that leveraging other modality sensors could bring benefits. As for the time-consumption issue, inspired by recent 6-DoF pose estimation works [26, 61], we aim to directly match 3D sub-map and 2D query image in one-shot with self- and cross-attention modules.

Multi-sensor Visual Localization. As GPS roughly provides absolute 3-DoF locations in outdoor environments, some methods employ the GPS signals as an extra constraint in optimization to improve the global accuracy of VIO [31, 45, 56, 75] and visual SLAM [10, 14, 57], while others treat GPS as a prior to simplify the image retrieval

dataset	scale	equipment cost	operation cost	changes	additional sensors	groundtruth solution	accuracy
Cambridge [30]	small	low	low	people,weather	No	SfM	$> dm$
Phototourism [28]	small	low	low	people,construction	No	SfM	$\approx m$
San Francisco [9]	large	low	low	people,construction	GPS	SfM+GPS	$\approx m$
Aachen [53]	large	low	high	people,day-night	No	SfM+manual	$> dm$
NCLT [6]	median	high	low	weather	LiDAR, IMU, GPS	VIO+LiDAR	$\approx dm$
ADVIO [15]	median	median	high	people	IMU, depth, GPS	VIO+manual	$\approx m$
ETH3D [55]	small	high	high	No	LiDAR	LiDAR+manual	$\approx mm$
LaMAR [48]	large	high	low	people,weather, day-night,construction	LiDAR, IMU, WiFi, Bluetooth, depth, infrared	LiDAR+ SL+VIO	$\approx cm$
SensLoc	large	low	low	people,car,weather, day-night,construction	GPS, IMU, WiFi, Bluetooth, compass, gravity	SL+VIO+ RTK+Gravity	$< dm$

Table 1. **Overview of existing outdoor datasets.** The following attributes are taken in account: environment (scale, changes), cost (equipment, operation), groundtruth (solution, accuracy), and whether contains sensor data. The equipment and operation costs refer to capital and labor overhead when building the dataset.

task for visual localization [35, 72, 73]. Gravity direction measured from IMU sensors is a commonly used prior for pose estimation considering its high accuracy [33]. Previous researches usually extract the knowledge of known gravity direction to enhance the PnP solver by bootstrapping minimal solutions [1, 33, 63, 65] or incorporating extra regularizers [20, 66]. However, to our knowledge, no previous work has considered multiple sensors jointly. In fact, modern cell phones and other smart devices have been equipped with a large variety of sensors, including gyroscopes and accelerometers, compass, GPS, Wifi, Bluetooth and so on. A localization algorithm to fully exploit multiple sensors is desired.

Datasets. The majority of existing datasets for visual localization do not provide [28, 30, 53, 76] or just provide limited types of sensor data [6, 9, 15, 46, 55, 58, 62, 67]. The closest to our work is the very recently released benchmark LaMAR [48] for augmented reality. However, LaMAR uses expensive and professional LiDAR devices to reconstruct the reference map and establish the ground-truth. Besides, LaMAR only provides radio signals from Wifi and Bluetooth, which are not always available in outdoor environments, and do not provide GPS and compass data.

To label the query GT, many previous works rely on offline SfM systems [54] upon unstructured photo collections [28, 30, 53]. However, when it comes to the circumstance where a huge appearance change exists between query and reference images, merely adopting SfM algorithms may lead to limited registration accuracy. Some other datasets propose to use fiducial markers [6, 8, 27], depth information (e.g., RGB-D camera [58] or LiDAR [6, 67]) as additional pose constraints, or even introduce manual annotations [15, 55, 67]. The requirement of external equipment and manual labeling greatly increase the capture cost and thus limit the dataset scalability. Recently, LSFb [39] and LaMAR [48] employ an automatic and low-cost framework for AR devices’ GT generation. Although we similarly fuse diverse sensor sources

to automatically generate GT, we devise a solution for hard cases where query images exhibit dramatic appearance differences compared with the reference map. Our approach builds an auxiliary 3D map to simplify feature matching and introduces more constraints such as RTK positions and gravity directions.

A detailed comparison of representative outdoor datasets is given in Table 1.

3. Visual Localization with Mobile Sensors

An overview of the proposed method is exhibited in Fig. 2. Given the reference map consisting of images with known camera poses $\{\mathbf{I}^r, \xi^r\}$ and point clouds $\{\mathbf{P}_j\}$, our objective is to estimate the 6-DoF pose ξ^q for the query image \mathbf{I}^q . To achieve this, we present a novel three-stage pipeline that first leverages GPS and compass to narrow the search space for image retrieval (Section 3.1). Then we build 2D-3D correspondences in a coarse-to-fine manner for query image \mathbf{I}^q (Section 3.2), and finally solve the camera pose ξ^q by a gravity-guided PnP RANSAC (Section 3.3).

3.1. Sensor-guided Image Retrieval

Given a query image \mathbf{I}^q , the retrieval stage aims to find a set of covisible images $S_{covis} = \{\mathbf{I}_1^r, \mathbf{I}_2^r, \dots, \mathbf{I}_k^r\}$ among the reference image collection $\{\mathbf{I}^r\}$, where k represents the number of retrieved items. A commonly-used method is to first use an embedding function $f(\cdot)$ to map images $\{\mathbf{I}^q, \mathbf{I}^r\}$ into a compact feature space and then search for nearest neighbors of \mathbf{I}^q using a distance metric $d(f(\mathbf{I}^q), f(\mathbf{I}^r))$.

However, this method is completely dependent on the expressiveness of global features. When images are acquired from a site with significant visual changes, e.g., the *SensLoc* dataset, it may fail. To address this issue, we propose to use the metadata from the mobile sensors as priors to filter incorrect candidates beforehand, as shown in Fig. 3. We denote the prior pose of the query image as ${}_s\xi^q$, which comprises translation from the built-in GPS signal and rotation from integrating the gravity and compass data. For each \mathbf{I}^q , we

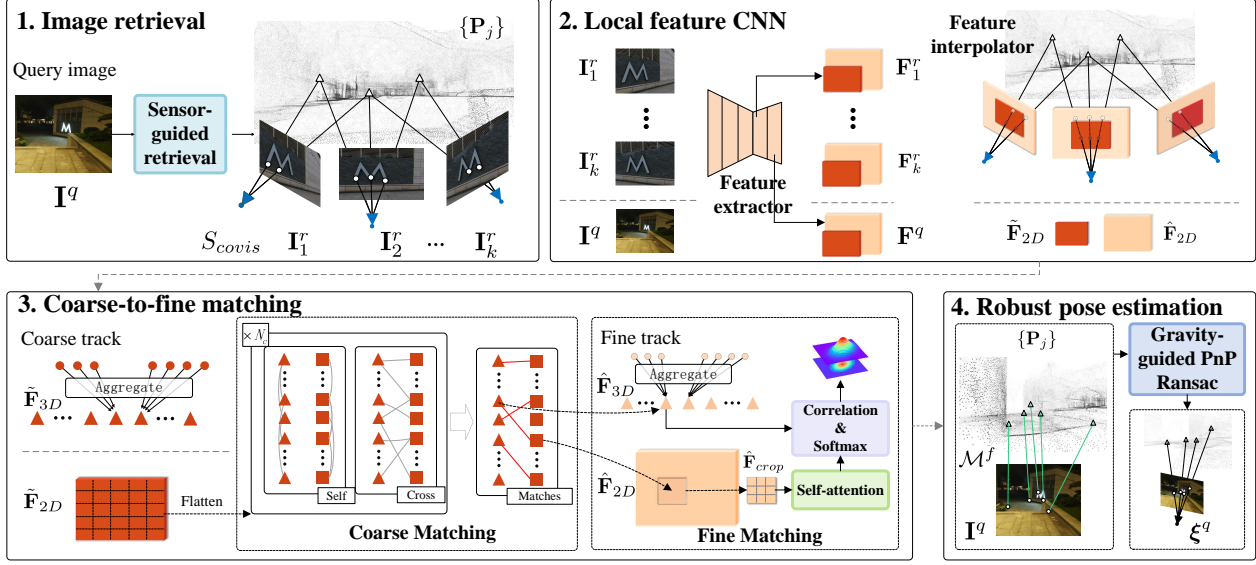


Figure 2. **Overview of the proposed method.** 1. For each query image \mathbf{I}^q , our sensor-guided retrieval algorithm first finds covisible references S_{covis} . 2. A local feature CNN extracts coarse-level and fine-level feature maps $\tilde{\mathbf{F}}_{2D}$ and $\hat{\mathbf{F}}_{2D}$. 3. 2D descriptors are aggregated to 3D descriptors for the feature tracks, followed by self- and cross-attention modules to build coarse correspondences and a fine matching module to refine these correspondences to sub-pixel positions. 4. The camera pose ξ^q is estimated via gravity-guided PnP RANSAC.

narrow the retrieval candidate set ${}^q\mathcal{I}$ as follows:

$${}^q\mathcal{I} = \{ \mathbf{I}_i^r \mid \forall \| \xi_i^r(t_{xy}) - {}_s\xi^q(t_{xy}) \| \leq \tau_t, \arccos(\xi_i^r(\mathbf{p}) \cdot {}_s\xi^q(\mathbf{p})) \leq \tau_o \}, \quad (1)$$

where τ_t and τ_o are translation and orientation thresholds, respectively. We notice that the altitude value of GPS is unstable, so we only use the x-y coordinate t_{xy} in reference pose ξ_i^r and query prior pose ${}_s\xi^q$ to constrain the search space. Besides, we compute the angle between principal axes of query pose prior (${}_s\xi^q(\mathbf{p})$) and reference images ($\xi_i^r(\mathbf{p})$), to ensure that camera directions are similar. The camera's principal axis is defined as a line perpendicular to the image plane that extends from the camera center. We use principal axes instead of calculating the camera orientations because images rotated around their principal axis still observe the same scene. Based on the filtering step, we can efficiently and accurately determine the k-nearest neighbors S_{covis} according to $d(\mathbf{f}(\mathbf{I}^q), \mathbf{f}(\mathbf{I}^p))$, where $\mathbf{I}^p \in {}^q\mathcal{I}$. The local point cloud $\{\mathbf{P}_j\}$ observed by images in S_{covis} are used for the latter feature matching step.

3.2. Direct 2D-3D Matching

For a query \mathbf{I}^q , assuming that we have obtained its covisible neighbors S_{covis} and relevant point cloud $\{\mathbf{P}_j\}$, we focus on establishing 2D-3D matches between sub-pixels in \mathbf{I}^q and points in $\{\mathbf{P}_j\}$.

Preprocess. First, we hierarchically extract coarse- and fine-level dense feature maps $\{\mathbf{F}^q, \{\mathbf{F}_1^r, \mathbf{F}_2^r, \dots, \mathbf{F}_k^r\}\}$ from

images $\{\mathbf{I}^q, \{\mathbf{I}_1^r, \mathbf{I}_2^r, \dots, \mathbf{I}_k^r\}\}$. For clarity, we use $\tilde{\cdot}$ and $\hat{\cdot}$ to denote the coarse-level and fine-level features, so that extracted multi-level features can be denoted as $\tilde{\mathbf{F}}_{2D} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_c}$ for the coarse level and $\hat{\mathbf{F}}_{2D} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C_f}$ for the fine level, where H, W indicate image height and width, and C_c, C_f represent the dimensions of the coarse and fine features, respectively. Next, every point \mathbf{P}_j is reprojected into the feature maps in S_{covis} and the corresponding feature descriptors are extracted using bilinear interpolation. This step yields a feature track $\mathcal{T}_j = \{(\tilde{\mathbf{F}}_{2D}(k), \hat{\mathbf{F}}_{2D}(k)) \mid k = 1 \dots N_{j_k}\}$ for each 3D point \mathbf{P}_j , where N_{j_k} is the track length. Finally, an average pooling operation is used for each track \mathcal{T}_j to form a 3D descriptor $(\tilde{\mathbf{F}}_{3D}(j), \hat{\mathbf{F}}_{3D}(j))$.

Coarse-to-fine matching. Based on the 2D and 3D features $\{\tilde{\mathbf{F}}_{2D}, \hat{\mathbf{F}}_{2D}\}$ and $\{\tilde{\mathbf{F}}_{3D}, \hat{\mathbf{F}}_{3D}\}$, we initially determine whether a 3D point in sub-map $\{\mathbf{P}_j\}$ is observable by the query \mathbf{I}^q and search for a rough location $\hat{\mathbf{u}}^i$ in the coarse level. Afterward, the coarse correspondence is refined to a sub-pixel position $\hat{\mathbf{u}}^i$ by a fine-matching module.

Similar to [26, 60], we first use positional encoding to augment the features $\{\tilde{\mathbf{F}}_{2D}, \hat{\mathbf{F}}_{3D}\}$, resulting in $\{\tilde{\mathbf{F}}_{2D}^{pe}, \hat{\mathbf{F}}_{3D}^{pe}\}$. This process involves 2D pixel extension following [5] and 3D coordinate encoding with Multilayer Perceptron (MLP) [43]. We flatten the 2D coarse map $\tilde{\mathbf{F}}_{2D}^{pe}$ and apply linear self-attentions and cross-attentions [29] upon $\{\tilde{\mathbf{F}}_{2D}^{pe}, \hat{\mathbf{F}}_{3D}^{pe}\}$ N_c times to obtain transformed features $\{\tilde{\mathbf{F}}_{2D}^t, \hat{\mathbf{F}}_{3D}^t\}$. Subsequently, a probability matrix \mathcal{P}^c is calculated by a dual-softmax operation on the correlation volume

C, which is derived from the inner product of 2D and 3D features $\{\tilde{\mathbf{F}}_{2D}^t, \tilde{\mathbf{F}}_{3D}^t\}$:

$$\mathcal{P}^c(i, j) = \text{softmax}(C(i, \cdot))_j \cdot \text{softmax}(C(\cdot, j))_i, \quad (2)$$

where $C(i, j) = \frac{1}{\tau} \cdot \langle \tilde{\mathbf{F}}_{2D}^t(i), \tilde{\mathbf{F}}_{3D}^t(j) \rangle.$

$\langle \cdot, \cdot \rangle$ indicates the inner product, τ represents the temperature parameter of softmax, and i, j are the indices of a pixel in the flattened coarse map and a 3D point, respectively. The coarse 2D-3D correspondences, denoted as \mathcal{M}^c , are constructed from \mathcal{P}^c by selecting correspondences that meet a confidence threshold θ and satisfy the mutual nearest neighbor (MNN) criteria:

$$\mathcal{M}^c = \{(i, j) \mid \forall (i, j) \in \text{MNN}(\mathcal{P}^c), \mathcal{P}^c(j, i) \geq \theta\}, \quad (3)$$

For each coarse correspondence $(\tilde{\mathbf{u}}^i, \mathbf{P}^j)$ identified by \mathcal{M}^c , a local window $\hat{\mathbf{F}}_{crop}$ of size $w \times w$ is cropped around $\tilde{\mathbf{u}}^i$ in the fine feature $\hat{\mathbf{F}}_{2D}$. A similar self-attention and cross-attention layer is then applied to transform the cropped feature map $\hat{\mathbf{F}}_{crop}$ and its corresponding 3D fine feature $\hat{\mathbf{F}}_{3D}(j)$ N_f times. Finally, we correlate the transformed 3D fine feature $\hat{\mathbf{F}}_{3D}^t(j)$ with all elements in the transformed crop feature $\hat{\mathbf{F}}_{crop}^t$ to produce a heatmap representing the matching probability in the local region. By estimating the 2D expectation over the probability distribution, we gain the fine correspondence position $\hat{\mathbf{u}}^i$ with respect to \mathbf{P}^j .

Following [26, 60], we supervise the network by a focal loss [38] for coarse matching and variance-weighted ℓ_2 loss for fine refinement. More details are provided in the supplementary material.

3.3. Gravity-guided PnP RANSAC

After establishing the 2D-3D fine matches \mathcal{M}^f between the query image \mathbf{I}^q and the local point cloud \mathbf{P}^j , previous works commonly estimate the pose ξ^q with the Perspective-n-Point (PnP) [34] technique in a RANSAC [18] framework.

However, such methods may have trouble with scenes containing large illumination changes as they highly rely on the performance of local feature matching. To this end, we propose to insert a simple yet effective verification module into a LO-RANSAC [13] scheme to ensure the correctness of the gravity direction. In detail, for each pose hypothesis ξ_{hyp} estimated from a random sample of three 2D-3D correspondences, we compute the difference of gravity directions d_ϵ between sensor pose $s\xi$ and hypothesis pose ξ_{hyp} , as

$$d_\epsilon = \arccos(s\xi(\mathbf{g}) \cdot \xi_{hyp}(\mathbf{g})). \quad (4)$$

The notations $s\xi(\mathbf{g})$ and $\xi_{hyp}(\mathbf{g})$ refer to the gravity directions of sensor and hypothesis poses. During RANSAC iterations, an earlier termination is permitted if $d_\epsilon \geq \tau_\epsilon$, where τ_ϵ is a predefined maximum gravity error.

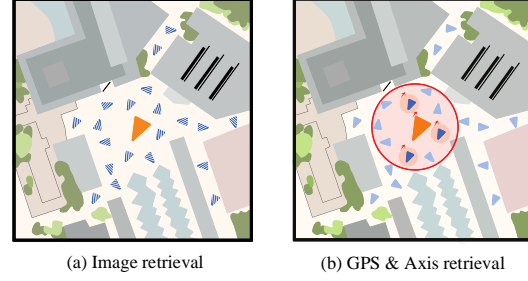


Figure 3. **Using the GPS and compass for image retrieval.** (a) Query views (orange triangle) against reference views (blue triangles). (b) Reducing search space by considering the GPS distance and axis direction.

According to this strict criterion, we not only suppress potential false hypotheses and return high-quality results, but also could handle overload situations when there are too many tentative 2D-3D correspondences. This is achieved by filtering out a large fraction of incorrect poses with $d_\epsilon \geq \tau_\epsilon$ in advance.

3.4. Implementation details

ResNet-18 [25] is used as the feature extractor and we set N_c to 4 and N_f to 1 for the transformer modules. The temperature value τ is 0.1. The crop window size w is equal to 5. As for training, we select the same sequences on MegaDepth [36] with [17], but retriangulate the 3D structure with Superpoint [16] and Superglue [51] to increase the point density. We randomly sample or pad 10 covisible images S_{covis} and extract a visible local sub-map $\{\mathbf{P}_j\}$ with 4000 points for batch fetching. The backbone of our model is fixed with the LoFTR-DS [60] outdoor model, while the other parts are randomly initialized and backpropagated. The model is trained using AdamW with an initial learning rate of 8×10^{-3} and a batch size of 24 on 4 NVIDIA GTX 3090 GPUs. During testing, we set the confidence threshold θ to 0.05 in the daytime and 0.005 in the night-time, as SensLoc has the ability to handle tentative matches brought by such a low threshold. The translation and orientation thresholds τ_t and τ_o in Eq. 1 are 20 meters and 60 degrees for sensor-guided retrieval. The gravity error threshold τ_ϵ is set to 2 degrees for gravity-guided PnP RANSAC. All testing experiments are operated on an NVIDIA TITAN RTX GPU.

4. Dataset

The released dataset *SensLoc* includes a large city park (approximately $31,250m^2$), containing vegetation, rivers, buildings and walkways. As a publicly accessible area, unavoidably, *SensLoc* experiences various scene changes all the time, e.g., moving people, vehicles or even new infrastructure constructions. The query images were captured with different illuminations, seasons and weather, compared to

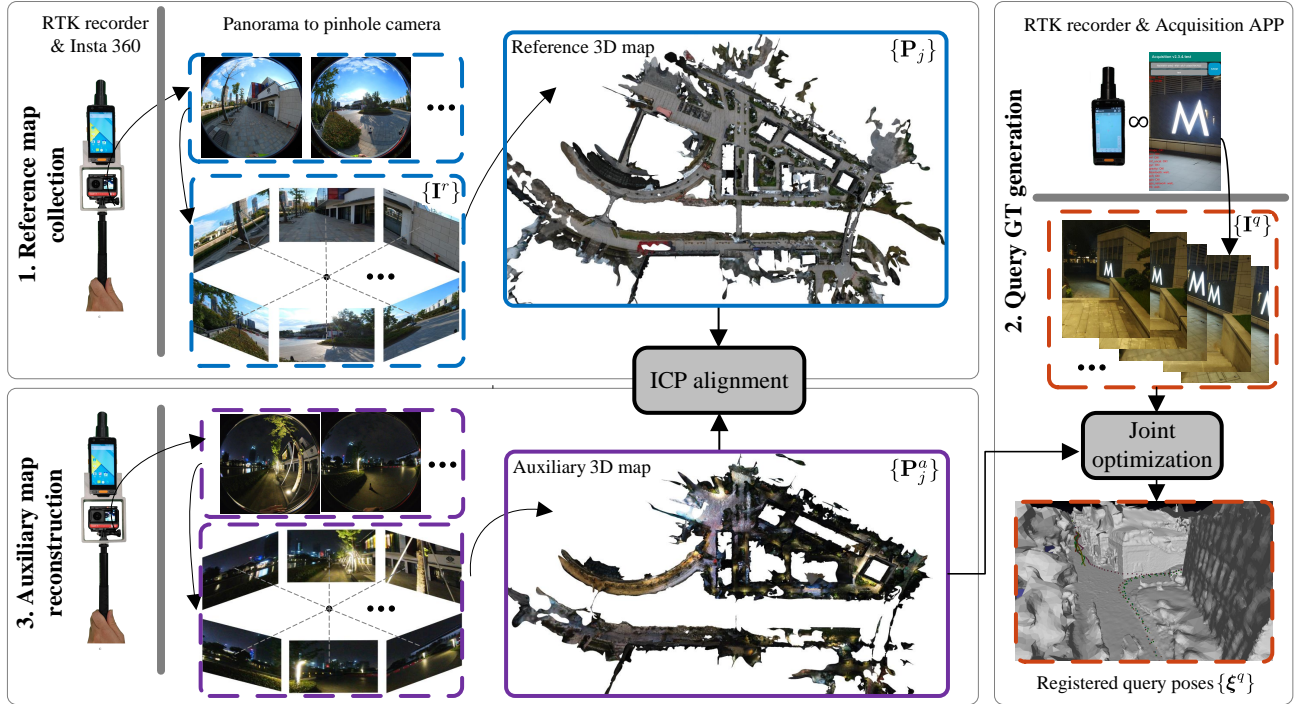


Figure 4. **Overview of dataset collection.** The dataset construction can be divided into three parts: 1. Building a large-scale reference map with Insta 360 and RTK. 2. Collecting query images by a mobile phone bounded with an RTK recorder. 3. At the same time with query image collection, building an auxiliary map with Insta 360 and RTK to facilitate query-reference matching. The GT poses of query images are generated by first aligning the auxiliary map to the reference map by ICP and then registering query images to the auxiliary map by joint optimization based on multi-sensor information.

the reference images. A visualization of the dataset can be found in Fig. 1. The supplementary material provides statistics on the number of query images for each condition.

4.1. Reference map collection

We created a digital map upon *SensLoc* environment with a panoramic camera, Insta360 ONE, due to the efficiency of 3D reconstruction from panoramic imaging over monocular acquisition. Specifically, given reference images $\{I^r\}$ captured by Insta360, we applied SfM techniques to reconstruct sparse point cloud $\{P_j \in \mathbb{R}^3\}$ and feature track $\{T_j\}$, where j is the point index. After converting 7,958 database images from panorama to pinhole, *SensLoc* dataset results in 47,748 reference images and 0.31M 3D points triangulated from 6.20M local features. In order to recover the scale of the reconstruction and align it with the real geographic world, we pre-calibrated and equipped the Insta360 ONE with a RTK recorder, T38P-E, to document the geographical locations. The pipeline of the reference map reconstruction is illuminated in the top area of Fig. 4.

4.2. Query image collection

For query image collection, we freely walked through the *SensLoc* environment several times after half a year, and captured videos using Huawei P40pro and Xiaomi Mix3 smartphones with a developed Android-powered Acquisition Application (APP). Since the *SensLoc* algorithm focuses on image localization, we sampled all sequences into single frames and totally generated 1,576 day-time and 362 night-time queries. Note that without downsampling, our dataset can also support sequence localization tasks. The Acquisition APP records raw data that measure motion, orientation, and various environmental conditions from the phones' built-in sensors. It includes but not limited to accelerometer, gyroscope, gravity, compass, Bluetooth, WiFi and GPS measurements. Please refer to the supplementary material for a detailed introduction. Besides, we employed a handheld RTK device attached to the phone rigidly to record shooting locations, which could provide GT locations at centimeter-level accuracy [70]. All sensors have been hardware synchronized and carefully calibrated.

Auxiliary map reconstruction. We found that it was quite difficult to directly generate pseudo-GT poses against the pre-

built reference map $\{\mathbf{P}_j\}$ due to large appearance changes between query and reference images. To solve this problem, we propose to reconstruct an auxiliary 3D map $\{\mathbf{P}_j^a\}$ during the query time. This process is detailed in the median part of Fig. 4. Specifically, we first replicate the reference map construction process by utilizing the Insta360 and RTK recorder to capture and build a 3D auxiliary model. The auxiliary model $\{\mathbf{P}_j^a\}$ is then aligned with the base reference model $\{\mathbf{P}_j\}$. In detail, both P_j and P_j^a are initially aligned with the real geographic world by RTK with an error less than 0.5m. Fine alignment is achieved using the ICP [47], which is not likely to get stuck in local minima due to the good initialization. The RMSE of all inlier 3D-3D matches is around 1.5cm and the alignment quality can be visually inspected in supplementary materials. Based on the alignment of two maps, pseudo-GT generation in $\{\mathbf{P}_j\}$ is equal to recovering accurate poses in $\{\mathbf{P}_j^a\}$, where registering query images to a 3D map collected at the same time is much easier.

Query GT generation. We apply a joint optimization algorithm to find the pseudo-GT pose parameters $\{\xi_i^q\}$ that meet the requirements: 1) the camera self-localization, 2) motion constraints from IMU measurements, and 3) sensor prior constraints regarding translation and orientation. The self-localization recovers camera poses $\{\xi_i^q\}$ by finding 2D-3D correspondences between point cloud $\{\mathbf{P}_j^a\}$ and query image $\{\mathbf{I}^q\}$. To impose temporal constraints, VIO is employed as another constraint to refine the localization $\{\xi_i^q\}$. To further increase the pose accuracy, we use the measured RTK positions and gravity directions as regularization. The entire GT generation is formulated as the following joint optimization [71] function:

$$\begin{aligned} \xi_i^{q*} = \operatorname{argmin}_{\xi_i^q, \mathbf{X}_j} & \\ w_{st} \sum_i \sum_j & \|\mathbf{p}_{ij}^q - \pi(\xi_i^q, \mathbf{P}_j)\|_{\sigma_{st}}^2 + \\ w_{vo} \sum_i \sum_j & \|\mathbf{x}_{ij}^q - \pi(\xi_i^q, \mathbf{X}_j)\|_{\sigma_{vo}}^2 + \\ w_{io} \sum_i & \|\mathbf{h}(\xi_i^q, \xi_{i+1}^q, \mathbf{v}_i, \mathbf{b}_a, \mathbf{b}_g)\|_{\sigma_{io}}^2 + \\ w_t \sum_i & \|\mathbf{t}_i^q - \mathbf{t}_i^q\|_{\sigma_t}^2 + w_g \sum_i (\arccos(\hat{\mathbf{g}}_i^q \cdot \mathbf{g}_i^q))_{\sigma_g}^2, \end{aligned} \quad (5)$$

where $\pi(\cdot)$ represents a pinhole projection function. The sets $\{(\mathbf{P}_j, \mathbf{p}_{ij}^q)\}$ and $\{(\mathbf{X}_j, \mathbf{x}_{ij}^q)\}$ indicate 2D-3D matches in self-localization and visual odometry respectively. The energy function $\mathbf{h}(\cdot)$ [19] is used to measure the difference between consecutive poses $\{\xi_i^q, \xi_{i+1}^q\}$ and the integration of IMU measurements $\{\mathbf{b}_a, \mathbf{b}_g\}$ and velocity \mathbf{v}_i . We employ the x-y coordinate value $\{\mathbf{t}_i^q\}$ from RTK to direct the movement of camera trajectory $\{\mathbf{t}_i^q\}$, where $\mathbf{t}_i^q = \xi_i^q(\mathbf{t}_{xy})$. Furthermore, we impose a misalignment penalty on query

Global feature	Prior	R@1	R@5	P@5	R@10	P@10	R@20	P@20
-	Axis+GPS	34.73	77.29	34.20	82.04	34.38	87.20	33.91
AP-GeM [23]	-	51.87	67.75	44.46	74.45	39.60	79.71	32.58
	Axis	53.10	71.67	47.63	79.31	43.16	85.35	36.80
	GPS	61.56	82.40	54.91	89.27	50.03	93.81	43.12
	Axis+GPS	64.96	88.13	60.27	93.19	55.92	96.70	49.59
NetVLAD [2]	-	59.80	73.87	53.87	77.80	48.10	82.20	38.97
	Axis	63.57	80.90	58.26	85.81	52.77	89.83	43.75
	GPS	70.85	90.30	66.21	94.69	61.08	97.06	52.39
	Axis+GPS	74.51	93.29	70.33	96.59	65.44	98.91	57.27
OpenIBL [21]	-	68.42	82.49	63.44	86.03	57.58	89.19	47.55
	Axis	68.57	83.75	63.81	87.82	58.02	91.07	48.92
	GPS	75.12	91.59	70.99	95.77	66.16	97.88	57.09
	Axis+GPS	77.50	92.93	73.49	96.65	69.22	98.56	60.73

Table 2. **Image retrieval results.** We report the top- k recall and precision when $k = 1, 5, 10, 20$ using different image features and sensor priors. A retrieval result is deemed correct if it falls within a distance of $\tau_t = 10m$ from the query view and the orientation between them is less than $\tau_o = 30^\circ$.

gravity direction $\{\mathbf{g}_i^q\}$ based on reliable gravity sensor output $\{\hat{\mathbf{g}}_i^q\}$, where $\mathbf{g}_i^q = \xi_i^q(\mathbf{g})$. The term $\|\mathbf{e}\|_\sigma^2 = \mathbf{e}^T \mathbf{e} / \sigma^2$ represents the squared norm of the error vector. Further details can be found in the supplementary material.

The major errors in our generated GT may include: the error of the RTK recorder (1cm to 3cm), calibration error between the RTK recorder and mobile phone (approximately 2cm), and misalignment between the reference and auxiliary maps (approximately 1.5cm), which make our pseudo-GT accuracy lie in the decimeter level.

5. Experiments

5.1. Image Retrieval

Baselines and metrics. We evaluate global descriptors computed by AP-GeM [23], NetVLAD [2] and OpenIBL [21], which are representative of the field [44]. In order to learn the effect of the provided priors (i.e., axis direction and GPS xy position), we carry out controlled experiments on the basis of distinct global features. For evaluation, a retrieval result is deemed correct if it satisfies two requirements: 1) it is within $\tau_t = 10$ meters from the query position, and 2) it shares a similar perspective with the query pose, where $\tau_o < 30^\circ$. We report recall and precision metrics on the behalf of the top k retrieved items, and k ranges from 1 to 20.

Results. The retrieval results are presented in Table 2. Even though sensor prior itself cannot output accurate results, it always improves the retrieval performance over vanilla feature-only methods, regardless of the top k threshold. Combining the global feature OpenIBL with Axis (orientation) and GPS (position) priors leads the benchmark, which is adopted to find the covisible reference set for the

	Day			Night			Time (Net)	Time (PnP)
	(25cm, 2°)	(50cm, 5°)	(1m, 10°)	(25cm, 2°)	(50cm, 5°)	(1m, 10°)		
HLoc (<i>SIFT</i> + <i>NN</i>)	62.56	74.56	80.52	22.65	29.28	35.08	706	18
HLoc (<i>SPP</i> + <i>NN</i>)	73.73	83.19	86.99	34.81	39.23	41.71	1361	18
HLoc (<i>D2Net</i> + <i>NN</i>)	16.31	25.06	32.74	1.10	2.76	3.87	1987	10
HLoc (<i>SPP</i> + <i>SPG</i>)	86.04	93.34	95.18	70.99	77.90	80.94	2419	92
HLoc (<i>LoFTR</i> *)	78.81	89.53	92.70	72.65	87.85	91.16	2880	170
Pixloc (<i>Retrieval initial</i>)	31.85	37.75	43.78	20.99	28.18	35.08	1237 (<i>Direct alignment</i>)	
SensLoc	85.09	93.02	94.99	80.94	92.27	93.65	66	33
SensLoc + Gravity	85.28	93.02	95.05	81.49	92.27	94.48	66	8

Table 3. **Visual localization results.** Our method is compared with HLoc [50] with different feature matchers and Pixloc [49]. The “Net” and “PnP” times refer to the running time in *ms* of the feature matching and PnP RANSAC processes, respectively. Note that all the methods use the same multi-sensor retrieval method, i.e., OpenIBL+Axis+GPS in Table 2.

following localization methods.

5.2. Visual Localization

Baselines and metrics. We compare our approach with the following baselines in two categories: 1) Feature matching pipelines HLoc [50], using different keypoint descriptors (SIFT [40], SuperPoint [16] and D2-Net [17]), and matchers (Nearest Neighbour and learned SuperGlue [51]), as well as the detector-free matcher LoFTR [60]. 2) The end-to-end direct alignment approach Pixloc [49]. Note that LoFTR outputs semi-dense correspondence between reference images, which brings unacceptable memory overhead to BA optimization in our large-scale dataset *SensLoc*. We thus implement an interval sampling upon raw tentative matches for sparse point cloud reconstruction (denoted as HLoc (*LoFTR**)). We follow the standard localization evaluation procedure [69], and report the localization recall at thresholds (25cm, 2°), (50cm, 5°), and (1m, 10°). The testing time-cost is also compared in terms of network inference and PnP RANSAC, as it is an important indicator for the navigation performance of mobile devices in the real world.

Results. We report the results in Table 3. The state-of-the-art method HLoc (SuperPoint+SuperGlue) obtains the best results during the daytime. It is reasonable because well-lit images have rich textures, benefiting the keypoint detection of SuperPoint. In the challenging condition of night, our approach outperforms other baselines by a large margin thanks to the keypoint-free design of 2D-3D matching. In addition, as we execute sparse-to-dense correspondence searching in one-shot, the network runs about 30x faster than the second competitor SuperPoint+SuperGlue. Furthermore, it is obvious that our proposed gravity validation module not only improves the localization accuracy, but also speeds up the PnP RANSAC process by 4 times. As we can see, even in severe long-term conditions, according to tightly coupling visual localization with multiple mobile sensors, we could obtain a satisfactory pose in real-time on a desktop computer.

Retrieval	Day			Night		
	(25cm, 2°)	(50cm, 5°)	(1m, 10°)	(25cm, 2°)	(50cm, 5°)	(1m, 10°)
w/o sensor-guided	75.38 / 81.54 / 82.87	57.46 / 62.15 / 64.64				
w sensor-guided	85.28 / 93.02 / 95.05	81.49 / 92.27 / 94.48				
ground-truth	87.44 / 95.81 / 97.27	87.02 / 93.92 / 96.41				

Table 4. **Ablation study.** We quantitatively evaluate the influence of image retrieval on SensLoc.

5.3. Ablation Studies

We conduct experiments to exploit the sensitivity of SensLoc to image retrieval, as shown in Tab. 4. Increasing the retrieval accuracy using sensor-guided retrieval or ground-truth markedly increases the localization recall at (25cm, 2°)/(50cm, 5°)/(1m, 10°). The ablation study demonstrates that the retrieval stage, which selects local point clouds in Section 3.1, plays a critical role in the pose estimation of our method. More ablation studies are provided in the supplementary material.

6. Conclusion

This paper presents SensLoc, a new approach for long-term visual localization based on multi-modality sensors on mobile phones. To overcome the difficulties of image matching in temporally-varying outdoor environments, we propose to leverage additional mobile sensors, mainly GPS, compass and gravity sensors, to assist both image retrieval and pose estimation. We also introduce a new outdoor dataset that provides a variety of mobile sensor data and strong appearance changes between query and reference images. Our method significantly outperforms the state-of-the-art localization methods in terms of both accuracy and time-cost.

Acknowledgements. The authors would like to acknowledge the support from the National Key Research and Development Program of China (No. 2020AAA0108901) and NSFC (No. 62171451, No. 62101576, No. 62071478).

References

- [1] Cenek Albl, Zuzana Kukelova, and Tomas Pajdla. Rolling shutter absolute pose problem with known vertical direction. In *CVPR*, 2016. 2, 3
- [2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 2016. 1, 2, 7
- [3] Daniel Barath, Jiri Matas, and Jana Noskova. MAGSAC: marginalizing sample consensus. In *CVPR*, 2019. 1
- [4] Eric Brachmann and Carsten Rother. Visual camera re-localization from RGB and RGB-D images using DSAC. *T-PAMI*, 2021. 1
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 4
- [6] Nicholas Carlevaris-Bianco, Arash K. Ushani, and Ryan M. Eustice. University of michigan north campus long-term vision and lidar dataset. *IJRR*, 2016. 3
- [7] Robert Castle, Georg Klein, and David W. Murray. Video-rate localization in multiple maps for wearable augmented reality. In *ISWC*, 2008. 1
- [8] Simone Ceriani, Giulio Fontana, Alessandro Giusti, Daniele Marzorati, Matteo Matteucci, Davide Migliore, Davide Rizzi, Domenico G Sorrenti, and Pierluigi Taddei. Rawseeds ground truth collection systems for indoor self-localization and mapping. *Auton. Robots*, 2009. 3
- [9] David M. Chen, Georges Baatz, Kevin Köser, Sam S. Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, Bernd Girod, and Radek Grzeszczuk. City-scale landmark identification on mobile devices. In *CVPR*, 2011. 3
- [10] Xiao Chen, Weidong Hu, Lefeng Zhang, Zhiguang Shi, and Maisi Li. Integration of low-cost GNSS and monocular cameras for simultaneous localization and mapping. *Sensors*, 2018. 2
- [11] Wentao Cheng, Weisi Lin, Kan Chen, and Xinfeng Zhang. Cascaded parallel filtering for memory-efficient image-based localization. In *ICCV*, 2019. 1
- [12] Ondřej Chum and Jiří Matas. Optimal randomized RANSAC. *T-PAMI*, 2008. 1
- [13] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized RANSAC. In *DAGM*, 2003. 1, 5
- [14] Benjamin Congram and Timothy D. Barfoot. Relatively lazy: Indoor-outdoor navigation using vision and gnss. In *CRV*, 2021. 2
- [15] Santiago Cortés, Arno Solin, Esa Rahtu, and Juho Kannala. ADVIO: an authentic dataset for visual-inertial odometry. In *ECCV*, 2018. 3
- [16] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, 2018. 2, 5, 8
- [17] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable CNN for joint description and detection of local features. In *CVPR*, 2019. 2, 5, 8
- [18] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 1981. 1, 5
- [19] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. In *Robotics: Science and Systems*, 2015. 7
- [20] Victor Fragoso, Joseph DeGol, and Gang Hua. gdl*: Generalized pose-and-scale estimation given scale and gravity priors. In *CVPR*, 2020. 2, 3
- [21] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *ECCV*, 2020. 1, 2, 7
- [22] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. S2DNet: Learning accurate correspondences for sparse-to-dense feature matching. In *ECCV*, 2020. 1
- [23] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *IJCV*, 2017. 1, 2, 7
- [24] Bert M. Haralick, Chung-Nan Lee, Karsten Ottenberg, and Michael Nölle. Review and analysis of solutions of the three point perspective pose estimation problem. *IJCV*, 1994. 1
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5
- [26] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. In *NeurIPS*, 2022. 2, 4, 5
- [27] Jinyong Jeong, Younggun Cho, Young-Sik Shin, Hyunchul Roh, and Ayoung Kim. Complex urban dataset with multi-level sensors from highly diverse urban environments. *IJRR*, 2019. 3
- [28] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *IJCV*, 2021. 3
- [29] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, 2020. 4
- [30] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 3
- [31] Hojun Kim, Kyoungah Choi, and Impyeong Lee. High accurate affordable car navigation using built-in sensory data and images acquired from a front view camera. In *IV*, 2014. 2
- [32] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *CVPR*, 2011. 1
- [33] Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Closed-form solutions to minimal absolute pose problems with known vertical direction. In *ACCV*, 2010. 2, 3
- [34] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Eppn: An accurate $O(n)$ solution to the pnp problem. *IJCV*, 2008. 5

- [35] Yicheng Li, Zhaozheng Hu, Yuezhi Hu, and Duanfeng Chu. Integration of vision and topological self-localization for intelligent vehicles. *Mechatronics*, 2018. 2, 3
- [36] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 5
- [37] Hyon Lim, Sudipta N. Sinha, Michael F. Cohen, and Matthew Uyttendaele. Real-time image-based 6-dof localization in large-scale environments. In *CVPR*, 2012. 1
- [38] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5
- [39] Haomin Liu, Mingxuan Jiang, Zhuang Zhang, Xiaopeng Huang, Linsheng Zhao, Meng Hang, Youji Feng, Hujun Bao, and Guofeng Zhang. LSFb: A low-cost and scalable framework for building large-scale localization benchmark. In *ISMAR*, 2020. 3
- [40] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2, 8
- [41] Simon Lynen, Torsten Sattler, Michael Bosse, Joel A. Hesch, Marc Pollefeys, and Roland Siegwart. Get out of my lab: Large-scale, real-time visual-inertial localization. In *Robotics: Science and Systems*, 2015. 1
- [42] Simon Lynen, Bernhard Zeisl, Dror Aiger, Michael Bosse, Joel Hesch, Marc Pollefeys, Roland Siegwart, and Torsten Sattler. Large-scale, real-time visual-inertial localization revisited. *IJRR*, 2020. 1
- [43] Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta numerica*, 1999. 4
- [44] Noé Pion, Martin Humenberger, Gabriela Csurka, Johann Cabon, and Torsten Sattler. Benchmarking image retrieval for visual localization. In *3DV*, 2020. 7
- [45] Tong Qin, Shaozu Cao, Jie Pan, and Shaojie Shen. A general optimization-based framework for global pose estimation with multiple sensors. *arXiv preprint arXiv:1901.03642*, 2019. 2
- [46] Jerome Revaud, Jon Almazán, Rafael S. Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *ICCV*, 2019. 1, 2, 3
- [47] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the ICP algorithm. In *3DIM*, 2001. 7
- [48] Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L. Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. Lamar: Benchmarking localization and mapping for augmented reality. In *ECCV*, 2022. 1, 3
- [49] Paul-Edouard Sarlin, Ajaykumar Unagar, Måns Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the feature: Learning robust camera localization from pixels to pose. In *CVPR*, 2021. 8
- [50] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 1, 2, 8
- [51] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1, 2, 5, 8
- [52] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *T-PAMI*, 2016. 1, 2
- [53] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomás Pajdla. Benchmarking 6dof outdoor visual localization in changing conditions. In *CVPR*, 2018. 1, 3
- [54] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3
- [55] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 3
- [56] Markus Schreiber, Hendrik Königshof, André-Marcel Hellmund, and Christoph Stiller. Vehicle localization with tightly coupled GNSS and visual odometry. In *IV*, 2016. 2
- [57] Yun Shi, Shunping Ji, Zhongchao Shi, Yulin Duan, and Ryosuke Shibasaki. Gps-supported visual SLAM with a rigorous sensor model for a panoramic camera in outdoor environments. *Sensors*, 2012. 2
- [58] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew W. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*, 2013. 3
- [59] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 2
- [60] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 2, 4, 5, 8
- [61] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without CAD models. In *CVPR*, 2022. 2
- [62] Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A dataset for benchmarking image-based localization. In *CVPR*, 2017. 3
- [63] Linus Svärm, Olof Enqvist, Fredrik Kahl, and Magnus Oskarsson. City-scale localization for cameras with known vertical direction. *T-PAMI*, 2016. 2, 3
- [64] Linus Svärm, Olof Enqvist, Magnus Oskarsson, and Fredrik Kahl. Accurate localization and pose estimation for large 3d models. In *CVPR*, 2014. 2
- [65] Chris Sweeney, John Flynn, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. Efficient computation of absolute pose for gravity-aware augmented reality. In *ISMAR*, 2015. 2, 3
- [66] Chris Sweeney, Victor Fragoso, Tobias Höllerer, and Matthew Turk. gdl: A scalable solution to the generalized pose and scale problem. In *ECCV*, 2014. 3
- [67] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, 2018. 1, 3

- [68] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *CVPR*, 2019. 2
- [69] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, et al. Long-term visual localization revisited. *T-PAMI*, 2020. 1, 8
- [70] Miguel Tradacete, Álvaro Sáez, Juan Felipe Arango, Carlos Gómez Huélamo, Pedro Revenga, Rafael Barea, Elena López-Guillén, and Luis Miguel Bergasa. Positioning system for an electric autonomous vehicle based on the fusion of multi-gnss RTK and odometry by using an extended kalman filter. In *WAF*, 2018. 6
- [71] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment - A modern synthesis. In *ICCV*, 1999. 7
- [72] Kumar Vishal, C. V. Jawahar, and Vishes Chari. Accurate localization by fusing images and gps signals. In *CVPRW*, 2015. 2, 3
- [73] Olga Vysotska, Tayyab Naseer, Luciano Spinello, Wolfram Burgard, and Cyrill Stachniss. Efficient and effective matching of image sequences under substantial appearance changes exploiting GPS priors. In *ICRA*, 2015. 2, 3
- [74] Shen Yan, Maojun Zhang, Shiming Lai, Yu Liu, and Yang Peng. Image retrieval for structure-from-motion via graph convolutional network. *Inf. Sci.*, 2021. 2
- [75] Yang Yu, Wenliang Gao, Chengju Liu, Shaojie Shen, and Ming Liu. A gps-aided omnidirectional visual-inertial state estimator in ubiquitous environments. In *IROS*, 2019. 2
- [76] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference pose generation for long-term visual localization via learned features and view synthesis. *IJCV*, 2021. 1, 3