Benchmarking Pedestrian Odometry: The Brown Pedestrian Odometry Dataset (BPOD)

David Charatan
Brown University
Providence, RI
david_charatan@alumni.brown.edu

Hongyi Fan
Brown University
Providence, RI
hongyi_fan@brown.edu

Benjamin Kimia
Brown University
Providence, RI
benjamin_kimia@brown.edu

Abstract

This paper presents the Brown Pedestrian Odometry Dataset (BPOD) for benchmarking visual odometry algorithms on data from head-mounted sensors. This dataset was captured with stereo and RGB streams from RealSense cameras with rolling and global shutters in 12 diverse indoor and outdoor locations on Brown University's campus. Its associated ground-truth trajectories were generated from third-person videos that documented the recorded pedestrians' positions relative to stick-on markers placed along their paths. We evaluate the performance of canonical approaches representative of direct, feature-based, and learning-based visual odometry methods on BPOD. Our finding is that current methods which are successful on other benchmarks fail on BPOD. The failure modes correspond in part to rapid pedestrian rotation, erratic body movements, etc. We hope this dataset will play a significant role in the identification of these failure modes and in the design, development, and evaluation of pedestrian odometry algorithms.

1. Introduction

Visual Odometry (VO) is the process of measuring egomotion using image data. Specifically, VO uses visual data to recover a navigating agent's path relative to its position at an earlier time. This is in contrast to odometry based on other sensory data such as wheel sensors, step counters, global positioning system (GPS), inertial measurement units (IMUs), sonar, infrared, radio frequency (RF) receivers, laser range finders (LIDAR), RGB-D cameras, and others [1, 2]. Visual odometry has become predominant given the versatility and relatively low cost of cameras [3, 4]. Challenges of ambiguous scale, motion blur during rapid rotations, low or repeated texture, and large dynamic range have encouraged the fusion of visual odometry with low-cost and versatile IMUs, resulting in visual-inertial odometry (VIO) [5, 6]. Applications of visual odometry are vast and span planetary exploration, unmanned aerial vehicles (drones or MAVs), [7, 8], autonomous driving [9, 10], augmented reality applications [11], mobile mapping [12], service robotics [13], simultaneous localization and mapping (SLAM) [14], *etc*. A significant application of VO is tracking ground-level pedestrian trajectories, critical in a number of scenarios: *(i)* navigation for the visually impaired, *(ii)* monitoring elderly people navigating in indoor and outdoor environments [15], *(iii)* tracking first responders [16], *(iv)* passive guided shopping in supermarkets and large stores, *(v)* indoor navigation guidance in airports, train stations, and hospitals [17], *(vi)* personalized guided tours in exhibitions, museums, and galleries [18], and *(vii)* pandemic contract tracing [19, 20].

The key to advancing the state of the art in VO is the availability of challenging, high-quality, broadly represented, and task-driven benchmarks. A case in point is the rapid development resulting from the introduction of KITTI [10] and a number of other datasets [21, 22, 23] for the task of autonomous driving. The utility of each benchmark, however, is necessarily limited to the application for which it was designed. For example, autonomous driving benchmarks contain paths that are planar and mostly straight, with a small number of turns that have limited accelerations and radii of curvature. In contrast, other agents such as drones and pedestrians exhibit rapid rotations, high acceleration, and more general paths of motion [24]. The introduction of drone datasets such as EuRoc [25] has led to rapid development of algorithms for drone odometry. However, there has not been a dataset targeting pedestrians, who in particular rotate very rapidly, leading to blur in images, Figure 2, and move more erratically than cars and drones. Pedestrian ground-level tracking is an application for which VO datasets are not available to the best of our knowledge. Existing datasets are commonly

The authors gratefully acknowledge the support of NSF award 1910530. The authors further wish to thank Cameron Fiore, Chiang-Heng Chien and Paul Waltz for their help in collecting and annotating the data.

Figure 1: BPOD spans a diverse set of indoor and outdoor scenes ranging from texture-rich historic buildings to modern construction.

vehicle-mounted [10, 26, 27], Segway-mounted [28], MAV-mounted [25, 29], or hand-held [30, 31, 32, 33]. The rare exception is one sequence (Campus-run) obtained from a head-mounted Velodyne LIDAR, but not from visual cameras. The inclusion of this sequence was to highlight the added importance of the IMU in head-mounted situations which depict erratic movements [34]. As such, the proposed dataset, the Brown Pedestrian Odometry Dataset (BPOD), fills a gap in a significant application area. Our dataset shows that VO techniques which are largely successful on existing datasets, Table 1, do not perform well on BPOD. The new dataset enables the identification of areas where existing techniques fail, thereby paving the way for the development of VO techniques for capturing ground-level pedestrian trajectories.

Benchmarking is an elusive task with numerous subtleties. VO techniques can generally be classified into three categories, namely, feature-based methods [40], direct methods [4, 41], and deep learning methods [42, 43, 44, 45]. Certain datasets can be more suitable for one or the other. Textured, feature-rich scenes favor feature-based methods, while textureless scenes with large homogeneous areas favor direct methods. Illumination variation (*e.g.*, the requirement to illuminate darker environments like those found underground or underwater) impacts the photometric invariance assumption of direct methods [33]. Similarly, a dataset whose images have not been photometrically calibrated (*i.e.,* where exposure times, the camera response function, and lens vignetting have been measured) disfavors direct methods. These and other nuances have led to an abundance of benchmarks with varying targets [35, 25, 38, 39, 36, 28, 29]. Benchmarks are inherently task-oriented and must be constructed carefully to satisfy the requirement of the application at hand.

The main contribution of this paper is the design and development of a dataset to benchmark algorithms recovering **ground-level pedestrian trajectories.** The dataset does not document ground-truth camera pose, which would be welcome in AR/VR applications, but would also require laser odometry or a complex network of cameras, which is unnecessarily complicated and would limit the range and extent of environments for which the trajectory of a pedestrian can be observed. Rather, the application mentioned above requires a pedestrian's ground-level location. The

Figure 2: From left to right, typical turning sequences from vehicle-mounted, hand-held, and head-mounted cameras demonstrate the potential for severe blur in head-mounted sequences.

choice of a head-mounted camera as a surrogate for a projected ground-level body-center position necessarily introduces some errors. However, these errors are acceptable for pedestrian odometry, where avoiding long-term drift is more important than achieving centimeter-scale accuracy at any given time. In fact, our results show that all three types of canonical odometry methods fail to capture pedestrian ground-level trajectories to any reasonable level of accuracy. We measure the quality of our ground-truth data through *(i)* consistency over multiple traversals of the same path, and *(ii)* comparison with successful SfM reconstructions (see supplementary material for cases where SfM reconstruction instead fails).

The creation of a VO-focused dataset like BPOD requires several components: *(i)* A sensory platform for data acquisition: we selected two synchronized stereo cameras, one with a rolling shutter and one with a global shutter mounted on a helmet. Embedded within the latter is an IMU. Note that we do not provide hardware synchronization between the rolling shutter and stereo shutter cameras; synchronization between cameras uses a clapperboard-like method (see Section 4). In addition, one camera can acquire depth images, although the current version of BPOD omits these in favor of higher resolution images and higher frame rates. *(ii)* Video sequences: ours are obtained from indoor and outdoor scenes by a pedestrian wearing the helmet following a path annotated on the ground by small markers. Additional diversity is introduced by including both forward and backward traversal of the same path. An auxiliary camera records the pedestrian's movement. *(iii)* Data processing to correlate pedestrian position to a map of markers to generate ground-truth trajectory data. *(iv)* Experiments to evaluate the performance of VO algorithms: we tested three classes of odometry approaches on BPOD. The BPOD dataset is curated by the Brown University Library and is currently available for public use[1].

2. Related Odometry Datasets

Datasets focused on a variety of settings have been proposed to evaluate the performance of visual odome-

[1] https://repository.library.brown.edu/studio/item/bdr:p52vqgtg/

Table 1: A comparison of representative related datasets.

Dataset	Mounting	Environment	Shutter Type	Image	Ground Truth
KITTI [35]	Car	Outdoors	Global	Stereo/Auto Exposure	GPS
EuRoc [25]	MAV	Indoors	Global	Stereo/Auto Exposure	Motion Capture
TUM [36]	Hand-held	Indoors	Global	Mono./Auto Exposure	Motion Capture
ICL-NIUM [37]	Synthetic	Synthetic	Synthetic	Mono./Synthetic	Synthetic
UMA-VI [38]	Hand-held	Mix	Global	Stereo/Auto Exposure	SfM
PennCOSY [39]	Hand-held	Mix	Mix	Stereo/Auto Exposure	Fiducial/SfM
ADVIO [32]	Hand-held	Mix	Rolling	Mono/ Auto Exposure	IMU
BPOD (Ours)	Head-mount	Mix	Mix	Stereo/Mix Exposure	Marker

try systems. Perhaps the most well-known is KITTI [35], which focuses on outdoor car-mounted stereo imaging, with ground-truth trajectories obtained via GPS. While this dataset showcases a diverse set of driving scenes, it is inherently limited by its outdoor, vehicle-focused setting. The EuRoC MAV dataset [25] provides stereo images from a drone in an indoor environment. Its ground truth is captured by an external motion capture system and a laser tracker. While this provides high-quality ground-truth trajectories, the necessary equipment requires a controlled environment, which prevents the creation of a large-scale dataset for indoor environments. The TUM RGB-D dataset [46] was one of the first datasets focused on hand-held cameras. It features a large set of hand-held video sequences, but the ground-truth time-stamps are not well aligned with the camera. The ground truth of this dataset is also captured by an external motion capture system, so this dataset only covers a limited range of scenes. Additionally, the motion of the hand-held camera is slow and smooth, which is appropriate for hand-held scanning applications, but not for pedestrian ego-motion estimation. Fiducial markers have been used to create ground-truth trajectories in a number of datasets [39] in order to overcome the limitations of external motion capture systems. However, these prominent markers generate extra feature correspondences within the scene (note that the markers used for our ground truths are non-invasive). Several datasets have used SfM to generate ground-truth camera trajectories, *e.g.*, [38] and [47] use COLMAP and Pix4D, respectively, to generate their ground-truths. However, unlike these carefully constructed hand-held and MAV datasets, our dataset contains rapid blur and generate feature-less segments that make SfM-based ground truth generation infeasible. Other datasets omit ground-truth trajectories entirely, instead relying on loop closure. For example, the TUM MonoVO [36] dataset, which features a large set of hand-held, photometrically calibrated, monocular footage, proposes an evaluation metric based on loop closure to evaluate VO without ground-truth trajectories.

In the face of these difficulties in generating ground truths for real cameras, synthetic datasets have also been devised for VO development. The ICL-NUIM dataset [37] consists of 8 sequences of photorealistic synthetic indoor scenes in a monocular setting. However, it is hard for these synthetic datasets to simulate features of real scenes, such as motion blur, erratic movements and pedestrian head bobbing. A summary of representative datasets is listed in Table 1.

3. BPOD Sensory Platform

The camera mount most appropriate for pedestrian visual odometry is a head mount, since cameras with chest-mounted or hand-held configurations are often occluded by the subject's arms or the presence of other pedestrians. The BPOD camera rig is a ski helmet outfitted with a standard GoPro mount. Figure 3 shows the frame we designed and 3D-printed to attach cameras to the helmet.

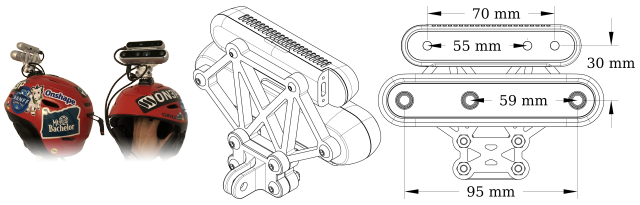

Figure 3: Two views of the camera assembly designed and 3D-printed, shown together with CAD drawings and approximate inter-camera measurements.

Camera selection was guided by several constraints, requiring *(i)* synchronized stereo cameras to compare monocular and stereo configurations for pedestrian odometry; *(ii)* high image quality, resolution, and frame rate; *(iii)* exposure control to allow for precise photometric calibration and to allow for comparison to auto exposure; *(iv)* both rolling shutter and global shutter cameras for comparison; *(v)* a universal driver for compatibility with a variety of devices; *(vi)* an IMU; and, *(vii)* consumer-level pricing.

Given these constraints, we chose a pair of Intel RealSense cameras [48] whose specifications are outlined in

Table 2: The streams we recorded for each sensor.

Camera	Sensors	Resolution	Frame Rate (Hz)	Sensor Aspect Ratio	Focal Length	FOV	Baseline	Shutter Type
D455	Color	1280 ×800	30	16:10	1.88mm	77°	95mm	Global
	Stereo Monochrome	1280 ×800	30	8:5	1.93mm	100.6°		
D415	Color	1920 ×1080	30	16:9	1.88mm	77°	55mm	Rolling
	Stereo Monochrome	1280 ×720	30	16:9	1.88mm	77°		

Table 2. An additional advantage of this choice is the ability to capture depth maps, although we chose not to incorporate this feature into this first version of BPOD, mainly to meet data transfer bandwidth constraints.

Intrinsic and Extrinsic Camera Calibration: Calibration of a multi-camera system's intrinsic and extrinsic parameters is very important in building a dataset. We use the Kalibr calibration package [49] to calibrate the intrinsic and the extrinsic parameters of our rolling shutter and global shutter cameras. The 6×6 calibration pattern used with Kalibr is shown in Figure 4.

Photometric and Vignette Calibration: We calibrated the photometric parameters, and vignette parameters for each camera independently. For each camera, we captured two calibration sequences: *(i)* a static scene with sweeping exposure for photometric calibration, and *(ii)* An image sequence with an ArUco tag [50] for vignette calibration, as shown in Figure 4. We estimated the photometric and vignette parameters using the code provided by [36].

IMU Calibration: While our dataset and evaluation focus on visual odometry, we recognize the importance of IMU data in recent works. Our dataset includes the raw IMU outputs. The intrinsic parameters and the noise analysis is generated using Intel's IMU calibration tool [51] and the IMU noise analysis tool [52], while the extrinsic parameters are extracted from the factory calibrated extrinsic values.

The sequences used to create the aforementioned camera and IMU calibrations are included in the published dataset.

Figure 4: Left to right: A sample Kalibr calibration pattern image, an image in the sweep exposure set, and a sample vignette image used for calibration.

4. Data Acquisition Protocol

Scene Selection: One of BPOD's goals is to capture a diverse set of images ranging from texture-rich to textureless, well-illuminated to dimly illuminated, illuminated by natural light or by artificial light, in simple to complex scenes, *etc.* Campus buildings, which vary in construction date and architectural style, provide an ideal setting to generate a diverse selection of imagery. We selected twelve scene trajectories from buildings on Brown University's campus and recorded forward and backward traversals of each path with both fixed and automatic exposure, for a total of four sequences per trajectory, Table 3.

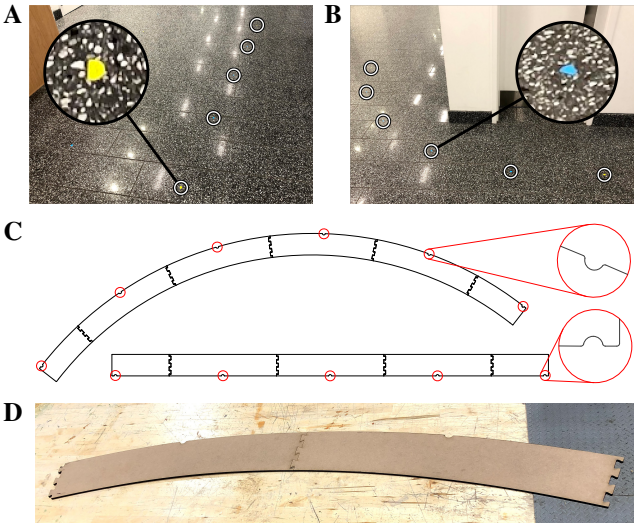

Figure 5: Laser-cut templates and color stick-on markers are used to generate ground-truth trajectories. A: Markers placed using the straight path segment, with a key marker magnified. B: Markers placed using the arc segment, with an intermediate marker magnified. C: CAD drawings of the templates used to place markers. Indentations for marker placement, which are spaced 30 inches apart, are highlighted and shown in detail views. D: Two laser-cut components for an arc template.

Defining Trajectories: We used sequences of sticker markers to define intended trajectories for our video sequences. We placed these markers at 30-inch intervals using two laser-cut templates as shown in Figure 5. The first template is a 120-inch straight template with slots for four markers. The second template is a 90-degree circular arc template with a 95.9 inch radius and slots for five markers. Our trajectories are closed loops, which allows the subject to tra-

Table 3: A summary of the BPOD dataset. The four columns represent four sequences: *(i)* forward, auto exposure; *(ii)* forward, fixed exposure; *(iii)* backward, auto exposure; *(ii)* backward, fixed exposure. Entries marked x represent unusable sequences.

Location (Abbrev.)	Sequence Length (sec.)				Trajectory Length (m)			
Applied Math Building (APMA)	86	86	84	84	68.4	68.4	68.4	68.3
85 Waterman St. (BERT)	102	96	106	100	78.4	78.4	78.4	78.4
Brown Design Workshop (BDW)	109	113	110	109	125.2	125.2	125.2	125.2
CIC Office Balcony (CIC)	122	141	140	145	125.4	125.4	125.4	125.4
Center of Information Technology (CIT 2nd Floor)	75	73	80	80	78.2	77.5	78.3	78.3
Center of Information Technology (CIT 4th Floor)	91	91	96	94	84.1	84.1	84.1	84.1
Lobby of Engineering Research Center (ERC)	109	110	x	x	108.8	108.8	x	x
Lobby of Friedman Hall (Friedman)	95	97	99	98	95.7	95.7	95.7	95.7
Lobby of MacMillan Hall (MacMillian)	95	90	96	93	86.2	86.2	86.2	86.2
Science Library (SciLiTables)	68	67	69	74	89.7	89.7	89.7	89.7
Science Library (SciLiShutters)	84	82	100	84	72.3	72.3	72.3	72.3
Smith-Buonanno Hall (SmithBuonanno)	98	92	106	96	105.7	105.7	105.7	105.7

verse them multiple times in succession, usually 2-3 times.

Mapping Trajectories: We used fixed pairwise distance measurements between sequential and co-visible markers in conjunction with trilateration and gradient-descent-based optimization to construct two-dimensional trajectory maps. We partitioned the markers into key markers and intermediate markers. Key markers used a distinct sticker color and were placed at the ends of templates, while intermediate markers filled the templates' middle slots. From each key marker, we recorded distance measurements to all other non-occluded key markers using a hand-held Bosch Blaze Pro GLM165-40 laser distance measurement device. Typically, about half of the other key markers were visible. We then constructed the ground-truth trajectory as detailed in the supplementary material. Figure 6 shows the errors between measured distances and final optimized distances in our trajectories. These are generally below 1 cm, demonstrating that our ground truth does not suffer from drift over large distances.

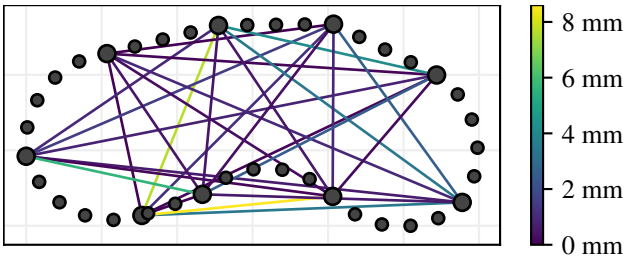

Figure 6: A representative example of post-optimization measurement errors for the SciLiShutters location's ground truth map. Grid cells are two meters across.

Extracting the data: Each sequence produces two distinct Robot Operating System (ROS) bag files, one per camera. We used a Docker container running an Ubuntu image with ROS to extract PNG images, camera calibration parameters, and IMU data along with the corresponding timestamps.

Capturing video sequences: We captured four videos for each trajectory along two axes of variation: forward vs. backward trajectory traversal and fixed vs. auto exposure. Each trajectory follows a loop that is traversed 2 to 3 times. The idea of varying exposure is to probe the performance of various methods on scenes with a very high dynamic range of illumination. For example, the sequence "4th floor of Brown" transitions between a brightly lit atrium with skylights and a dimly lit set of hallways. We noted that the default setting of the auto-exposure mode is set relatively low, mainly because it is intended to work with a laser projector. For the manually set exposure, we chose settings to balance overexposed (blown-out) regions and underexposed (completely black) regions.

The subject is captured on a third-person video which is primarily aimed at the subject's feet but has some of the background as well. The subject is localized with respect to the location on the trajectory on each video frame.

Why Markers? The generation of ground-truth trajectory is the most difficult challenge in creating a dataset. The use of the more complex external trackers [36] or a collection of video camera limits the scope of navigation. Our use of a single video camera to capture the ground-truth location with respect to a set of easily installed fiducials is a compromise of temporal accuracy for flexibility, generality and diversity of navigation paths. This temporal loss in accuracy can be justified considering the goal is to measure drift over long sequences.

Image Position Annotation: Our third-person smartphone videos allow us to identify the subject's location with respect to the marker-defined map at each point in time. We annotated the times at which the subject crosses individual markers and synchronized the resulting timestamps times with the two cameras' video streams. To synchronize the stereo cameras with the third-person captures used to generate our ground truth, we recorded the snap of a hand from all cameras at the beginning of each sequence. This technique is similar in principle to the use of clapperboards for synchronizing video and audio in filmmaking. We measured this technique to be accurate within one frame (1/30 sec-

ond), which we believe is sufficiently accurate given our emphasis on measuring long-term drift. The cameras further provide microsecond-scale timestamps for each frame and IMU data point; a comparison with the snap timestamps suggests that these are aligned.

Verification of the ground truth: We used COLMAP [53] to verify our ground-truth trajectories. COLMAP has been previously been used to generate ground-truth trajectories for datasets in Table 1, *e.g.* [39, 38]. Unfortunately, due to our dataset's erratic and drastic pedestrian movement, neither COLMAP nor Pix4D can reliably reconstruct complete sequences (see the supplementary material for more details). Nevertheless, in four sequences, COLMAP suceeded after downsampling by a factor of 10 to 3 frames per second, Figure 7. For these four sequences, our ground-truth trajectories can be compared to the ground-level projections of COLMAP generated trajectories, showing remarkable similarity, about 0.12m point-to-point average distance when the length of the trajectories is on the order of 95m. In fact a detailed observation of the trajectories makes it clear that our ground-truth tracks are more consistent with the second and third traversal of the loop than COLMAP's. The COLMAP reconstructions are also noisy, while the BPOD ground-truth is smoother and more structurally consistent with true trajectories.

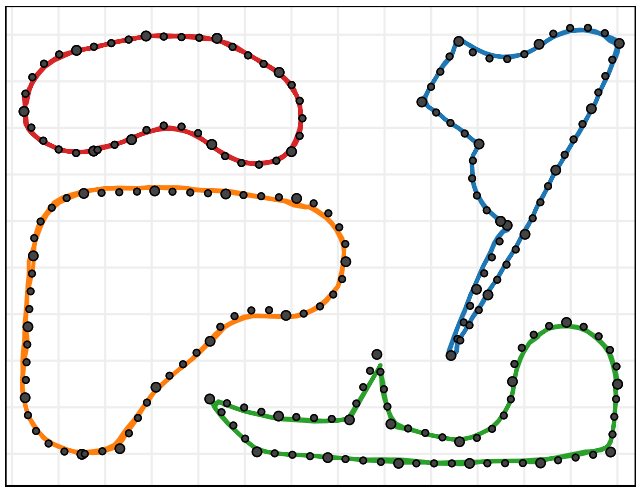

Figure 7: A comparison between our ground-truth scheme (circles) and COLMAP trajectories (lines) for a selection of sequences where COLMAP is successful. Grid cells are two meters across. Note that these sequences are plotted together for compactness although they are from separate real-world locations.

5. Experiments

In this section, we present the performance of three canonical categories of visual odometry algorithms on BPOD. We select a representative algorithm for each category: ORB-SLAM3 [54] for feature-based approaches, DSO [55] for direct approaches, and TrianFlow [45] (trained on the TUM-RGBD dataset) for deep learning approaches. We use the authors' original implementations of their respective odometry algorithms for all experiments. Note that both TrianFlow and DSO are restricted to monocular inputs, so that we compare these with the monocular mode of ORB-SLAM.

The evaluation of monocular odometry must address the inherent scale ambiguity in the resulting reconstructions: metric ambiguity in reconstruction and pose implies that scale as well as translation and rotation need to be matched optimally to compare two trajectories, *i.e.* two trajectories must by aligned under a similarity transformation. Note that the ground truth of BPOD is in 2D while the computed odometry paths are in 3D, though fairly planar. However, there are small pseudo-sinusoidal depth variations that correspond to the bobbing head movements inherent in pedestrian motion. Figure 8 compares the BPOD ground-truth trajectory to the reconstructed paths for each of the three approaches under the optimal similarity alignment using Horn's Method. Qualitatively, ORB-SLAM performs well on some sequences, but loses tracks on the others; DSO results suffer from quite severe scale drifting; and TrianFlow performs poorly. More results, including quantification using the standard absolute trajectory error (ATE) metric, are shown in the supplementary material. A quantitative characterization of these results is shown in Table 4. We use three quantitative measurements to evaluate the performance of the three methods. **Endpoint Distances:** A global evaluation metric d_e measures the distance between the ground-truth and the estimated trajectories' endpoints. Observe that only three of twelve locations are completed by ORB-SLAM, while DSO completes over 85% of the locations and TrianFlow completes all; the endpoint error, d_e, however, is quite high for DSO and TrianFlow. A closer look at the reconstructed path reveals that scale drift is a significant issue, especially for DSO. This implies that a single scaling of the reconstruction path will lead to a large endpoint difference d_e, even though the shape is correctly estimated.

We also propose two local metrics to decouple the error caused by scale drift from the error in reconstructing the shape of the path which works specifically for our 2D ground truth. Figure 9 illustrates the idea: consider the samples along each path. Aligning the first two samples adjusts the local scale of the estimated path to be that of the ground truth. Then the local difference of the third point can be decomposed into two parts: **the angular difference between the bearing vectors,** θ, reveals the performance on local rotational estimation and **the difference between the actual positions** d_l, reveals the local performance on scale estima-

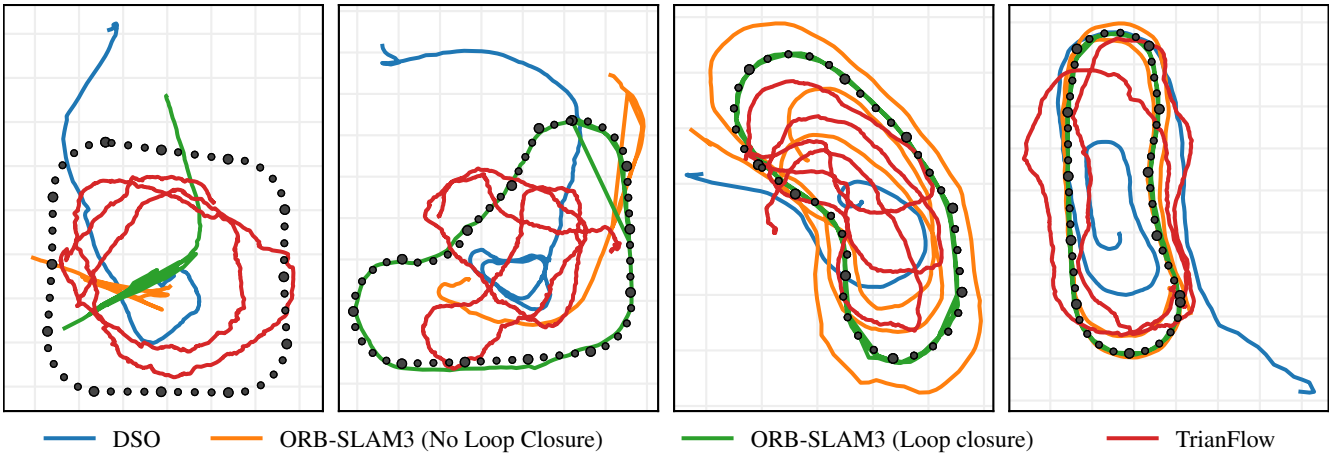

Figure 8: A comparison of estimated camera trajectories for four sample sequences, with ground truth markers shown in black. Grid cells are two meters across, and sequences are aligned using Horn's Method.

tion. Note that these metrics are similar in principle to relative pose error (RPE) used by [39, 35], while being adapted to ground-truth data without rotational information.

In addition to the metrics presented, we also calculate absolute trajectory error (ATE) on our monocular image sequences. Adapting the conventions of [39] to translation-only trajectories, we calculate the ATE between ground-truth trajectories $\mathbf{q}_{1:N}$ in world coordinate frame W and odometric trajectories $\mathbf{p}_{1:N}$ in world coordinate frame W'. Since our ground-truth trajectories are two-dimensional, we define $\mathbf{q}_{q:N}$ by concatenating each point value with $z = 0$. To calculate ATE, we use Horn's Method [56] to find a transformation $\mathbf{S}$ that optimally aligns (via translation, rotation, and scaling) $\mathbf{p}_{1:N}$ and $\mathbf{q}_{1:N}$. We then compute the root mean square of the resulting pose errors $\mathbf{f}_i = \mathbf{S}\mathbf{p}_i - \mathbf{q}_i$, $i \in \{1, \cdots, N\}$, weighted by the time Δt_i between poses i and $i-1$.

$$\mathrm{ATE}(\mathbf{f}_{1:N}) = \left(\frac{1}{T} \sum_{i=1}^{N} \Delta t_i ||\mathbf{f}||^2 \right)^{1/2}$$

We match the trajectory point pairs $(\mathbf{p}_i, \mathbf{q}_i)$ using a temporal threshold of 0.1 seconds. Because our non-interpolated ground truth trajectories are temporally sparse (with values appearing roughly every 0.5 seconds), non-deterministic behavior in the odometry algorithms we tested can impact the number of matches and the quality of the results. To address this problem, we ran each sequence several times and report average ATE values across exposure settings (auto vs. fixed) and walking directions (forward vs. backward). Note that because we omit failed runs with fewer than 20 matches ($N < 20$) along the trajectories from our results, the actual number of trials over which we average each result is in practice lower. Our ATE values are shown in Table 5.

The conclusions include *(i)* ORB-SLAM is the most accurate in reconstructing a trajectory, but it frequently fails to do so; this is due to fast rotations that cause motion blur and the subsequent loss of feature tracking. *(ii)* DSO frequently completes the trajectory but suffers from scale drift despite accurate calibration and use of global shutter cameras; and for some sequences, DSO recovers the direction of the local motion with a high accuracy, *i.e., lower local angular error*; *(iii)* TrianFlow completes all trajectories but is the least accurate.

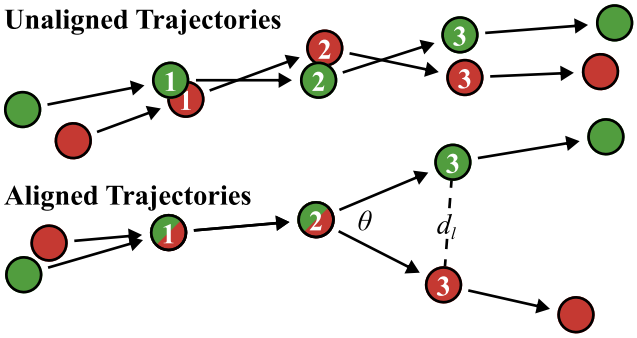

Figure 9: An illustration of our local error metrics.

Our experiments all used global shutter images, as the odometry results are generally better with these. Figure 10 compares ORB-SLAM3's estimated trajectories for global and rolling shutter images. This significant difference highlights the need for explicit models for rolling shutter cameras [57] and the usefulness of BPOD to evaluate the effectiveness of such explicit models.

Table 4: Local error metric results. X indicates a lost track; — indicates that the estimated trajectory is too short for analysis.

Locations	ORB-SLAM w/o LC			ORB-SLAM w/ LC			DSO			TrianFlow		
	d_e(m)	$\theta(^\circ/s)$	$d_l(m/s)$	d_e(m)	$\theta(^\circ/s)$	$d_l(m/s)$	d_e(m)	$\theta(^\circ/s)$	$d_l(m/s)$	d_e(m)	$\theta(^\circ/s)$	$d_l(m/s)$
APMA	X	1.44	**0.09**	X	**1.15**	0.09	**3.73**	4.13	0.27	4.20	8.80	0.43
BDW	X	—	—	X	—	—	7.62	**3.23**	**0.22**	**6.48**	6.27	0.42
BERT	X	0.83	**0.12**	X	**0.52**	**0.12**	4.25	14.2	0.69	7.29	7.75	0.53
CIC Balcony	X	**6.3**	0.63	X	7.00	1.99	**4.33**	8.64	**0.11**	5.21	8.22	0.67
CIT 2nd Floor	X	—	—	X	—	—	**7.02**	**3.88**	0.68	8.86	5.87	**0.40**
CIT 4th Floor	6.50	4.15	0.18	**0.05**	**1.07**	0.22	7.79	1.78	**0.21**	7.82	3.75	0.30
ERC	X	6.02	**0.18**	X	5.96	**0.18**	**6.95**	**3.00**	0.48	8.73	5.62	0.41
Friedman Hall	X	**4.55**	**0.23**	X	5.94	**0.23**	X	**4.53**	0.45	9.92	10.77	0.54
MacMillan	X	—	—	X	—	—	X	**5.57**	**0.57**	**9.87**	8.18	**0.57**
SciLiShutters	3.74	0.72	0.11	**0.15**	**0.54**	**0.06**	5.90	3.57	0.50	4.22	5.10	0.56
SciLiTables	1.39	0.40	0.06	**0.09**	**0.41**	**0.06**	5.34	0.80	0.14	5.45	2.73	0.41
SmithBuonanno	3.23	10.22	0.62	**0.15**	5.99	0.24	6.27	9.79	0.332	5.70	11.5	0.44
Mean	3.82	3.65	0.24	**0.11**	**2.85**	**0.36**	5.92	5.26	0.39	6.97	7.05	0.48

Table 5: Absolute trajectory error (ATE) results. Each cell contains a mean value along with a standard deviation and number of successful ($N \geq 20$) trials in parentheses. We run ORB-SLAM3 and DSO for a total of 5 trials per sequence and TrianFlow for a total of 2 trials per sequence to account for variance introduced by the sparsity of their outputs, then average across all trials (with varying exposures and path directions) at each location.

Location	DSO	ORB-SLAM3 (Loop Closure)	ORB-SLAM3 (No Loop Closure)	TrianFlow
APMA	4.129 (0.844, 14)	**2.150** (2.158, 10)	4.897 (2.162, 7)	3.829 (0.275, 8)
BDW	4.787 (0.633, 20)	**0.154** (0.016, 16)	1.110 (1.338, 12)	4.574 (0.527, 4)
BERT	3.250 (1.010, 12)	3.718 (0.626, 6)	3.522 (1.061, 5)	**3.009** (0.839, 8)
CIC_Balcony	4.461 (0.152, 20)	**0.745** (1.112, 19)	0.990 (1.178, 15)	4.330 (0.264, 7)
CIT_2nd_Floor	**4.242** (0.898, 19)	4.620 (1.322, 23)	4.807 (0.961, 18)	4.896 (0.876, 8)
CIT_4th_Floor	4.248 (0.642, 15)	**0.642** (0.848, 19)	1.066 (1.114, 18)	4.430 (0.629, 6)
ERC	6.621 (0.384, 10)	**0.581** (0.846, 10)	1.247 (1.814, 10)	3.233 (0.942, 4)
Friedman_Hall	2.792 (1.546, 20)	0.889 (0.877, 5)	**0.380** (0.289, 4)	4.260 (0.956, 8)
MacMillan	5.240 (1.343, 20)	2.645 (2.673, 9)	**1.573** (1.666, 6)	3.413 (0.861, 8)
SciLiShutters	3.679 (0.299, 20)	**0.118** (0.022, 20)	0.776 (0.600, 20)	2.900 (0.599, 8)
SciLiTables	2.766 (0.911, 20)	**0.120** (0.012, 19)	0.633 (0.929, 19)	3.361 (1.114, 8)
SmithBuonanno	4.336 (0.711, 14)	**0.722** (0.528, 12)	3.136 (1.612, 12)	4.887 (0.104, 4)

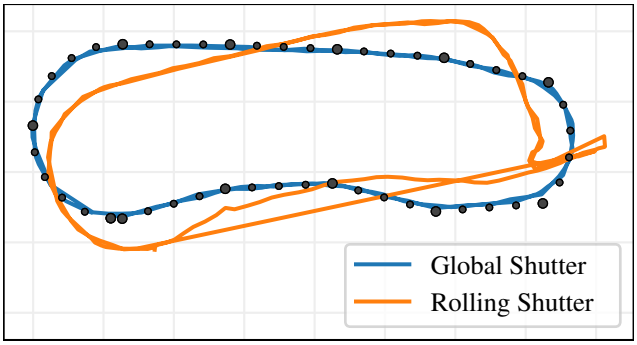

Figure 10: A comparison of ORB-SLAM3's performance on equivalent global and rolling shutter sequences in the SciLiTables location, with ground-truth markers shown. Grid cells are two meters across. The rolling shutter path's perceived misalignment occurs due to errors in the Z axis.

6. Conclusion

We present a novel ground-level pedestrian trajectory dataset based on 12 locations that cover a diverse set of scenes and illumination conditions. The dataset contains calibrated synchronized stereo data from both rolling shutter and global shutter cameras, enabling a comparison of the two for future algorithm development, and is also supplemented with IMU data. Tests on three representative feature-based, direct, and deep learning methods reveal that the dataset contains challenges not previously addressed by existing datasets. This reveals that pedestrian odometry is beyond the reach of the state-of-the-art algorithms. We hope that the BPOD dataset will facilitate the identification of failure modes of each algorithm and motivate the development of more reliable and accurate odometry methods to address the unique challenges of head-mounted pedestrian odometry.

References

[1] David Nistér, Oleg Naroditsky, and James Bergen. Visual odometry. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004.*, volume 1, pages I–I. IEEE, 2004.

[2] Davide Scaramuzza and Friedrich Fraundorfer. Visual odometry [tutorial]. *IEEE robotics & automation magazine*, 18(4):80–92, 2011.

[3] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017.

[4] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. DTAM: Dense tracking and mapping in real-time. In *2011 international conference on computer vision*, pages 2320–2327. IEEE, 2011.

[5] Konstantine Tsotsos, Alessandro Chiuso, and Stefano Soatto. Robust inference for visual-inertial sensor fusion. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5203–5210. IEEE, 2015.

[6] Davide Scaramuzza and Zichao Zhang. Visual-inertial odometry of aerial robots. *arXiv preprint arXiv:1906.03289*, 2019.

[7] Michael Blösch, Stephan Weiss, Davide Scaramuzza, and Roland Siegwart. Vision based mav navigation in unknown and unstructured environments. In *2010 IEEE International Conference on Robotics and Automation*, pages 21–28. IEEE, 2010.

[8] Igor Cvišić, Josip Ćesić, Ivan Marković, and Ivan Petrović. SOFT-SLAM: Computationally efficient stereo visual simultaneous localization and mapping for autonomous unmanned aerial vehicles. *Journal of field robotics*, 35(4):578–595, 2018.

[9] Henning Lategahn, Andreas Geiger, and Bernd Kitt. Visual SLAM for autonomous ground vehicles. In *2011 IEEE International Conference on Robotics and Automation*, pages 1732–1737. IEEE, 2011.

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[11] Denis Chekhlov, Andrew P Gee, Andrew Calway, and Walterio Mayol-Cuevas. Ninja on a plane: Automatic discovery of physical planes for augmented reality using visual SLAM. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 153–156. IEEE, 2007.

[12] Samer Karam, George Vosselman, Michael Peter, Siavash Hosseinyalamdary, and Ville Lehtola. Design, calibration, and evaluation of a backpack indoor mobile mapping system. *Remote sensing*, 11(8):905, 2019.

[13] Xuesong Shi, Dongjiang Li, Pengpeng Zhao, Qinbin Tian, Yuxin Tian, Qiwei Long, Chunhao Zhu, Jingwei Song, Fei Qiao, Le Song, et al. Are we ready for service robots? the openloris-scene datasets for lifelong SLAM. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3139–3145. IEEE, 2020.

[14] Tim Bailey and Hugh Durrant-Whyte. Simultaneous localization and mapping (SLAM): Part ii. *IEEE Robotics & Automation Magazine*, 13(3):108–117, 2006.

[15] Huthaifa Obeidat, Wafa Shuaieb, Omar Obeidat, and Raed Abd-Alhameed. A review of indoor localization techniques and wireless technologies. *Wireless Personal Communications*, 119(1):289–327, 2021.

[16] Carl Fischer and Hans Gellersen. Location and navigation support for emergency responders: A survey. *IEEE Pervasive Computing*, 9(01):38–47, 2010.

[17] Panos Kourouthanassis, Leda Koukara, Chris Lazaris, and Kostas Thiveos. Last-mile supply chain management: Mygrocer innovative business and technology framework. In *the Proceedings of the 17th International Logistics Congress: Strategies and Applications, Thessaloniki, Greece*, pages 264–273, 2001.

[18] Kostas Fouskas, George Giaglis, Panos Kourouthanassis, Adamantia Pateli, and Argiris Tsamakos. On the potential use of mobile positioning technologies in indoor environments. *BLED 2002 Proceedings*, page 33, 2002.

[19] Yingying Wang, Hu Cheng, Chaoqun Wang, and Max Q-H Meng. Pose-invariant inertial odometry for pedestrian localization. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021.

[20] Jorge Joo-Nagata, José Rafael García-Bermejo Giner, and Fernando Martínez-Abad. Mobile pedestrian navigation and augmented reality in the virtualization of the territory: Cities of salamanca and santiago de chile. In *Information Technology Trends for*

a Global and Interdisciplinary Research Community, pages 268–301. IGI Global, 2021.

[21] Gaurav Pandey, James R McBride, and Ryan M Eustice. Ford campus vision and lidar data set. *The International Journal of Robotics Research*, 30(13):1543–1552, 2011.

[22] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.

[23] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[24] Amado Antonini, Winter Guerra, Varun Murali, Thomas Sayre-McCord, and Sertac Karaman. The blackbird dataset: A large-scale dataset for uav perception in aggressive flight. In *International Symposium on Experimental Robotics*, pages 130–139. Springer, 2018.

[25] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, 2016.

[26] José-Luis Blanco-Claraco, Francisco-Angel Moreno-Duenas, and Javier González-Jiménez. The málaga urban dataset: High-rate stereo and lidar in a realistic urban scenario. *The International Journal of Robotics Research*, 33(2):207–214, 2014.

[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. *The International Journal of Robotics Research*, 38(6):642–657, 2019.

[28] Nicholas Carlevaris-Bianco, Arash K Ushani, and Ryan M Eustice. University of michigan north campus long-term vision and lidar dataset. *The International Journal of Robotics Research*, 35(9):1023–1035, 2016.

[29] András L Majdik, Charles Till, and Davide Scaramuzza. The Zurich urban micro aerial vehicle dataset. *The International Journal of Robotics Research*, 36(3):269–273, 2017.

[30] Bernd Pfrommer, Nitin Sanket, Kostas Daniilidis, and Jonas Cleveland. Penncosyvio: A challenging visual inertial odometry benchmark. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3847–3854. IEEE, 2017.

[31] David Schubert, Thore Goll, Nikolaus Demmel, Vladyslav Usenko, Jörg Stückler, and Daniel Cremers. The TUM VI benchmark for evaluating visual-inertial odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1680–1687. IEEE, 2018.

[32] Santiago Cortés, Arno Solin, Esa Rahtu, and Juho Kannala. Advio: An authentic dataset for visual-inertial odometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 419–434, 2018.

[33] Mike Kasper, Steve McGuire, and Christoffer Heckman. A benchmark for visual-inertial odometry systems employing onboard illumination. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5256–5263. IEEE, 2019.

[34] Frank Neuhaus, Tilman Koß, Robert Kohnen, and Dietrich Paulus. MC2SLAM: Real-time inertial lidar odometry using two-scan motion compensation. In *German Conference on Pattern Recognition*, pages 60–72. Springer, 2018.

[35] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[36] Jakob Engel, Vladyslav Usenko, and Daniel Cremers. A photometrically calibrated benchmark for monocular visual odometry. *arXiv preprint arXiv:1607.02555*, 2016.

[37] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *2014 IEEE international conference on Robotics and automation (ICRA)*, pages 1524–1531. IEEE, 2014.

[38] David Zuñiga-Noël, Alberto Jaenal, Ruben Gomez-Ojeda, and Javier Gonzalez-Jimenez. The UMA-VI dataset: Visual–inertial odometry in low-textured and dynamic illumination environments. *The International Journal of Robotics Research*, 39(9):1052–1060, 2020.

[39] Bernd Pfrommer, Nitin Sanket, Kostas Daniilidis, and Jonas Cleveland. Penncosyvio: A challenging visual inertial odometry benchmark. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3847–3854. IEEE, 2017.

[40] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[41] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017.

[42] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2043–2050. IEEE, 2017.

[43] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 7286–7291. IEEE, 2018.

[44] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1281–1292, 2020.

[45] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9161, 2020.

[46] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[47] András L Majdik, Charles Till, and Davide Scaramuzza. The zurich urban micro aerial vehicle dataset. *The International Journal of Robotics Research*, 36(3):269–273, 2017.

[48] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10, 2017.

[49] Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286, 2013.

[50] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[51] rs-imu-calibration tool. `https://github.com/IntelRealSense/librealsense/`.

[52] IMU Utils. `https://github.com/gaowenliang/imu_utils`.

[53] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[54] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D Tardós. ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM. *IEEE Transactions on Robotics*, 2021.

[55] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017.

[56] Berthold K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. 1988.

[57] Cenek Albl, Zuzana Kukelova, Viktor Larsson, and Tomas Pajdla. Rolling shutter camera absolute pose. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1439–1452, 2019.