Benchmarking Egocentric Visual-Inertial SLAM at City Scale

Anusha Krishnan[1*] Shaohui Liu[1*] Paul-Edouard Sarlin[2*] Oscar Gentilhomme[1]
David Caruso[3] Maurizio Monge[3] Richard Newcombe[3] Jakob Engel[3] Marc Pollefeys[1,4]
[1]ETH Zürich [2]Google [3]Meta Reality Labs Research [4]Microsoft Spatial AI Lab

Figure 1. **We introduce a dataset to benchmark multi-sensor VIO/SLAM with egocentric data at city scale.** We record hours and kilometers of trajectories with Project Aria devices [21], with pose annotations derived from centimeter-accurate surveyed control points. We cover unique challenges of egocentric data, including low light, exposure changes, moving platforms, time-varying calibration, *etc*.

Abstract

Precise 6-DoF simultaneous localization and mapping (SLAM) from onboard sensors is critical for wearable devices capturing egocentric data, which exhibits specific challenges, such as a wider diversity of motions and viewpoints, prevalent dynamic visual content, or long sessions affected by time-varying sensor calibration. While recent progress on SLAM has been swift, academic research is still driven by benchmarks that do not reflect these challenges or do not offer sufficiently accurate ground truth poses. In this paper, we introduce a new dataset and benchmark for visual-inertial SLAM with egocentric, multi-modal data. We record hours and kilometers of trajectories through a city center with glasses-like devices equipped with various sensors. We leverage surveying tools to obtain control points as indirect pose annotations that are metric, centimeter-accurate, and available at city scale. This makes it possible to evaluate extreme trajectories that involve walking at night or traveling in a vehicle. We show that state-of-the-art systems developed by academia are not robust to these challenges and we identify components that are responsible for this. In addition, we design tracks with different levels of difficulty to ease in-depth analysis and evaluation of less mature approaches. The dataset and benchmark are available at lamaria.ethz.ch.

*indicates equal contribution

1. Introduction

Estimating the precise location of a camera over time is a fundamental problem in computer vision. Algorithms like Visual-Inertial Odometry (VIO) or Simultaneous Localization and Mapping (VI-SLAM) can estimate a 6 Degrees-of-Freedom (DoF) pose for each image of a sequence, often aided by inertial sensors. Positioning plays a crucial role in ensuring the persistence of digital content over time and enabling seamless sharing across devices, which is especially important for applications like AI assistants and augmented reality. Progress in mobile computing has fueled the development of wearable devices that are equipped with various sensors, including multiple color or depth cameras, inertial units, and radio receivers. The egocentric, multi-modal data that they capture presents unique challenges that are often overlooked in computer vision research, which typically relies on curated datasets with controlled viewpoints and motions tailored to algorithms or visual content of interest.

Differently, egocentric data is passive and accidental: it does not constrain the user's actions but rather endures them. As a result, this data exhibits significantly more diversity in motion patterns, viewpoints, and environments than typically found in computer vision datasets. Moreover, egocentric devices aspire to be all-day wearables that capture data over extended durations, in which factors like sensor calibration

dataset	data			sensors			ground-truth		challenges		
	motion	environment	multi-seq	multi-cam	IMU	others	source	accuracy	duration	dynamics	lighting
EuRoC [10]	drone	small	no	yes	yes	no	mocap	∼cm	<3 min	no	partial
TartanAir [67]	random	large	no	yes	no	depth,LiDAR	synthetic	perfect	→20 min	yes	yes
4Seasons [68]	car	large	yes	yes	yes	GNSS	VI+GNSS	>dm	350 km	moderate	yes
VBR [8]	car,handheld	large	partial	yes	yes	LiDAR, GNSS	LiDAR+IMU +RTK-GNSS	∼cm	→50 min	moderate	partial
TUM-RGBD [61]	handheld	small	no	no	no	depth	mocap	<cm	<3 min	no	no
TUM-VI [60]	handheld	medium	no	yes	yes	no	mocap	<cm	→25 min	no	no
ADVIO [53]	handheld	medium	yes	no	yes	no	VI-SLAM	∼dm	∼3 min	moderate	no
ETH3D-SLAM [59]	handheld	small	no	yes	yes	depth	mocap	<cm	<4 min	moderate	yes
NewerCollege [52, 70]	handheld	medium	yes	yes	yes	LiDAR	LiDAR-SLAM	∼cm	→26 min	no	no
Hilti-Oxford [71]	handheld	medium	yes	yes	yes	LiDAR	surveying	<cm	→17 min	no	partial
Hilti-UZH [45]	robot,handheld	medium	yes	yes	yes	LiDAR	surveying	<cm	→12 min	no	partial
LaMAR [56]	head-mounted handheld	medium	yes	yes	uncalibrated	GNSS, WiFi, BT	V-SLAM +LiDAR	∼dm	∼5 min	moderate	yes
LaMAria (ours)	head-mounted handheld	large	yes	yes	yes (×2)	GNSS, WiFi, BT	surveying	∼cm	→45 min	people,tram funicular	yes

Table 1. **Overview of existing datasets.** LaMAria is the first egocentric dataset that is recorded in city-scaled large environments, has multiple calibrated camera and IMU sensors, includes long sequences up to 48 minutes, and covers all challenges including dynamic environments, moving platforms, and varying lighting conditions, while still providing centimeter accurate pose annotations from surveying.

can change over time. Finally, the wearability and consumer adoption of these devices limits the size, weight, and cost of these sensors, and thus their quality.

Academic research in VIO/SLAM is mainly driven by benchmarks that do not exhibit the characteristics of egocentric data. Often originating from the robotics community [10, 68], their data is recorded by expensive, industrial-grade sensors mounted on robots with limited locomotion capabilities. The robot's motion can also often be adapted to accommodate the limitations of the perception algorithms, as in active perception. Additionally, datasets that offer sufficiently accurate ground-truth (GT) camera poses are often limited to smaller environments than the ones found in egocentric applications [45, 59, 60, 71]. The datasets that offer egocentric data [4, 38, 39, 56] do not have sufficiently accurate GT poses to measure improvements in VIO/SLAM algorithms without saturation.

In this paper, we introduce LaMAria, a new dataset and benchmark[1] to track progress in egocentric SLAM (Fig. 1). We record data with Project Aria devices [21], which capture rich multi-sensor streams in a glasses-like form-factor, such that they can be worn over extended durations and distances without impeding the wearer's motion. The dataset thus exhibits all key characteristics of egocentric data, with a focus on challenges that break existing algorithms: extremely low illumination, fast motion, large distances, transition between indoors and outdoors, time-varying calibration, and dynamic content – the wearer's own body, other people, or even moving environments such as elevators and vehicles. The trajectories cover the large area of a city center, with some of them spanning kilometers. They benefit from a metric, centimeter-accurate ground-truth based on sparse control points (CPs) widely used in the surveying community.

We evaluate state-of-the-art VIO/SLAM systems with over 22 h and 70 km of egocentric data under different sensor configurations and across different difficulty levels and types of challenges. Our results suggest that the top methods developed by academia are still far from solving this benchmark, while exhibiting a significant gap against Aria's SLAM API. Additional sequences offer gradually increasing difficulty levels between controlled hand-held motion, as exhibited by most academic datasets, and challenging unrestricted head-mounted motion. All evaluated methods perform well with controlled motion but significantly break down as it becomes more natural and egocentric.

Our results shed light on the limitations of existing systems while our dataset opens new avenues for multi-sensor SLAM. The dataset and benchmark is publicly released to ease tracking progress in this direction.

2. Related work

Odometry and SLAM: Visual odometry (VO) [47] estimates motion incrementally in a causal manner, while SLAM extends it by incorporating loop closure to correct drift and can be formulated as a batch optimization problem. Both are related to Structure-from-Motion (SfM) [1, 18, 49, 58], which more generally handles unordered image collections in an offline manner. While early visual SLAM methods [16] employ Extended Kalman Filtering (EKF), PTAM [29] and early ORB-SLAM systems [42, 43] have utilized bundle adjustment with non-linear least square optimization and achieved improved accuracy. While these approaches rely on image matching via sparse interest points [11, 37, 55], others align pixel intensities [19, 46] and achieve good stability when a good initialization is available. These have been later developed into sparse SLAM systems [20, 23].

With the reduction in size and cost of inertial sensors, visual-inertial odometry and SLAM have gained significant attention. Early methods were mainly filter-based, pioneered

[1] All data collection, storage, and hosting was performed by ETH Zürich.

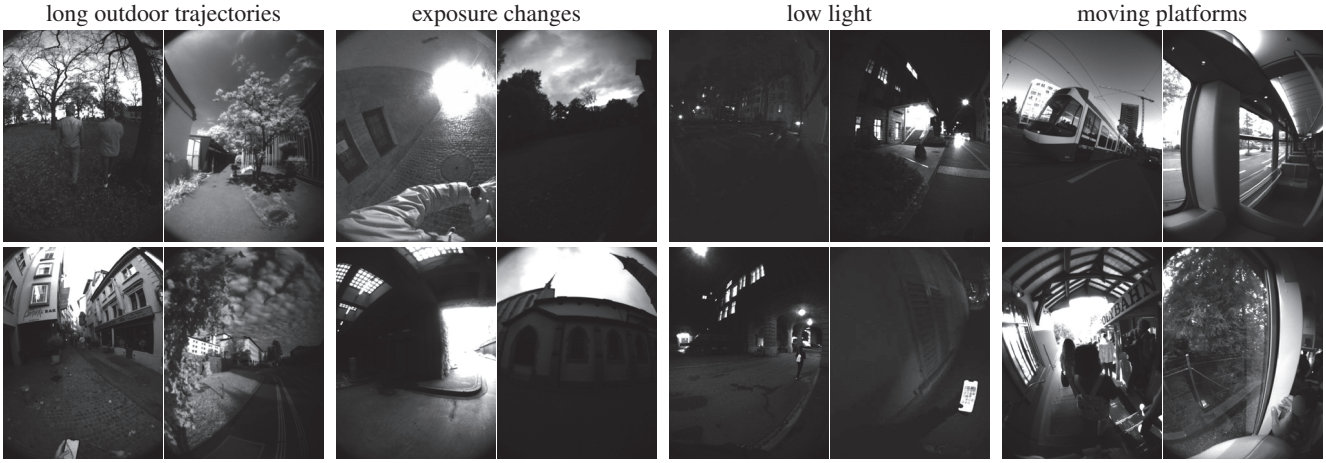


Figure 2. **Challenges:** LaMAria includes sensor data recorded by head-mounted devices following outdoor and indoor trajectories in diverse conditions and environments that impair the perceived visual information and are thus challenging for existing algorithms.

by MSCKF [41] and later extended into full-fledged systems [6, 26, 57]. Other approaches rely on factor graph optimization [2, 17, 28, 30] to achieve higher performance with batch VI optimization [12, 31, 32, 50, 54]. Direct approaches have also been shown to benefit from inertial constraints [64, 65]. As multi-camera system become ubiquitous, many systems also support stereo (horizontally aligned/rectified) [54, 66] or binocular [26, 31] modes.

Deep learning too can increase the robustness of VO/SLAM using learned geometric priors [7, 15, 33, 44] and differentiable optimization [13, 35, 62, 63], especially in dynamic environments [5, 13, 69, 72]. It can also benefit systems that rely on inertial constraints [9, 36, 74]. While many of these approaches exhibit higher robustness and accuracy on short videos, they are computationally expensive and cannot scale to trajectories that span kilometers.

Our dataset makes it possible to evaluate different sensor configurations that include one or multiple cameras and inertial sensors, on both short and long trajectories.

Datasets and benchmarks: There exists many well-established datasets for VIO/SLAM, as shown in Tab. 1. Most are captured with sensors mounted on cars, robots, or handheld devices, and thus exhibit significantly different motion than devices head-worn in everyday activities. While mocap systems and indoor surveying instruments are very accurate, they severely limit the scale of the data capture. Few datasets cover large scale environments, and they generally are automotive and rely on bulky GNSS receivers [8, 68]. Moreover, egocentric data exhibits a large number of unique challenges, such as time varying calibration, low-light conditions, dynamic environments, moving platforms, *etc*. None of the existing VIO/SLAM datasets reflects the constraints and opportunities of the egocentric setup. LaMAR [56] offers egocentric data but is mainly designed for benchmarking offline localization and mapping. Its ground truth poses and sensor calibration are thus not sufficiently accurate to evaluate VIO/SLAM. On the other hand, existing datasets recorded with Aria devices focus on semantic tasks in small indoor spaces [4, 38, 39]. In this paper, we present a city-scale dataset that covers typical challenges found in egocentric data, with centimeter-level pose annotations to reliably evaluate modern VIO/SLAM systems.

3. Dataset

We now give an overview of the content of our dataset.

Device and sensors: We leverage data collection devices of Project Aria [21], which embed multiple sensors in a glasses-like form-factor. These sensors include two synchronized grayscale global-shutter cameras (640×480, 20 FPS), a rolling-shutter RGB camera (1408×1408, 10 FPS), two inertial measurement units (IMUs, 1 kHz and 800 Hz), a magnetometer (10 Hz), a barometer (50 Hz), a thermometer, a GNSS receiver (1 Hz), and WiFi and Bluetooth transceivers (0.1 Hz). They are mounted on a rigid frame, factory-calibrated, and accurately timestamped. This makes the resulting data ideal for multi-sensor odometry and SLAM. The recording is controlled by a user-friendly mobile app and is based on the efficient VRS file format [40], which is optimized for long recordings.

Environment and setup: We recorded data in the city center of a mid-sized European city over 6 months. It spans an area of approximately 1.5 km^2 and over 50 m of elevation. It includes an old town, river banks, several busy tram stations, and a university campus. Participants were given devices and asked to record multiple sequences through the city. Each trajectory observes 5 to 30 unique fiducial markers located on control points, whose positions are accurately known. We rely on them to compute GT device poses, as later explained in Sec. 4. We obtained 63 sequences, each spanning on average 1.5 km and 26 min. The longest one reaches 2.87 km and 48 min.

Challenges: Participants were not familiar with the robustness of existing VIO/SLAM algorithms, ensuring that the data distribution is not biased towards benchmarking. They most often walked but occasionally also traveled in a tram or a funicular. This *moving platform* scenario is very challenging as it introduces a discrepancy between visual and inertial constraints but is rarely found in academic benchmarks. Sequences were recorded at different times of the day but also at night, yielding low-light, uninformative images. The device is head-worn in most sequences but is occasionally hand-held, with participants removing the device and putting it back on. This introduces deformations of the frame that may be accounted for with time-varying extrinsic calibration. The intrinsic calibration can also vary over long sequences as the device temperature increases, which particularly affects the focal length. Finally, some sequences also traverse indoor buildings and exhibit indoor-outdoor transitions that often introduce over or underexposure (Fig. 2).

Controlled experimental set: Alongside the main dataset, we collect a curated experimental set that serves as an entry-level testbench for VIO/SLAM systems. This helps understand why egocentric data is much more challenging than existing benchmarks. This set contains sequences with 4 difficulty levels, as follows:

- **Level I:** platform-based data, with controlled motion and only in-plane rotation.
- **Level II:** platform-based data, with controlled motion and both in-plane and out-of-plane rotation.
- **Level III:** platform-based data, with fast and complex motion / rotation but controlled initial motion.
- **Level IV:** egocentric data, with controlled initial motion.

Level II that has four sequences while the other levels each have three sequences. The gradually increasing difficulty helps identify system failure cases and bridges the gap between academic datasets and our full egocentric data.

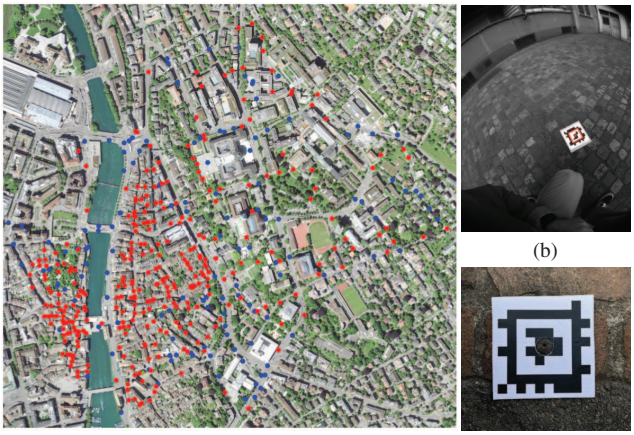
Privacy: We blur visible faces and license plates using Aria's EgoBlur [51].

4. Ground-truth generation

We can sparsely evaluate any trajectory based on highly-accurate but sparse ground control points measured independently. We can further recover a pseudo-ground-truth for the device trajectory through sensor fusion.

4.1. Control points

Characteristics: Control points (CPs) are 3D points commonly used in the fields of surveying and photogrammetry to anchor sensor measurements in a common reference system across time. Their position is measured in a geographic coordinate system with a survey instrument like a GNSS-RTK rover, which corrects the error of a GNSS measurement using a nearby base station. Their uncertainty includes the uncertainty of the base station and increases with the distance between the rover and the station. As with GNSS, the measurement is reliable only when there is a clear line-of-sight with multiple satellites. As such, it is not reliable in urban canyons, under overpasses, and indoors. In these scenarios, surveyors typically propagate constraints from nearby reliable CPs using total stations, which measure absolute distances and relative orientations between CPs.



(a) 483 2D and 3D control points cover the area. (b) (c)

Figure 3. **Control points** (a) are measured with centimeter-accuracy by surveying instruments and are (b) automatically detected in Aria's images using (c) fiducial markers.

Public points: In many countries, public administrations maintain a registry of CPs in each city. They are typically used to anchor construction work on public or private ground. These CPs are generally defined on urban features that are standardized, easy to recognize, and stable over time, such as marked stones or metal bolts sealed in cement. The position of these CPs is publicly available and regularly updated. Their measurement process is often documented and their accuracy is guaranteed. In the area in which our dataset is captured, public CPs are located on average every 35 m, with an horizontal uncertainty of 1 cm.

Usage: Some public CPs do not provide height information but only a 2D horizontal constrain. We measured this missing information with a GNSS-RTK rover (Emlid Reach RS3), when reliable. We added and measured additional 3D CPs in areas where public CPs have a low density. In total, this results in 483 CPs (134 3D and 349 2D), shown in Fig. 3a. Our measurements have an uncertainty of $\sim$1.5 cm horizontally and 3 cm vertically. It is close to identical for all CPs because the distance of the base station is approximately constant (4.5 km). We measure each CP three times, on different days, to further validate its uncertainties (see Appendix D for details).

Detection: To automatically detect control points in Aria imagery, we attach fiducial markers on top of them. We choose the AprilTag markers [48] as they offer a special layout at the

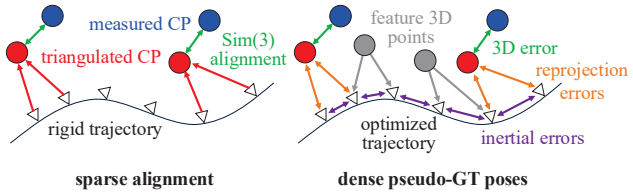


Figure 4. **Types of ground-truth.** Left: Any trajectory can be evaluated with high accuracy via sparse alignment against the GT control points (CPs). Right: We also compute GT camera poses, which are denser but less accurate, via a joint multi-sensor optimization.

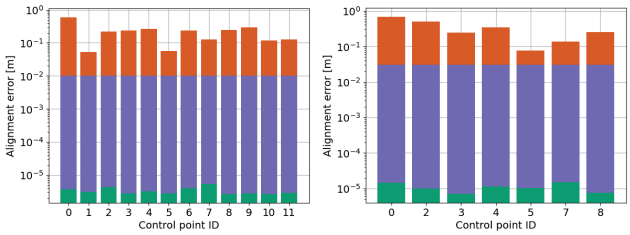


Figure 5. **Cross-validation of CPs for sparse alignment.** The uncertainty of the triangulations and CP measurements are orders of magnitude smaller than the CP error that we evaluate, in both 2D (left) and 3D (right), validating that our sparse GT is sufficiently accurate for evaluation.

center of which there is no information. We thus pierce a hole in the center of each marker and use it to accurately align the marker with the CP (Fig. 3c). Each marker is placed on its corresponding CP by a second helper participant shortly before being observed by the Aria device, ensuring that CP and markers are well-aligned. After constructing the dataset, we blur the detected fiducial markers to avoid biasing the evaluated VIO/SLAM algorithms.

4.2. Sparse alignment

We can readily use the detected control points to evaluate any given trajectory and its camera calibration by rigidly aligning them (Fig. 4). We denote $\mathcal{I}$ the set of image indices and, for an image $i \in \mathcal{I}$, its pose ${}_{\mathsf{L}}\mathbf{T}_i \in \mathrm{Sim}(3)$, which might include the composition of the device pose and rig extrinsics, in the local SLAM coordinate system L.

Process: For each of the N CPs, we use the 2D detections $\{\mathbf{p}_i^n \in \mathbb{R}^2 \; \forall i \in \mathcal{V}(n) \subset \mathcal{I}\}$ of fiducial markers to triangulate its 3D position $\mathbf{P}_{\mathsf{L}}^n$ in L, by minimizing the reprojection error:

$$E_{\text{tri}}^n = \sum_{i \in \mathcal{V}(n)} \left\| \Pi\left({}_{\mathsf{L}}\mathbf{T}_i^{-1} \cdot \mathbf{P}_{\mathsf{L}}^n, \mathbf{C}_i \right) - \mathbf{p}_i^n \right\|^2 \quad . \tag{1}$$

Here $\Pi(\cdot)$ is the image projection function, given camera intrinsics $\mathbf{C}_i$. We first find an initial estimate using a LO-RANSAC scheme [14, 22, 58] and subsequently refine it with non-linear least-squares.

Next, we aim to estimate the similarity transformation ${}_{\mathsf{W}}\mathbf{T}_{\mathsf{L}} \in \mathrm{Sim}(3)$ that best aligns the triangulated $\{\mathbf{P}_{\mathsf{L}}^n\}$ and measured $\{\mathbf{P}_{\mathsf{W}}^n\}$ CPs. Intuitively, this can be done by minimizing the alignment error:

$$E_{\text{sim}} = \sum_{n=1}^{N} \| \mathbf{P}_{\mathsf{W}}^n - {}_{\mathsf{W}}\mathbf{T}_{\mathsf{L}} \cdot \mathbf{P}_{\mathsf{L}}^n \|^2 \quad . \tag{2}$$

However, in practice triangulations may have different uncertainties and can get unreliable when the camera poses exhibit ill-posed configurations. To take this into consideration, we solve for the optimal similarity transformation ${}_{\mathsf{W}}\mathbf{T}_{\mathsf{L}}$ via joint optimization of ${}_{\mathsf{W}}\mathbf{T}_{\mathsf{L}}$ and proxy 3D points $\hat{\mathbf{P}}_{\mathsf{L}}^n$ over the two factors defined in Eqs. (1) and (2), weighted by their respective covariances. This formulation is able to account for uncertainties of the camera poses, although this is often not exposed by the algorithms that we evaluate.

Use for evaluation: We use the optimal ${}_{\mathsf{W}}\mathbf{T}_{\mathsf{L}} \in \mathrm{Sim}(3)$ from the alignment to transform the original triangulated point and measure its error against the corresponding CP: $\|\mathbf{P}_{\mathsf{W}}^n - {}_{\mathsf{W}}\mathbf{T}_{\mathsf{L}} \cdot \mathbf{P}_{\mathsf{L}}^n\|$. These alignment errors define the accuracy of the trajectory: more accurate camera poses better fit the CPs, up to their measurement uncertainty. Intuitively, for a trajectory of 1.5 km and CPs with an horizontal uncertainty of 3 cm, we can measure scale drift of 0.002 %, which is much lower than the error of state-of-the-art systems – stereo ORB-SLAM3 [12] and mono DM-VIO [64] both report 0.6 % on the EuRoC dataset [10].

Our approach is different from other large-scale SLAM benchmarks [53, 56, 68], which have dense but less accurate GT device poses. Hilti benchmarks [45, 71] also have sparse CPs but those are not detected in the images. Instead, the capture device is directly positioned on each CP. This makes its motion artificial, as the device remains static for a few seconds, and is tedious, as precise alignment takes time. Our solution has minimal effect on the motion pattern and only requires to observe each CP from different angles to maximize the triangulation accuracy. This can be easily performed by non-expert participants.

Reliability of the sparse evaluation: We aim to study whether the uncertainty in triangulation will affect the CP alignment error in our sparse evaluation. We use a trajectory estimated by the best approach (Aria's SLAM, as found later in our benchmark in Sec. 5). We perform a Leave-one-out Cross Validation (LOOCV) in which we iteratively ignore one CP in the alignment process and compare its alignment error with its aggregated uncertainty:

$$\mathbf{\Sigma} = \mathbf{\Sigma}_{\text{tri}_{\text{metric}}} + \hat{\mathbf{\Sigma}}, \tag{3}$$

where $\mathbf{\Sigma}_{\text{tri}_{\text{metric}}}$, $\hat{\mathbf{\Sigma}}$ are the covariances of the transformed triangulation and the corresponding CP respectively. The original triangulation covariance $\mathbf{\Sigma}_{\text{tri}}$ is estimated via the inverse of the Gauss-Newton Hessian from the triangulation problem, and can be transformed with scale s and rotation matrix $\mathbf{R}$ from ${}_{\mathsf{W}}\mathbf{T}_{\mathsf{L}} \in \mathrm{Sim}(3)$, where $\mathbf{\Sigma}_{\text{tri}_{\text{metric}}} = s^2 \mathbf{R} \mathbf{\Sigma}_{\text{tri}} \mathbf{R}^T$.

Fig. 5 shows the uncertainty (square root of the covariance spectral norm) of the aligned triangulation and CP measurement on a sequence with 12 CPs. The CP error is on average 70 times larger than its uncertainty, which confirms that our sparse GT is by far sufficiently accurate to benchmark the next generations of SLAM systems.

4.3. Dense ground-truth poses

Having accurate camera poses instead of sparse points is however useful *e.g.*, for fine-grained evaluation and analysis or other computer vision tasks like 3D reconstruction. As such, we compute pseudo-GT poses and sensor calibrations by fusing visual, inertial, and CP information, thus propagating the CP constraints to the poses connecting them.

Process: We start with an initial SLAM trajectory. While our approach is general, here we rely on the result of the proprietary VI-SLAM exposed by Aria's SLAM API since it performs the best compared to open-source systems (see Sec. 5 for more details). We then align the trajectory sparsely, as described in Sec. 4.2, perform steps typical in Structure-from-Motion: we select representative keyframes, extract local features [73], match them based on sequential pairs and image retrieval [3, 34], and finally triangulate a sparse 3D point cloud. Finally, we fix the time-varying intrinsics and refine the camera poses, inertial biases and speeds, triangulated CPs, and 3D points by jointly minimizing a) the CP triangulation error (Eq. (1)), b) the error between measured and triangulated CPs, as well as c) feature reprojection errors and d) inertial pre-integration constraints typically used in VI-SLAM [24]. This is a standard non-linear least-squares optimization problem.

We weight these four terms by their respective covariances. The covariances of the inertial terms are computed using pre-integration [24], with factory-calibrated IMU noise parameters. We use two different covariances for the detection of features and fiducial markers, which are iteratively estimated and refined based on the variance factor of the corresponding residuals [25]. Because we know that the CP measurements are unbiased, we deflate their covariance w.r.t. the one estimated by the surveying instrument.

For the most challenging sections, *e.g.*, involving moving platforms, neither Aria's SLAM nor any other baseline provides a usable initialization because the visual information is not reliable. In these cases, we rely only on the inertial and CP information, ignoring visual features.

Accuracy Validation: We consider the multi-factor optimization described above, which minimizes different costs $\mathbf{r}^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{r}$, with residuals $\mathbf{r} \in \mathbb{R}^d$ whitened by their measurement covariance $\hat{\mathbf{\Sigma}}$. We collect all whitened residuals for visual and inertial factors respectively. Fig. 6 shows that the distribution of residuals is close to a unit Gaussian distribution for both types, which validates the appropriate weighting

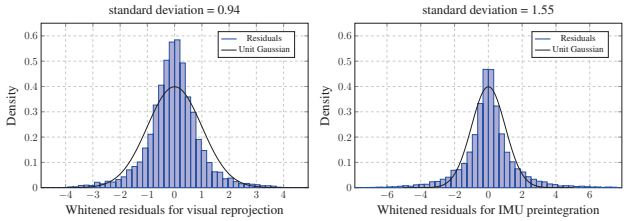
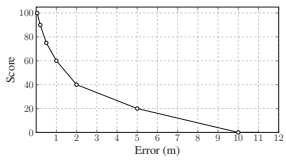


Figure 6. **Distribution of the whitened visual and IMU residuals.**



error e	5 cm	20 cm	50 cm	1 m
score $s(e)$	100	90	75	60
error e	2 m	5 m	10 m	>10 m
score $s(e)$	40	20	0	0

Figure 7. **Scoring function $s(e)$ for the sparse evaluation** given the measured alignment error. The function is piecewise linear as shown in the left plot, with its anchor points in the right table.

of the visual and inertial terms in the optimization [25].

We further validate this dense optimization by dropping the two CP-related factors and evaluating the trajectory sparsely (Sec. 4.2). We find that our pseudo-ground-truth has an overall accuracy of ∼20 cm (see Appendix D for details), which is sufficiently accurate to measure keyframe errors larger than 50 cm in trajectories spanning kilometers.

5. Evaluation

We can use the sparse alignment defined above to evaluate any estimated trajectory on our full set. In the following part, we present our benchmark results and analysis for various publicly-available visual (-inertial) odometry/SLAM systems under different sensor configurations.

Systems: We evaluate systems under monocular, monocular-inertial and multi-camera-inertial setups. The latter relies on the two global-shutter SLAM cameras, which cannot be used in a proper stereo (rectified) mode because they have little overlap in field of view. We try with our best efforts to cover most major popular systems, including those based on optimization (Kimera VIO [54], ORB-SLAM3 [12], OKVIS2 [31]), filtering (OpenVINS [26]), direct alignment (DSO [20], DM-VIO [64]), and deep learning (DPVO [63], DPV-SLAM [35]). In addition, we evaluate the offline VI optimization in Maplab [57] on top of the odometry output of OpenVINS [26]. We make sure to appropriately tune on our data the hyperparameters of each approach, often with the help of the respective authors.

Calibration: Unless stated otherwise, we use the factory calibration provided by the device. The systems are thus responsible for refining this calibration and modeling temporal variations. However, all systems that we consider, with the exception of OpenVINS [26], do not support online intrinsic calibration and implement only few simple camera

method	year	level I			level II				level III			level IV		
		seq 1	seq 2	seq 3	seq 4	seq 5	seq 6	seq 7	seq 8	seq 9	seq 10	seq 11	seq 12	seq 13
DSO	2016	0.76	0.21	×	1.95	19.80	×	30.12	×	×	×	×	×	×
ORB-SLAM3	2020	0.13	0.35	0.12	0.72	1.25	×	5.94	×	×	×	×	×	×
DPVO	2023	0.22	0.16	0.28	1.10	5.40	1.74	5.15	7.14	18.40	15.65	42.07	28.73	39.43
DPV-SLAM	2023	0.12	0.20	0.29	5.35	2.93	2.95	8.10	2.83	19.45	19.48	28.02	29.43	50.41
Kimera VIO	2020	0.82	2.38	0.74	7.23	4.06	11.06	12.59	9.70	32.24	×	7.98	13.11	×
ORB-SLAM3	2020	0.03	0.43	0.23	2.01	1.51	2.70	5.90	7.80	19.10	21.80	×	14.60	×
OpenVINS	2020	0.75	3.15	0.96	2.32	3.94	6.07	5.96	9.23	23.78	8.62	5.87	28.00	17.57
OpenVINS + Maplab	2022	0.71	3.09	0.77	1.23	3.92	6.00	4.86	8.87	23.90	7.73	5.92	26.30	15.44
DM-VIO	2022	0.75	0.34	0.26	0.78	32.03	×	×	×	×	×	10.70	×	×
OpenVINS	2020	0.66	2.36	0.68	0.94	1.43	1.35	2.96	4.25	4.31	8.01	1.04	18.72	10.35
OpenVINS + Maplab	2022	0.65	2.30	0.68	1.05	1.22	1.19	2.01	3.97	4.29	8.22	1.62	16.59	8.37
OKVIS2	2024	0.02	0.72	0.03	1.36	0.80	3.78	×	6.81	5.32	7.06	1.85	16.55	6.65

Table 2. **Results for the controlled experimental set.** We evaluate systems on monocular, monocular+inertial, and multi-camera+inertial inputs. We report the ATE RMSE [27] (lower is better) in meters for each sequence. Failures to output a valid trajectory are marked as ×.

method	causal	short			medium			long			challenge – low-light			challenge – moving platform		
		score ↑	CP@1m ↑	R@5m ↑	score ↑	CP@1m ↑	R@5m ↑	score ↑	CP@1m ↑	R@5m ↑	score ↑	CP@1m ↑	R@5m ↑	score ↑	CP@1m ↑	R@5m ↑
DPVO	✓	9.4	1.7	21.3	5.2	1.0	10.8	1.2	0.0	1.9	3.4	0.2	7.5	2.4	0.1	–
DPV-SLAM	x	7.5	1.5	14.8	5.2	1.4	10.1	0.4	0.0	0.7	1.9	0.4	3.5	1.7	0.0	–
Kimera VIO	✓	6.3	2.9	12.6	6.6	1.7	15.1	6.3	1.7	14.3	4.2	2.7	6.4	7.1	1.6	–
ORB-SLAM3	x	28.3	13.4	67.1	20.3	4.4	57.0	14.2	2.3	40.6	6.2	0.6	12.5	15.7	4.1	–
OpenVINS	✓	18.1	4.4	45.7	10.9	2.3	27.9	4.7	0.5	12.3	7.9	2.4	17.6	2.4	0.6	–
OpenVINS + Maplab	x	22.9	8.1	50.8	13.1	4.1	29.0	5.8	1.3	13.3	9.6	2.9	19.3	3.7	1.2	–
OpenVINS	✓	22.2	6.2	57.9	17.8	5.7	46.1	10.6	1.7	25.8	16.9	6.2	38.2	11.5	2.4	–
OpenVINS + Maplab	x	26.0	9.5	61.1	21.3	7.3	50.6	12.6	1.9	30.3	16.5	4.6	37.9	13.0	3.0	–
OKVIS2	x	24.2	12.0	54.7	13.6	6.8	28.2	3.6	2.7	7.2	15.4	5.4	38.6	4.2	2.8	–
Aria's SLAM	x	90.7	99.2	–	78.5	87.4	–	70.8	75.9	–	84.2	91.6	–	53.6	51.2	–

Table 3. **Main evaluation.** We evaluate systems on monocular, monocular+inertial, and multi-camera+inertial inputs. We also include the closed-source SLAM system exposed by Aria's SLAM API. We report the average score (defined in Fig. 7), recall w.r.t. control points at 1 meter (CP@1m), and recall w.r.t. pseudo-GT at 5 meters (R@5m), all of which are evaluated in 2D.

models. When necessary, we therefore undistort the images to pinhole cameras, using the factory calibration, and run the algorithms on the resulting images.

5.1. Controlled experimental set

We first evaluate the systems on our controlled experimental set to better understand their failure patterns.

Setup: We rely on our dense pseudo-GT poses for sequences in level IV but, for sequences in level I-III, since they are short and do not include CPs, we use the output of Aria's SLAM API as pseudo-GT. Following common practice in existing benchmarks [10, 60], we align the output trajectories with the GT poses and calculate the average of the Absolute Trajectory Error (ATE) [27] for each sequence. We report a failure if the system crashes or only produces results that span less than half of the total sequence duration.

Results: Tab. 2 shows that all approaches work reasonably well on level I. This is consistent with the findings of commonly-used academic datasets like EuRoC [10], whose data exhibits similar controlled motions. However, as the motion becomes more natural/egocentric, most systems start to break down by either reporting failures or suffering from drifts and large trajectory error. In particular, the two direct methods and ORB-SLAM3 mono easily lose track due to fast and complex motion patterns. This motivates the need for a VIO/SLAM benchmark that captures such motions in unrestricted egocentric recordings with a larger span in length and duration.

5.2. Main benchmark

Setup: We categorize 63 sequences in our main dataset into short (up to 15 CPs, 18 sequences), medium (16-22 CPs, 10 sequences), long (more than 22 CPs, 16 sequences) and highlight specific hard challenges with low-light (9 sequences) and moving platforms (10 sequences). We provide more statistics in the supplementary material. Inspired by the Hilti challenge [45, 71], we design a scoring function for each of the CP error (Fig. 7). We employ a piecewise linear scoring function, where we can confidently evaluate until 5 cm with our survey-grade control points.

Along with the score averaged over the sequences, we also report the recall of the CP alignment error at 1 m and the recall of the device position error w.r.t. our dense pseudo-GT poses at 5 m. Unlike the standard ATE, these three metrics can be computed even when systems estimate no or only partial trajectories. We do not report the pose recall for Aria's SLAM, as it is used to initialize the optimization of pseudo-GT poses, and for sequences involving moving platforms, as we have limited guarantees on the accuracy of their pseudo-GT poses (see Appendix E for details).

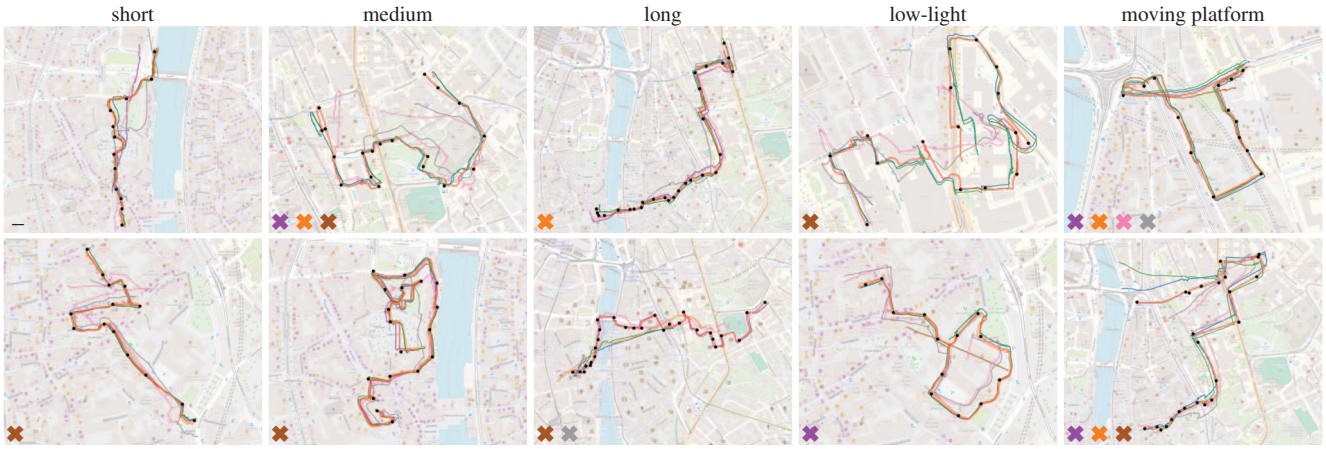


Figure 8. **Visualizations of the trajectories estimated by different systems:** OpenVINS ●, OpenVINS+Maplab ●, ORB-SLAM3 ●, OKVIS2 ●, Kimera VIO ●, DPVO ●, DPV-SLAM ●, and Aria's SLAM ●. Failures are shown as ✖ and control points in black ●.

We evaluate all methods that produce valid results on level IV of the experimental set, according to Tab. 2. Note that the sequences in the short-sequence group are overall more difficult than the level IV recordings due to the presence of non-controlled initial motion, which brings additional practical challenges to bootstrapping and IMU initialization. We perform all evaluation in 2D to make use of all the control points. To account for randomness, we average the metrics over 3 runs of each system.

Results: We show the main results in Tab. 3. As expected, relying on multiple cameras and inertial sensors significantly benefits each VIO/SLAM system. Overall, ORB-SLAM3 scores the best in the mono-inertial category, while OpenVINS achieves the most promising results when considering multi cameras. The non-linear optimization of Maplab also consistently boosts the accuracy of the estimates. However, all academic solutions suffer largely on the defined challenges and are far behind Aria's commercial SLAM system, which is non-causal and includes online calibration and a full visual-inertial bundle adjustment. On the other hand, the evaluation of Aria's SLAM indicates that our benchmark is not saturated even for a heavily engineered system, especially in sequences that include moving platforms. This further highlights that our dataset provides a good benchmark for all practices relevant to developing VIO/SLAM solutions for unconstrained egocentric data.

Qualitative results in Fig. 8 show that most evaluated systems face drift and failures in long recordings. They are also more prone to failure when facing low-light conditions and moving platforms, while Aria's SLAM produces the most reasonable trajectories that best fit the control points.

5.3. Discussions

Our dataset opens up opportunities for more principled iterations of multi-sensor SLAM developments on top of uncontrolled egocentric recordings. Following our empirical studies on all the evaluated systems, we highlight several promising research directions to explore for addressing the unique characteristics of egocentric data:

- Online optimization of time-varying calibration, which distinguishes Aria's SLAM from existing academic baselines, adapts to the always-on nature of the wearable devices.
- Loop closure detection and VI bundle adjustment, to reduce odometry drift in open loop predictions.
- Robust outlier removal, tailored strategies for moving platforms, and better handling of tracking loss.
- Advanced image matching and point tracking based on machine learning models trained on large datasets.

To specifically quantify the importance of online optimization, we analyze the online calibration results from Aria's SLAM. We observe that the variation range of focal length is comparably larger in the long sequences (0.11%) than in medium (0.09%) and short (0.08%) ones. Fixing the calibration to factory calibration in our VI optimization also empirically leads to a large decrease in pose accuracy, when validated with our survey-grade control points.

6. Conclusion

We have introduced a new dataset and benchmark for visual-inertial odometry and SLAM for egocentric, multi-modal data captured by future head-mounted computing devices. The dataset covers multiple challenges that no existing dataset covers: extreme low-light, crowded or moving environments like vehicles, and kilometer-long trajectories with time-varying calibration. The dataset provides sparse, centimeter-accurate pose annotations from surveying for tens of kilometers of trajectories spanning a large city center. This annotation is orders of magnitude more accurate than the best existing visual-inertial odometry/SLAM systems and will enable the development and benchmarking of the next generations of multi-sensor algorithms. Our results show that much progress remains to be made, which will be supported by our public evaluation.

References

[1] Sameer Agarwal, Noah Snavely, Steven M Seitz, and Richard Szeliski. Bundle Adjustment in the Large. In *ECCV*, 2010. 2

[2] Sameer Agarwal, Keir Mierle, and The Ceres Solver Team. Ceres Solver. http://ceres-solver.org, 2023. 3

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016. 6

[4] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing HOT3D: An Egocentric Dataset for 3D Hand and Object Tracking. *arXiv:2406.09598*, 2024. 2, 3

[5] Berta Bescos, José M Fácil, Javier Civera, and José Neira. DynaSLAM: Tracking, Mapping and Inpainting in Dynamic Scenes. *IEEE RA-L*, 2018. 3

[6] Michael Bloesch, Michael Burri, Sammy Omari, Marco Hutter, and Roland Siegwart. Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback. *IJRR*, 2017. 3

[7] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. CodeSLAM - Learning a Compact, Optimisable Representation for Dense Visual SLAM. In *CVPR*, 2018. 3

[8] Leonardo Brizi, Emanuele Giacomini, Luca Di Giammarino, Simone Ferrari, Omar Salem, Lorenzo De Rebotti, and Giorgio Grisetti. VBR: A Vision Benchmark in Rome. In *ICRA*, 2024. 2, 3

[9] Russell Buchanan, Varun Agrawal, Marco Camurri, Frank Dellaert, and Maurice Fallon. Deep IMU Bias Inference for Robust Visual-Inertial Odometry with Factor Graphs. *IEEE RA-L*, 2022. 3

[10] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The EuRoC micro aerial vehicle datasets. *IJRR*, 2016. 2, 5, 7

[11] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: Binary Robust Independent Elementary Features. In *ECCV*, 2010. 2

[12] Carlos Campos, Richard Elvira, Juan J. Gómez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM. *IEEE T-RO*, 2021. 3, 5, 6

[13] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. LEAP-VO: Long-term Effective Any Point Tracking for Visual Odometry. In *CVPR*, 2024. 3

[14] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally Optimized RANSAC. In *GCPR*, 2003. 5

[15] Jan Czarnowski, Tristan Laidlow, Ronald Clark, and Andrew J Davison. DeepFactors: Real-Time Probabilistic Dense Monocular SLAM. *IEEE RA-L*, 2020. 3

[16] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE TPAMI*, 2007. 2

[17] Frank Dellaert. Factor graphs and gtsam: A hands-on introduction. *Georgia Institute of Technology, Tech. Rep*, 2(4), 2012. 3

[18] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3R-SfM: a Fully-Integrated Solution for Unconstrained Structure-from-Motion. *arXiv:2409.19152*, 2024. 2

[19] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *ECCV*, 2014. 2

[20] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE TPAMI*, 2017. 2, 6

[21] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project Aria: A New Tool for Egocentric Multi-Modal AI Research, 2023. 1, 2, 3

[22] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *ACM*, 24(6): 381–395, 1981. 5

[23] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: Fast Semi-Direct Monocular Visual Odometry. In *ICRA*, 2014. 2

[24] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. On-Manifold Preintegration for Real-Time Visual-Inertial Odometry. *IEEE T-RO*, 2017. 6

[25] Wolfgang Förstner and Bernhard P. Wrobel. *Photogrammetric Computer Vision*. Springer, 2016. 6

[26] Patrick Geneva, Kevin Eckenhoff, Woosik Lee, Yulin Yang, and Guoquan Huang. OpenVINS: A Research Platform for Visual-Inertial Estimation. In *ICRA*, 2020. 3, 6

[27] Michael Grupp. evo: Python package for the evaluation of odometry and SLAM. https://github.com/MichaelGrupp/evo, 2017. 7

[28] Michael Kaess, Hordur Johannsson, Richard Roberts, Viorela Ila, John J Leonard, and Frank Dellaert. iSAM2: Incremental Smoothing and Mapping Using the Bayes Tree. *IJRR*, 2012. 3

[29] Georg Klein and David Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *ISMAR*, 2007. 2

[30] Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. g2o: A General Framework for Graph Optimization. In *ICRA*, 2011. 3

[31] Stefan Leutenegger. OKVIS2: Realtime Scalable Visual-Inertial SLAM with Loop Closure. *arXiv:2202.09199*, 2022. 3, 6

[32] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-Based Visual-Inertial SLAM Using Nonlinear Optimization. *IJRR*, 2015. 3

[33] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, Fast, and Robust Structure and Motion from Casual Dynamic Videos. *arXiv:2412.04463*, 2024. 3

[34] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 6

[35] Lahav Lipson, Zachary Teed, and Jia Deng. Deep Patch Visual SLAM. In *ECCV*, 2024. 3, 6

[36] Wenxin Liu, David Caruso, Eddy Ilg, Jing Dong, Anastasios I Mourikis, Kostas Daniilidis, Vijay Kumar, and Jakob Engel. TLIO: Tight Learned Inertial Odometry. *IEEE RA-L*, 2020. 3

[37] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2):91–110, 2004. 2

[38] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria Everyday Activities Dataset. *arXiv:2402.13349*, 2024. 2, 3

[39] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A Massive Collection of Multimodal Egocentric Daily Motion in the Wild. In *ECCV*, 2024. 2, 3

[40] Meta. VRS: A file format designed to record and playback streams of XR sensor data. https://facebookresearch.github.io/vrs/, 2024. 3

[41] Anastasios I Mourikis and Stergios I Roumeliotis. A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation. In *ICRA*, 2007. 3

[42] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *IEEE T-RO*, 2017. 2

[43] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a Versatile and Accurate Monocular SLAM System. *IEEE T-RO*, 2015. 2

[44] Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors. In *CVPR*, 2025. 3

[45] Ashish Devadas Nair, Julien Kindle, Plamen Levchev, and Davide Scaramuzza. Hilti SLAM Challenge 2023: Benchmarking Single + Multi-session SLAM across Sensor Constellations in Construction. *IEEE RA-L*, 2024. 2, 5, 7

[46] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. DTAM: Dense Tracking and Mapping in Real-Time. In *ICCV*, 2011. 2

[47] David Nistér, Oleg Naroditsky, and James Bergen. Visual Odometry. In *CVPR*, 2004. 2

[48] Edwin Olson. AprilTag: A robust and flexible visual fiducial system. In *ICRA*, 2011. 4

[49] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global Structure-from-Motion Revisited. In *ECCV*, 2024. 2

[50] Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE T-RO*, 2018. 3

[51] Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M Parkhi. EgoBlur: Responsible Innovation in Aria. *arXiv:2308.13093*, 2023. 4

[52] Milad Ramezani, Yiduo Wang, Marco Camurri, David Wisth, Matias Mattamala, and Maurice Fallon. The Newer College Dataset: Handheld LiDAR, Inertial and Vision with Ground Truth. In *IROS*, 2020. 2

[53] Santiago Cortés Reina, Arno Solin, Esa Rahtu, and Juho Kannala. ADVIO: An authentic dataset for visual-inertial odometry. In *ECCV*, 2018. 2, 5

[54] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping. In *ICRA*, 2020. 3, 6

[55] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: an efficient alternative to SIFT or SURF. In *ICCV*, 2011. 2

[56] 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. 2, 3, 5

[57] Thomas Schneider, Marcin Dymczyk, Marius Fehr, Kevin Egger, Simon Lynen, Igor Gilitschenski, and Roland Siegwart. maplab: An Open Framework for Research in Visual-inertial Mapping and Localization. *IEEE RA-L*, 2018. 3, 6

[58] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. 2, 5

[59] Thomas Schops, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle Adjusted Direct RGB-D SLAM. In *CVPR*, 2019. 2

[60] David Schubert, Thore Goll, Nikolaus Demmel, Vladyslav Usenko, Jörg Stückler, and Daniel Cremers. The TUM VI Benchmark for Evaluating Visual-Inertial Odometry. In *IROS*, 2018. 2, 7

[61] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A Benchmark For The Evaluation of RGB-D SLAM Systems. In *IROS*, 2012. 2

[62] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. In *NeurIPS*, 2021. 3

[63] Zachary Teed, Lahav Lipson, and Jia Deng. Deep Patch Visual Odometry. In *NeurIPS*, 2023. 3, 6

[64] Lukas von Stumberg and Daniel Cremers. DM-VIO: Delayed Marginalization Visual-Inertial Odometry. *IEEE RA-L*, 2022. 3, 5, 6

[65] Lukas von Stumberg, Vladyslav Usenko, and Daniel Cremers. Direct Sparse Visual-Inertial Odometry using Dynamic Marginalization. In *ICRA*, 2018. 3

[66] Rui Wang, Martin Schworer, and Daniel Cremers. Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras. In *ICCV*, 2017. 3

[67] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A Dataset to Push the Limits of Visual SLAM. In *IROS*, 2020. 2

[68] Patrick Wenzel, Rui Wang, Nan Yang, Qing Cheng, Qadeer Khan, Lukas von Stumberg, Niclas Zeller, and Daniel Cremers. 4Seasons: A Cross-Season Dataset for Multi-Weather SLAM in Autonomous Driving. In *GCPR*, 2020. 2, 3, 5

[69] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion. In *ICLR*, 2025. 3

[70] Lintong Zhang, Marco Camurri, David Wisth, and Maurice Fallon. Multi-Camera LiDAR Inertial Extension to the Newer College Dataset. *arXiv:2112.08854*, 2021. 2

[71] Lintong Zhang, Michael Helmberger, Lanke Frank Tarimo Fu, David Wisth, Marco Camurri, Davide Scaramuzza, and Maurice F. Fallon. Hilti-Oxford Dataset: A Millimeter-Accurate Benchmark for Simultaneous Localization and Mapping. *IEEE RA-L*, 2022. 2, 5, 7

[72] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. ParticleSfM: Exploiting Dense Point Trajectories for Localizing Moving Cameras in the Wild. In *ECCV*, 2022. 3

[73] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter C. Y. Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation Measurement*, 72:1–16, 2023. 6

[74] Xingxing Zuo, Nathaniel Merrill, Wei Li, Yong Liu, Marc Pollefeys, and Guoquan Huang. CodeVIO: Visual-Inertial Odometry with Learned Optimizable Dense Depth. In *ICRA*, 2021. 3