Preprints.org

Article

Not peer-reviewed version

InCrowd-VI: A Realistic Visual-Inertial Dataset for Evaluating SLAM in Indoor Pedestrian-Rich Spaces for Human Navigation

Marziyeh Bamdad * , Hans-Peter Hutter , Alireza Darvishy

Posted Date: 25 November 2024

doi: 10.20944/preprints202411.1813.v1

Keywords: visual SLAM; blind and visually impaired navigation; crowded indoor environments; dataset

Article

InCrowd-VI: A Realistic Visual-Inertial Dataset for Evaluating SLAM in Indoor Pedestrian-Rich Spaces for Human Navigation

Marziyeh Bamdad [1,2,*], Hans-Peter Hutter [1] and Alireza Darvishy [1]

1 Institute of Applied Information Technology, Zurich University of Applied Sciences; marziyeh.bamdad@uzh.ch
2 Department of Informatics, University of Zurich;
* Correspondence: marziyeh.bamdad@uzh.ch

Academic Editor:

Abstract: Simultaneous localization and mapping (SLAM) techniques can be used to navigate the visually impaired, but the development of robust SLAM solutions for crowded spaces is limited by the lack of realistic datasets. To address this, we introduce InCrowd-VI, a novel visual-inertial dataset specifically designed for human navigation in indoor pedestrian-rich environments. Recorded using Meta Aria Project glasses, it captures realistic scenarios without environmental control. InCrowd-VI features 58 sequences totaling a 5 km trajectory length and 1.5 hours of recording time, including RGB, stereo images, and IMU measurements. The dataset captures important challenges such as pedestrian occlusions, varying crowd densities, complex layouts, and lighting changes. Ground-truth trajectories, accurate to approximately 2 cm, are provided in the dataset, originating from the Meta Aria project machine perception SLAM service. In addition, a semi-dense 3D point cloud of scenes is provided for each sequence. The evaluation of state-of-the-art visual odometry (VO) and SLAM algorithms on InCrowd-VI revealed severe performance limitations in these realistic scenarios, demonstrating the need and value of the new dataset to advance SLAM research for visually impaired navigation in complex indoor environments.

Keywords: visual SLAM; blind and visually impaired navigation; crowded indoor environments; dataset

1. Introduction

Navigation in crowded indoor public spaces presents major challenges for blind and visually impaired (BVI) individuals. Systems that support this navigation require real-time user localization, detailed environmental mapping, and enhanced spatial awareness. Robust solutions are necessary to cope with unfamiliar settings and provide safe and more independent mobility for people with visual disabilities. Simultaneous localization and mapping (SLAM) [1] offers promising capabilities for addressing these requirements. However, several hurdles must be overcome in order to make SLAM viable for visually impaired navigation, particularly in crowded public spaces. These settings are characterized by unpredictable pedestrian movements, varying lighting conditions, and reflective and transparent surfaces. Such dynamic and complex environments complicate reliable navigation significantly.

Although existing SLAM research has made significant advancements in domains such as robotics [2,3], autonomous driving [4], and aerial vehicles [5,6], these approaches do not adequately address the specific challenges of pedestrian-rich indoor navigation for the visually impaired.

The lack of realistic datasets tailored for human navigation in crowded environments has been a significant barrier in the development of robust SLAM systems tailored for visually impaired navigation. Current datasets often focus on vehicle-based scenarios or controlled environments and lack the diversity of scenes, dynamic complexity, and real-world conditions necessary for this application.

 doi:10.20944/preprints202411.1813.v1

To address this gap, we introduce InCrowd-VI[1], a visual-inertial dataset specifically designed for SLAM research in human-crowded indoor environments. Unlike existing datasets, InCrowd-VI captures sequences recorded at diverse indoor public locations, such as airports, train stations, museums, university labs, shopping malls, and libraries, representing realistic human motion patterns at typical walking speeds. The recorded sequences feature diverse settings, including varying crowd densities (from pedestrian-rich to static environments) and complex architectural layouts such as wide-open spaces, narrow corridors, (moving) ramps, staircases, and escalators. They present various challenges characteristic of real-world indoor spaces, including frequent occlusions by pedestrians, variations in lighting conditions, and presence of highly reflective surfaces. The dataset was collected with Meta Aria Project glasses worn by a walking person in pedestrian-rich environments, and thus incorporates realistic human motion, behavior, and interaction patterns.

The dataset has a total trajectory length of 4998.17 meters, with a total recording time of 1 h, 26 min, and 37 s. This dataset provides RGB images, stereo images, and IMU measurements. In addition, it includes a semi-dense 3D point cloud of scenes for further analysis. The ground-truth trajectories were provided by the Meta Aria project machine perception SLAM service [7], which offers a reliable benchmark for evaluating the accuracy of the SLAM algorithms.

To demonstrate the value of InCrowd-VI, several state-of-the-art classical and deep learning-based approaches for visual odometry (VO) and SLAM systems were evaluated. The analysis revealed the severe performance degradation of these systems in crowded scenarios, large-scale environments, and challenging light conditions, highlighting the key challenges and opportunities for future research to develop more robust SLAM solutions for visually impaired navigation.

The contributions of this paper are as follows.

- Introduction of InCrowd-VI, a novel visual-inertial dataset specifically designed for human navigation in indoor pedestrian-rich environments, filling a critical gap in existing research resources.
- Provision of ground-truth data, including accurate trajectories (approximately 2 cm accuracy) and semi-dense 3D point clouds for each sequence, enabling rigorous evaluation of SLAM algorithms.
- Evaluation of state-of-the-art visual odometry and SLAM algorithms using InCrowd-VI, revealing their limitations in realistic crowded scenarios.
- Identification of crucial areas for improvement in SLAM systems designed for visually impaired navigation in complex indoor environments.

2. Related Work

Evaluation of visual SLAM systems requires comprehensive datasets that capture the complexity and variability of real-world environments. The existing benchmark datasets for SLAM can be categorized according to their operational domains. Depending on the domain, different sensory data and varying degrees of ground truth accuracy have been provided [8]. Various datasets have been proposed for different operational domains, each with distinct sensor platforms, settings, and challenges. This section reviews state-of-the-art datasets, highlighting their characteristics and limitations in comparison with the newly proposed InCrowd-VI dataset. Datasets from robotics and autonomous systems as well as those focused on pedestrian odometry, are examined to assess their applicability to BVI navigation challenges. Table 1 provides an overview of these datasets, summarizing their key features and comparing them to InCrowd-VI.

[1] The InCrowd-VI dataset, along with custom tools to extract and process data from Meta .vrs recording files, will be made available online upon publication.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 November 2024 doi:10.20944/preprints202411.1813.v1

Table 1. Comparison of representative datasets.

Dataset	Environment	Carrier	Crowd Density	Ground Truth	# Sequence
Robotics and autonomous systems					
KITTI [9]	Outdoor	Car	Low	GPS/IMU	22
EuRoC MAV [10]	Indoor	Drone	Static	Motion capture	11
PennCOSYVIO [11]	In/Outdoor	Hand-held	Low	Fiducial markers	4
Zurich Urban [12]	Outdoor	Quadrotor	Not mentioned	Photogrammetric 3D reconstruction	2 km
InteriorNet [13]	Indoor	Simulated cameras	None	Synthetic	15K
TUM VI [14]	In/Outdoor	Handheld	Low	Partial motion capture	28
UZH-FPV [15]	In/Outdoor	Quadrotor	None	Leica Nova MS60 TotalStation	27+
Newer College [16]	Outdoor	Hand-held	None	6DOF ICP localization	4
VIODE [17]	In/Outdoor	Simulated quadrotor UAV	High	Synthetic	12
NAVER LABS [18]	Indoor	A dedicated mapping platform	Medium	LiDAR SLAM & SFM	5 datasets
ConsInv [19]	In/Outdoor	Not mentioned	Low	ORB-SLAM2	159
Hilti-Oxford [8]	In/Outdoor	Hand-held	Low	survey-grade scanner	16
CID-SIMS [20]	Indoor	Robot/Handheld	Low	GeoSLAM	22
Pedestrian odometry dataset					
Zena [21]	Outdoor	Head-mounted	High	Step estimation process using IMU	a 35-min dataset
ADVIO [22]	In/Outdoor	Hand-held	High	IMU-based + manual position fixes	23
BPOD [23]	In/Outdoor	Head-mounted	Not mentioned	Marker-based	48
InCrowd-VI (ours)	Indoor	Head-worn	High	Meta Aria Project SLAM service	58

2.1. *Robotics and Autonomous Systems*

Datasets play a crucial role in advancing SLAM research in various domains. The KITTI dataset [9], which is pivotal for autonomous vehicle research, has limited applicability to indoor pedestrian navigation, because it focuses on outdoor environments. Similarly, datasets like EuRoC MAV [10], TUM VI [14], and Zurich Urban [12], although offering high-quality visual-inertial data, do not fully capture the challenges of indoor pedestrian navigation in crowded environments. The CID-SIMS dataset [20], recorded from a ground-wheeled robot, provides IMU and wheel odometer data with semantic annotations and an accurate ground truth. However, it lacks the complexity of pedestrian-rich environments. VIODE [17] offers a synthetic dataset from simulated UAV navigation in dynamic environments; however, synthetic data cannot fully replace real-world data, particularly for safety-critical applications [24,25]. The ConsInv dataset [19] evaluates the SLAM systems with controlled

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 November 2024 doi:10.20944/preprints202411.1813.v1

dynamic elements. However, its controlled nature fails to capture the complexity of human-crowded public spaces. Moreover, its ground-truth method using ORB-SLAM2 with masked dynamic objects does not accurately represent the challenges faced by SLAM systems in crowded real-world settings.

2.2. Pedestrian Odometry Dataset

The Zena dataset [21] provides laser scan data and IMU measurements from helmet and waist-mounted sensors to support research on human localization and mapping in outdoor scenarios. However, the lack of visual data limits its utility in the assessment of visual SLAM systems. The Brown Pedestrian Odometry Dataset (BPOD) [23] provides real-world data from head-mounted stereo cameras in diverse indoor and outdoor environments. Although it captures challenges, such as rapid head movement and image blur, BPOD does not specifically focus on crowded indoor environments. ADVIO [22] is a dataset for benchmarking visual-inertial odometry using smartphone sensors in various indoor and outdoor paths. Although valuable for general visual-inertial odometry, the use of handheld devices limits its ability to capture natural head movements and gait patterns, which are crucial for realistic navigation scenarios. Additionally, ADVIO's ground truth, which is based on inertial navigation with manual position fixes, may have accuracy limitations.

Although these datasets provide valuable resources for evaluating SLAM systems, they do not fully address the specific challenges for human navigation in indoor pedestrian-rich environments. The InCrowd-VI dataset fills this gap by capturing realistic human motion patterns and complex environmental conditions, which are essential for developing robust SLAM solutions for visually impaired navigations.

3. InCrowd-VI Dataset

This section presents the InCrowd-VI dataset, specifically developed for evaluating SLAM in indoor pedestrian-rich environments for human navigation. The sensor framework used for data collection is first described, followed by an outline of the methodology employed to create the dataset. The process of obtaining and validating the ground-truth data is then explained. Finally, the captured sequences and the various challenges they represent are detailed.

3.1. Sensor Framework

The choice of platform for data collection was determined on the basis of the intended application of the SLAM system to be evaluated using the dataset. In the context of visually impaired navigation, wearable platforms are particularly appropriate because they can effectively capture human motion patterns during movements. Consequently, we employed a head-worn platform to collect the data. Head-mounted devices offer the advantage of capturing a forward-facing view, which is crucial for navigation, and more accurately representing how a visually impaired individual might scan their environment, including natural head movements and areas of focus. In this study, we utilized Meta Aria glasses as our sensor platform.

The Meta Aria glasses feature five cameras, including two mono scene cameras with less overlapping and large field of view, one RGB camera, and two eye-tracking (ET) cameras[2]. Additionally, the glasses are equipped with several non-visual sensors, including two Inertial Measurement Units (IMUs), a magnetometer, a barometer, a GPS receiver, and both Wi-Fi and Bluetooth beacons. The glasses also include a seven-channel spatial microphone array with a 48 kHz sampling rate, which can be configured to operate in stereo mode with two channels. It should be noted that the InCrowd-VI dataset includes data from only the RGB and mono cameras as well as IMU measurements. Other sensor data are not included in the dataset. One of the key features of Meta Aria glasses is their support for multiple recording profiles, allowing users to select which sensors to record with and

[2] https://facebookresearch.github.io/projectaria_tools/docs/tech_spec/hardware_spec

 doi:10.20944/preprints202411.1813.v1

Table 2. Specifications of the cameras on Meta Aria glasses [26], with the horizontal field of view (HFOV) in degrees, vertical field of view (VFOV) in degrees, instantaneous field of view (IFOV) in degrees per pixel, and maximum resolution in pixels. FPS represents the maximum frame rate. Note that InCrowd-VI only includes data from the RGB and Mono cameras.

Camera	HFOV	VFOV	IFOV	Max resolution	FPS	Shutter
Mono (x2)	150	120	0.26	640x480	30	global
RGB (x1)	110	110	0.038	2880x2880	30	rolling
ET (x2)	64	48	0.2	640x480	90	global

configure their settings accordingly. This flexibility makes these glasses particularly suited for diverse experimental conditions and requirements. Table 2 summarizes the specifications of the five cameras on Aria glasses.

3.2. Methodology

The data collection process was carefully designed to capture the diverse challenges of real-world indoor navigation. The strength of the dataset lies in its comprehensive representation of these challenges, which include frequent pedestrian occlusions, varying crowd densities, complex architectural layouts, wide open spaces, narrow corridors, (moving) ramps, staircases, escalators, texture-poor scenes, lighting variations, and highly reflective surfaces. These environments range from densely populated areas with more than 20 pedestrians per frame to empty cluttered spaces, offering a wide variety of scenarios for evaluating SLAM systems. During data collection, tactile paving (textured ground surfaces that BVI individuals can feel with their feet and cane) was followed where available, to mimic the walking patterns of visually impaired individuals and to enable later visual inspection of the accuracy of 3D point clouds, particularly in high-traffic areas. Data collection was conducted with the necessary permission from the relevant authorities, ensuring that ethical considerations were addressed.

3.3. Ground-Truth

The ground-truth trajectory and 3D reconstruction of the scene were produced using the Meta Aria machine perception SLAM service [7]. This service provides device trajectories generated by state-of-the-art VIO and SLAM systems, followed by offline postprocessing and refinement. It uses multiple sensors to improve accuracy and robustness, taking advantage of the precise knowledge of the sensor models, timing, and rigidity of Meta Aria devices. This allows for robust localization even under challenging real-world conditions such as fast motion, low or highly dynamic lighting, partial or temporary occlusion of cameras, and a wide range of static and dynamic environments.

Although the Meta Aria machine perception SLAM service achieves high accuracy, it does so under conditions that are not feasible for real-time visually impaired navigation: it operates offline with server-side processing, utilizes the full sensor suite, and performs extensive post-processing. By contrast, practical navigation systems for the visually impaired must operate in real time on resource-constrained wearable devices, provide immediate reliable feedback, and maintain robustness without the benefit of post-processing or cloud computation.

The trajectories produced by the Meta SLAM service have a typical global RMSE translation error of no more than 1.5 cm in room-scale scenarios. Additionally, Meta SLAM service provides a semi-dense point cloud of the scene, accurately reconstructing the static portion of the environment, even in highly dynamic and challenging situations [7].

The accuracy of this ground truth was further validated through manual measurements in several challenging scenarios, with results indicating a mean absolute error of approximately 2 cm, aligning with the reported accuracy of the Meta SLAM service. To validate the ground truth, a method was used that leverages the core principle of SLAM systems: the joint optimization of the 3D map and camera trajectory [27]. This approach involved identifying easily recognizable landmarks across the trajectory

 doi:10.20944/preprints202411.1813.v1

in both the real-world environment and semi-dense map. The 3D coordinates of these landmarks were recorded from the point cloud and the distances between them were calculated using the Euclidean distance formula. These calculated distances were then compared with actual measurements taken in the real world, allowing for a direct assessment of the accuracy of the map and reliability of the estimated trajectory. Figure 1 illustrates an example of the manual measurement process.

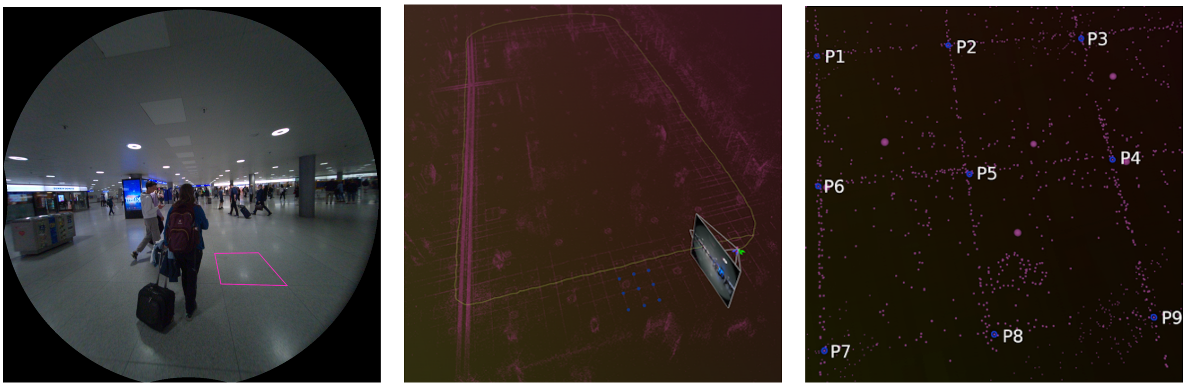


Figure 1. Sample of manual measurement process for ground-truth validation. Left: Real-world scene with a landmark floor tile highlighted by pink rectangle. Middle: Full 3D point cloud map of the scene with four adjacent floor tiles marked in blue. Right: Zoomed view of the marked corner of the tiles in the point cloud used for measurement.

Initially, manual measurements were conducted on selected crowded sequences. Following the evaluation of the state-of-the-art VO and SLAM systems presented in Section 4.2, we conducted additional manual measurements, specifically focusing on sequences where all systems exhibited failure or suboptimal performance, according to the metrics defined in Section 4.1. We used a variety of objects in each sequence, such as floor tiles, doors, advertising boards, and manhole covers, distributed across different spatial orientations throughout the trajectories to ensure robust validation. Figure 2 presents the relationship between real-world and measured distances from the second round of manual measurements. The strong linear correlation (indicated by the red trend line) and tight clustering of points around this line demonstrates that the Meta SLAM service maintains its accuracy even in challenging scenarios where contemporary SLAM systems struggle. The plot, incorporating measurements ranging from 30 cm to over 250 cm, shows that the reconstruction accuracy remained stable regardless of the measured distance, with deviations typically within 2 cm of the expected values.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 November 2024 doi:10.20944/preprints202411.1813.v1

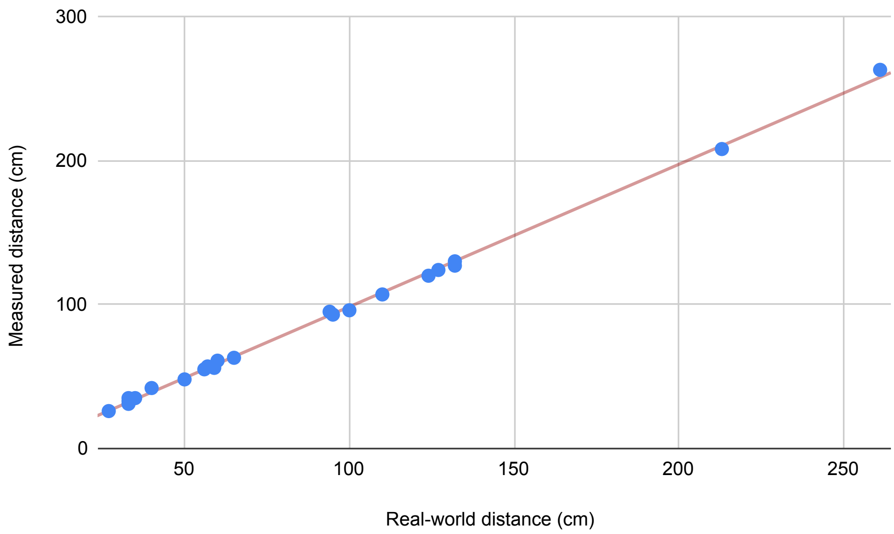


Figure 2. Correlation between real-world measurements and point-cloud-derived distances in challenging sequences, where state-of-the-art SLAM systems exhibited failure or suboptimal performance. The scatter plot demonstrates a strong linear relationship between real-world and measured distances (in centimeters), with an average error of 2.14 cm, standard deviation of 1.46 cm, and median error of 2.0 cm.

It is important to note that the manual measurement process itself introduces some level of error owing to the challenges in precisely identifying corresponding points. Despite this, the observed errors, which are slightly higher than the reported typical error, remain within a reasonable range, suggesting that the Meta SLAM service performs adequately in this specific scenario. In addition to quantitative metrics, a qualitative visual inspection of the estimated trajectories and maps was performed. This involved assessing the alignment of landmarks, plausibility of trajectories, and removal of moving pedestrians in the scene. Figure 3 illustrates the capability of the Meta Aria machine perception SLAM service to handle dynamic objects by showcasing a scenario where a pedestrian initially appears static relative to the camera while on an escalator but subsequently becomes dynamic. The image presents the refined 3D reconstruction produced by the SLAM service, which successfully identifies and removes dynamic pedestrians from the final point cloud, leaving only static elements of the environment. The dynamic object removal process improves the accuracy of 3D map reconstruction and trajectory estimation by ensuring that moving and temporary stationary objects are not incorporated into the static environment representation or ground truth.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 November 2024 doi:10.20944/preprints202411.1813.v1

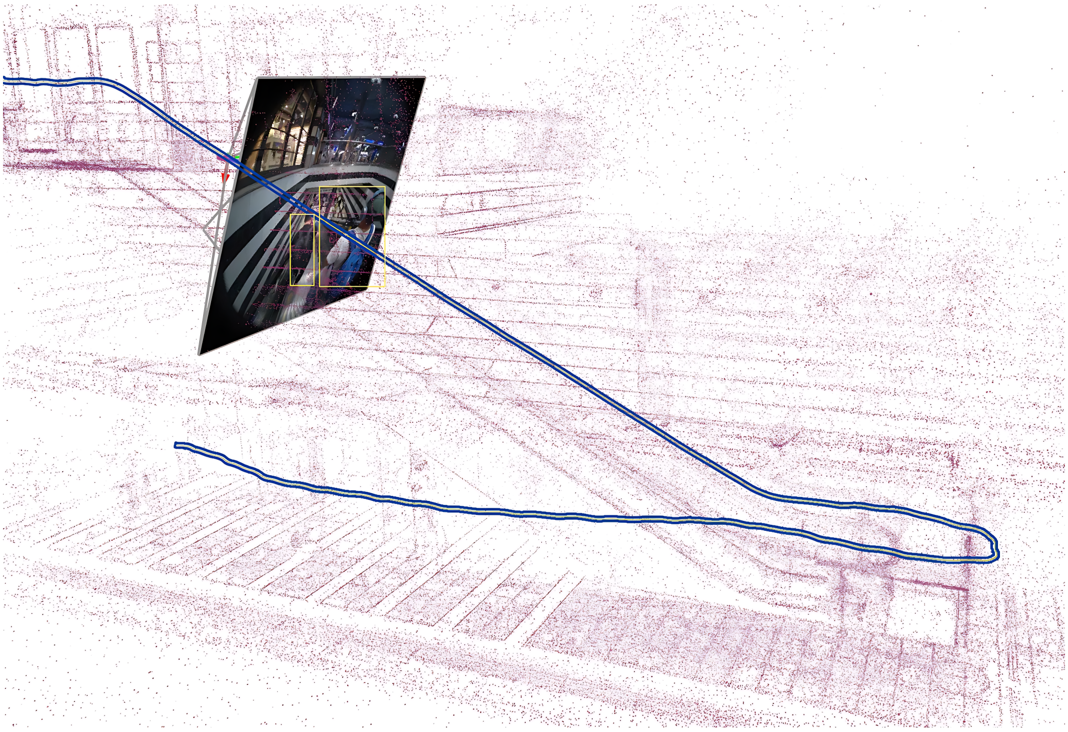


Figure 3. Refined 3D reconstruction demonstrating the removal of dynamic pedestrians that initially appeared static relative to the camera on the escalator.

3.4. Sequences and Captured Challenges

Each dataset sequence in InCrowd-VI includes timestamped RGB images at resolutions of 1408 × 1408 and 30 FPS, stereo pair images at resolutions of 640 × 480 and 30 FPS, two streams of IMU data at data rates of 1000 and 800 Hz, semi-dense point clouds, and an accurate trajectory ground truth. While the Meta Aria glasses provide data from multiple sensors, not all of them are included in the dataset, as they were not essential for the focus of this dataset. An example of the image data and their relative 3D map of the scene is shown in Figure 4.

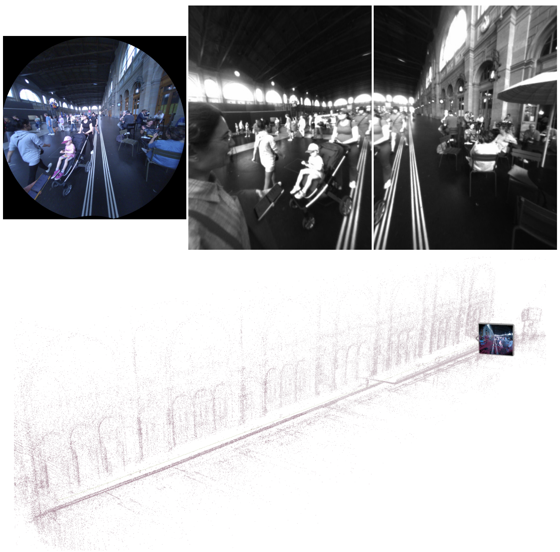


Figure 4. Example of image data and corresponding 3D map from a dataset sequence: The top-left image shows the RGB frame, and the top-middle and top-right images represent the left and right images of a stereo pair. The bottom image shows the 3D map of the scene.

 doi:10.20944/preprints202411.1813.v1

To facilitate a comprehensive evaluation of the SLAM algorithms, the dataset sequences are organized to represent different levels of pedestrian density and environmental complexity. This categorization allows for the evaluation of SLAM systems at varying levels of human density and environmental complexity. Table 3 provides an overview of the sequences, illustrating their categorization based on pedestrian density, diversity of venues, sequence lengths, durations, and the specific challenges encountered in each scenario. This structure ensures that the dataset offers a wide variety of scenarios, enabling a thorough evaluation of the adaptability and robustness of the SLAM algorithms in realistic scenarios. Additionally, trajectory length plays a crucial role in assessing the performance and robustness of visual SLAM systems. For indoor environments for visually impaired navigation, we considered sequences with lengths less than 40 m as short trajectories, those ranging from 40 m to 100 m as medium trajectories, and those of 100 m and beyond as long trajectories. Figure 5 shows a histogram of the trajectory lengths for the InCrowd-VI dataset.

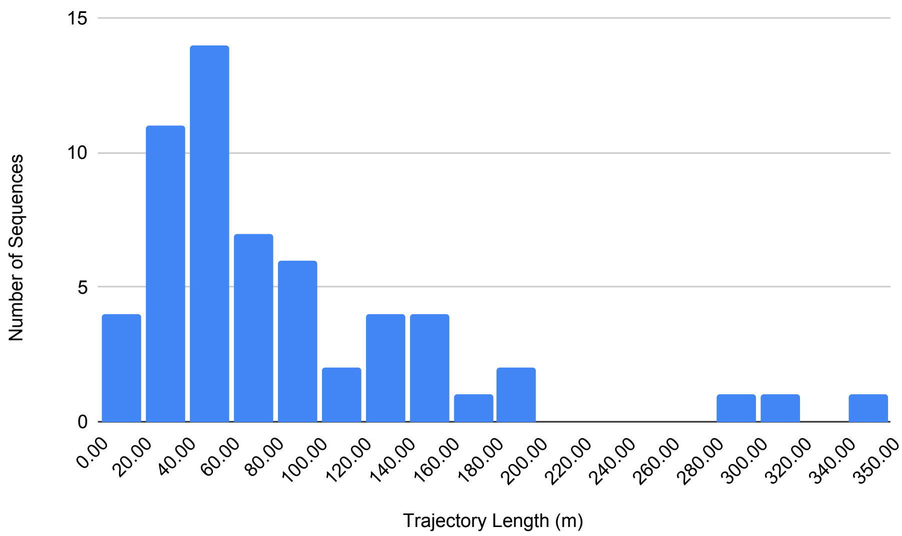


Figure 5. Histogram of trajectory length

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 November 2024 doi:10.20944/preprints202411.1813.v1

Table 3. InCrowd-VI Sequences.

Density	Sequence name	Venue	Length (m)	Span (mm:ss)	Main Challenges
High	Oerlikon_G7	Train station	133.36	02:04	Challenging light, ramp
	UZH_stairs	Museum	13.16	00:41	Stairs
	G6_exit	Train station	60.21	01:27	Stairs, flickering light
	G6_loop	Train station	103.74	02:07	Challenging light
	Getoff_pharmacy	Train station	159.50	03:06	Motion transition
	Service_center	Airport	131.89	02:18	Reflection, open area
	Ochsner_sport	Airport	70.57	01:14	Reflection
	Toward_gates	Airport	165.46	02:57	Reflection, open area
	Arrival2	Airport	74.96	01:23	Crowd
	Checkin2_loop	Airport	348.84	05:22	Long trajectory
	Shopping_open	Shopping mall	182.51	03:10	Challenging light, open area
	Turn180	Airport	41.95	00:46	Reflection
	Cafe_exit	Main train station	121.40	01:51	Reflection, open area
	G8_cafe	Main train station	294.05	04:14	Long trajectory
	Ground_53	Main train station	32.40	01:01	Escalator (motion transition)
	Kiko_loop	Shopping mall	314.68	05:50	Long trajectory
	Reservation_office	Main train station	95.15	01:28	Reflection, open area
	Orell_fussli	Shopping mall	57.98	01:16	Crowd
	Reservation_G17	Main train station	97.30	01:50	Challenging light, open area
	Short_Loop_BH	Main train station	104.94	02:29	Reflection, open area

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 November 2024 doi:10.20944/preprints202411.1813.v1

Table 3. *Cont.*

Density	Sequence name	Venue	Length (m)	Span (mm:ss)	Main Challenges
Medium	Migros	Shopping mall	84.18	01:40	Reflection, open area
	TS_exam_loop	Exam location	28.81	01:34	Glass wall, sitting people
	Museum_loop	Museum	66.99	02:02	Challenging light
	Ramp_checkin2	Airport	191.77	03:40	Moving ramp (motion transition), open area
	Airport_shop	Shopping mall	99.49	01:46	Reflection, open area
	Towards_checkin1	Airport	55.89	01:02	Reflection, open area, glass wall
	Entrance_checkin1	Airport	35.70	00:39	Reflection
	Shopping_loop	Shopping mall	151.91	02:36	Challenging light
	Towards_circle	Airport	127.46	02:06	Challenging light, repetitive structure
Low	AND_floor51	Library	40.23	00:53	Narrow aisle
	AND_floor52	Library	39.13	01:10	Narrow aisle, sitting people
	AND_liftAC	University building	71.15	01:33	Open area, glass wall
	ETH_HG	University building	99.00	01:56	Repetitive structure
	ETH_lab	Laboratory	56.70	01:24	Reflection
	Kriegsstr_pedestrian	Public building	31.97	00:59	Texture-poor
	Kriegsstr_same_dir	Public building	31.17	00:56	Texture-poor
	TH_entrance	University building	41.85	00:53	Challenging light, sitting people
	TS_entrance	University building	32.06	00:42	Reflection
	TS_exam	University building	44.93	01:01	Narrow corridor, texture-poor
	UZH_HG	University building	142.12	03:09	Long trajectory, repetitive structure
	Museum_1	Museum	69.30	01:51	Challenging light
	Museum_dinosaur	Museum	44.62	01:16	Challenging light
	Museum_up	Museum	12.20	00:34	Stairs
	Short_loop	Airport	36.92	00:46	Open area

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 November 2024 doi:10.20944/preprints202411.1813.v1

Table 3. *Cont.*

Density	Sequence name	Venue	Length (m)	Span (mm:ss)	Main Challenges
None	AND_Lib	Office building	52.19	01:18	Reflection
	Hrsaal1B01	Academic building	74.55	01:50	Challenging light
	ETH_FT2	Museum	40.97	00:56	Challenging light, reflection
	ETH_FTE	Museum	59.00	01:35	Challenging light, open area
	ETH_lab2	Laboratory	58.08	01:26	Texture-poor, reflection
	Habsburgstr_dark	Public building	36.08	01:05	Multi-floor, dimly lit
	Habsburgstr_light	Public building	87.99	02:46	Multi-floor
	IMS_lab	Laboratory	15.23	00:43	Cluttered scene
	IMS_TE21	Laboratory	42.43	01:23	Cluttered scene, challenging light
	IMS_LEA	Laboratory	19.35	00:43	Cluttered scene
	Kriegsstr	Public building	31.18	00:54	Texture-poor
	TH_loop	Office building	154.74	03:55	Reflection
	TS116	University building	59.11	01:24	Challenging light, reflection, glass wall
	TS_stairs	University building	27.67	00:52	Stairs, challenging light, glass wall

It is important to note that the walking speeds in our dataset reflect the typical navigation patterns of visually impaired individuals. Studies have shown that individuals with visual impairments tend to walk slower and have reduced stride lengths during both independent and guided walking compared to sighted people [28]. While the typical walking speed for sighted populations is generally between 1.11 and 1.4 m/s [28], the average walking speed in the InCrowd-VI dataset is 0.75 m/s. This lower average speed aligns with the expectations of the visually impaired navigation.

The dataset captures a wide range of challenges inherent to real-world indoor navigation scenarios, including the following.

- Dynamic obstacles: InCrowd-VI features sequences with moving pedestrians, capturing scenarios of crossing paths with groups and maneuvering around individuals moving in different directions. These sequences test the ability of the SLAM systems to handle unpredictable dynamic elements in real-world environments.
- Crowd density variation: Sequences capture a range of crowd densities from static to densely populated areas, testing the adaptability of SLAM systems to different levels of human activity.
- Frequent occlusions: The dataset includes sequences with frequent occlusions caused by moving pedestrians, luggage, and infrastructure, thereby creating significant challenges for maintaining accurate mapping and tracking.

- Reflective and transparent surfaces: The dataset includes scenes with glass and other reflective surfaces that can distort sensor readings and complicate the visual SLAM algorithms.
- Texture-poor areas: Scenes with minimal visual features, such as plain walls, challenge feature-based SLAM systems.
- Large-scale and complex environments: The dataset covers diverse architectural layouts, including open spaces, corridors, ramps, staircases, and escalators, to test the adaptability of SLAM to various spatial configurations.
- Lighting variations: Sequences incorporate sequences with varying lighting conditions, from well-lit atriums to dimly lit corridors or areas with flickering lights, to test the SLAM robustness under varying illumination conditions.
- Sudden viewpoint changes: Sequences capture user perspective shifts during corner turns and level transitions, thereby challenging SLAM tracking consistency.
- Motion transitions: Sequences include transitions between moving environments (escalators, moving ramps, and trains) and stationary areas, to test SLAM's ability to distinguish ego-motion from environmental motion.

These challenges collectively contribute to the realism and complexity of the InCrowd-VI dataset, making it a valuable resource for evaluating and advancing SLAM systems designed for visually impaired navigation in real-world indoor environments. Figure 6 shows the selection of images from the InCrowd-VI dataset.

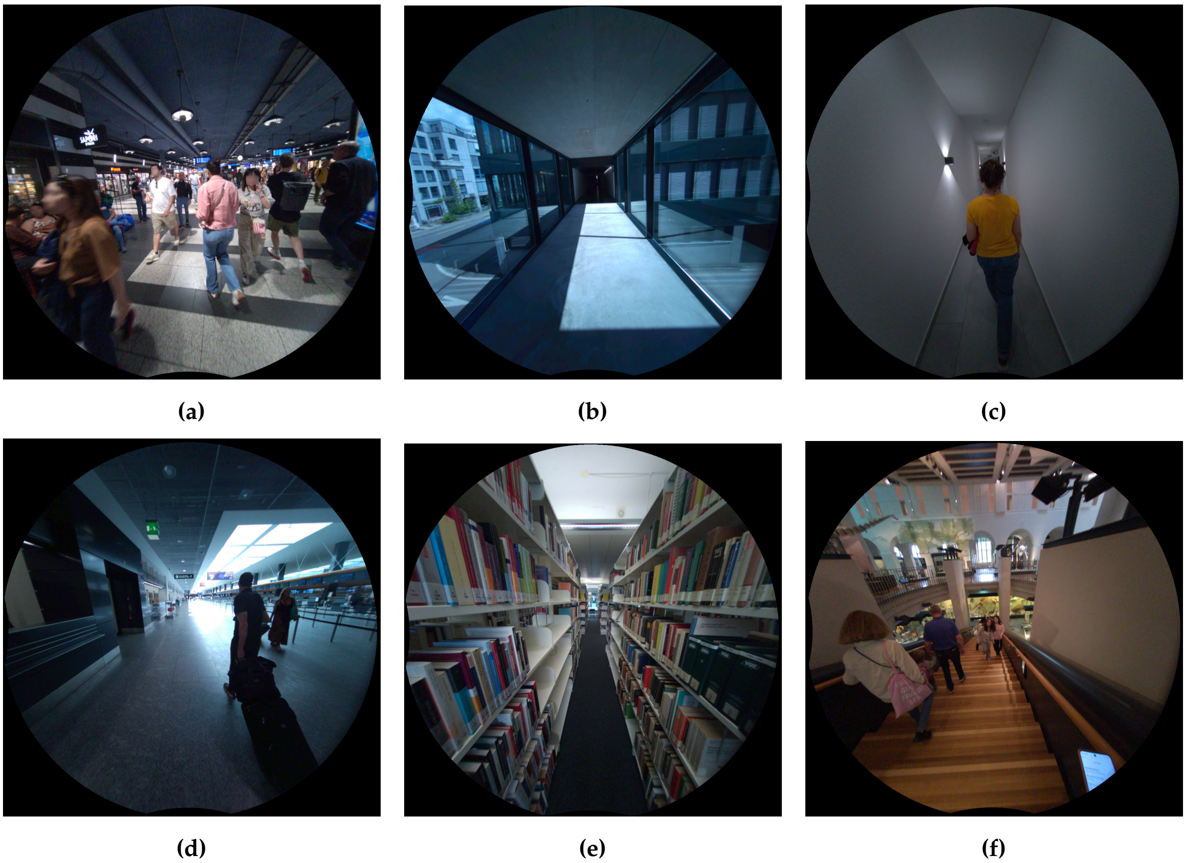


Figure 6. Example scenes from the InCrowd-VI dataset demonstrating various challenges: (a) high pedestrian density, (b) varying lighting conditions, (c) texture-poor surfaces, (d) reflective surfaces, (e) narrow aisles, (f) stairs.

4. Experimental Evaluation

To assess the accuracy and robustness of current state-of-the-art visual SLAM systems on the InCrowd-VI dataset, four representative algorithms were selected: two classical approaches and two

 doi:10.20944/preprints202411.1813.v1

Table 4. Characteristics of selected SLAM systems

SLAM system	Matching approach	Approach
DROID-SLAM	Dense optical flow-based	End-to-end
DPV-SLAM	Sparse patch-based	End-to-end
ORB-SLAM3	Feature-based	Classical SLAM
SVO	Semi-direct	Classical VO

deep learning-based methods. These systems were chosen because of their prominence in the field and diverse approaches to visual SLAM. As shown in Table 4, DROID-SLAM [29] and DPV-SLAM [30] use dense optical flow-based and sparse patch-based matching approaches, respectively, both of which leverage end-to-end learning. Conversely, ORB-SLAM3 [31] and SVO [32] employ the classical feature matching and semi-direct matching approaches, respectively. This evaluation serves to demonstrate that InCrowd-VI effectively captures challenging real-world scenarios that current state-of-the-art systems struggle to handle and underscores the dataset's value as a benchmark for advancing SLAM research in visually impaired navigation.

A selection of sequences from InCrowd-VI representing a range of difficulty levels from easy to hard was used for evaluation. These sequences were selected to represent the diverse challenges present in the dataset, including changing elevations, long trajectories, motion transitions, challenging lighting conditions, and crowds with different densities. This approach allows for a comprehensive assessment of the accuracy and robustness of SLAM systems across various scenarios, while maintaining a manageable evaluation process. ORB-SLAM3 was evaluated using the left camera images and left IMU measurements. The DROID-SLAM, DPV-SLAM, and SVO were evaluated using only the left camera images. Additionally, the intrinsic camera parameters were extracted, and the left IMU-to-left camera transformation was calculated using the calibration data provided by the Meta Aria project. For camera calibration, Project Aria used a sophisticated camera model (FisheyeRadTanThinPrism) [3], which includes six radial, two tangential, and four thin-prism distortion parameters. A custom tool was developed to convert the data from the VRS files recorded by the Meta Aria glasses to the required format for each VO and SLAM system, and system configuration parameter files were provided for each system to ensure compatibility with the dataset. The experimental setup ran Ubuntu 20.04 and 22.04 on a Lenovo laptop equipped with a 12th Gen Intel® Core™ i9-12950HX vPro® Processor, 64 GB DDR5 RAM, and an NVIDIA RTX™ A2000 8 GB GDDR6 graphics card. ORB-SLAM3, SVO, and DPV-SLAM incorporate random processes into their operations, such as RANSAC for outlier rejection and random feature selection. DPV-SLAM, which is based on DPVO [33], adds further variability through random keypoint selection and its probabilistic depth estimation approach. By contrast, DROID-SLAM uses a deterministic model that produces identical results across multiple runs, as observed in our experiments. To account for these algorithmic differences and ensure fair evaluation, we executed ORB-SLAM3, SVO, and DPV-SLAM five times each, reporting mean values, whereas DROID-SLAM required only a single execution per sequence. The results are summarized in Table 5.

[3] https://facebookresearch.github.io/projectaria_tools/docs/tech_insights/camera_intrinsic_models

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 November 2024 doi:10.20944/preprints202411.1813.v1

Table 5. Quantitative results of evaluated SLAM systems across various challenging scenarios, including different crowd densities, lighting conditions, long trajectories, elevation changes, and motion transitions. The results include the absolute trajectory error (ATE) in meters, drift percentage (DP) showing system drift as a percentage of total trajectory length, pose estimation coverage (PEC) as a percentage, system frames per second (FPS) indicating the number of frames processed by the system per second, and real-time factor (RTF). RTF provides a ratio that indicates how much faster the system processes data compared to the user's walking speed. The table includes the trajectory length (Trj. Len.) in meters and the average walking speed (AWS) in meters per second for each sequence. Sequences with high crowd density are marked in red, medium in orange, low in blue, and none in black. The failed sequences are denoted as ×.

Sequence	Trj Len (m)	AWS (m/s)	Classical Systems										Deep Learning-Based Systems									
			ORB-SLAM3					SVO					DROID-SLAM					DPV-SLAM			
			ATE (m)	DP (%)	PEC (%)	FPS	RTF	ATE (m)	DP (%)	PEC (%)	FPS	RTF	ATE (m)	DP (%)	PEC (%)	FPS	RTF	ATE (m)	DP (%)	PEC (%)	FPS
Crowd Density																					
Orell_fussli	57.98	0.76	2.95	5.08	90	23	0.80	x	x	x	174	5.82	1.07	1.84	100	16	0.53	0.21	0.36	98	10
Entrance_checkin1	35.70	0.92	0.77	2.15	92	25	0.82	6.78	18.99	80	128	4.27	0.08	0.22	100	16	0.53	0.08	0.22	97	11
Short_loop	36.92	0.80	0.22	0.59	96	24	0.83	3.71	10.04	70	134	4.51	0.04	0.10	100	15	0.51	0.32	0.86	97	11
IMS_lab	15.23	0.35	0.06	0.39	94	24	0.82	2.49	16.34	71	149	5.05	0.02	0.13	100	16	0.54	0.04	0.26	97	14
Lighting Variations																					
Reservation_G17	97.30	0.88	7.45	7.65	95	26	0.87	15.23	15.65	74	156	5.25	1.14	1.17	100	19	0.65	1.95	2.00	99	12
Toward_circle	127.46	1.01	x	x	97	25	0.86	x	x	x	175	5.86	0.88	0.69	100	15	0.50	0.59	0.46	99	10
Museum_1	69.30	0.62	0.34	0.49	98	25	0.87	3.82	5.51	36	171	5.75	6.21	8.96	100	18	0.61	8.06	11.63	99	10
TS116	59.11	0.70	x	x	57	21	0.71	3.99	6.75	31	159	5.32	3.77	6.37	100	15	0.52	1.08	1.82	98	10
Long Trajectory																					
Kiko_loop	314.68	0.90	22.81	7.24	98	26	0.87	20.12	6.39	59	153	5.11	2.51	0.79	100	17	0.56	26.79	8.51	99	8
Shopping_loop	151.91	0.97	3.02	1.98	97	25	0.85	21.41	14.09	89	138	4.63	2.52	1.65	100	16	0.55	12.21	8.03	99	9
UZH_HG	142.12	0.75	30.60	21.53	94	26	0.88	16.11	11.33	47	170	5.70	2.86	2.01	100	17	0.58	11.45	8.05	99	9
TH_loop	154.74	0.66	x	x	88	22	0.72	0.01	0.006	0.9	131	4.37	4.53	2.92	100	18	0.62	6.71	4.33	99	9
Changing Elevation (stairs) - Short Trajectories																					
UZH_stairs	13.16	0.32	0.06	0.45	77	23	0.78	1.52	11.55	37	138	4.62	0.04	0.30	100	22	0.78	0.05	0.37	97	15
G6_exit	60.21	0.69	6.67	11.07	97	26	0.86	x	x	x	153	5.11	0.41	0.68	100	17	0.57	1.91	3.17	97	12
Museum_up	12.20	0.36	0.42	3.44	82	22	0.75	0.03	0.24	6.2	164	5.44	0.03	0.24	100	18	0.61	0.02	0.16	98	12
TS_stairs	27.67	0.53	0.26	0.93	94	25	0.84	2.43	8.78	41	158	5.30	0.21	0.75	100	16	0.56	0.41	1.48	96	12
Motion Transition																					
Getoff_pharmacy	159.50	0.86	81.01	50.78	85	25	0.83	34.91	21.88	48	153	5.10	27.95	17.52	100	17	0.59	27.55	17.27	98	10
Ground_53	32.40	0.53	x	x	96	25	0.84	4.54	14.01	79	136	4.54	3.30	10.18	100	19	0.64	3.84	11.85	98	13
Ramp_checkin2	191.77	0.87	7.31	3.81	98	26	0.87	46.95	24.48	97	142	4.74	1.82	0.94	100	20	0.67	0.77	0.40	99	10

 doi:10.20944/preprints202411.1813.v1

4.1. Evaluation Metrics

To evaluate VO and SLAM systems, we used the root mean square error (RMSE) of absolute trajectory error (ATE) [34]. This metric quantifies the accuracy of the estimated trajectory compared with the ground truth by measuring the root mean square of the differences between the estimated and true poses. For ORB-SLAM3 and SVO, the absolute trajectory error was calculated using the TUM RGB-D benchmark tool [35] and RPG trajectory evaluation tool [36], respectively. ORB-SLAM3 outputs trajectories in TUM format, which is directly compatible with the TUM RGB-D benchmark tool, whereas SVO includes the RPG evaluation tool in its package, making it the most suitable choice for evaluating SVO's output. For DROID-SLAM and DPV-SLAM, custom scripts were developed to compute the same metrics using the EVO package [37], a flexible tool capable of handling their output formats.

Trajectory accuracy alone does not fully capture the performance of a SLAM system, particularly in challenging real-world scenarios. SLAM systems are susceptible to initialization failures and tracking loss, particularly under conditions such as motion blur, lack of visual features, or occlusions. Such disruptions lead to gaps in the estimated trajectory, thus affecting the overall reliability of the system. To address this, we introduced the pose estimation coverage (PEC) metric, calculated as (number of estimated poses/total number of frames) × 100. The PEC quantifies the system's ability to maintain consistent pose estimation throughout a sequence, offering insights into its robustness against initialization failures and tracking losses. A notably low PEC often indicates critical failure, which can make the ATE unreliable. To further evaluate each system, we introduced a drift percentage (DP) metric, defined as (ATE/trajectory length) × 100. This metric quantifies the drift of the system as a percentage of the total distance traveled. A lower value indicates better performance, with the system maintaining more accurate localization over extended trajectories. To ensure a comprehensive evaluation that considers both accuracy and robustness, we adopted a dual-criteria approach that incorporates both ATE and PEC. A sequence is considered successful if the drift percentage (DP) value is less than 1% of the distance traveled and its PEC exceeds 90%.

Beyond accuracy and robustness, real-time performance is crucial for visually impaired navigation applications because it directly influences user experience and safety. To directly relate system performance to the end-user experience, we developed a metric that compares the processing speed of SLAM systems with the user's walking speed. The fundamental idea is that, if the system can process more distance than the user walks in a given time interval (e.g., one second), it is considered to be performing in real time. First, we calculated the frames per second (FPS) processed by the system, indicating the number of frames handled per second. Next, we calculated the distance per frame (DPF), which represents the distance between consecutive camera frames. DPF was calculated using the formula DPF = AWS /camera_fps, where AWS is the average user walking speed and camera_fps is the camera's frame rate. Finally, we introduced the processed distance rate (PDR), which is obtained by multiplying the FPS and DPF. If the PDR meets or exceeds the average walking speed, the system is considered capable of keeping pace with the user's movement, indicating adequate real-time performance for BVI navigation. We quantified this capability through the real-time factor (RTF), defined as the ratio of PDR to AWS, where the values ≥ 1 demonstrate the real-time processing capacity.

4.2. Evaluation Results

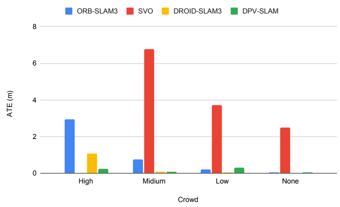
The evaluation results presented in Table 5 demonstrate that InCrowd-VI successfully captures challenging scenarios that push the limits of the current state-of-the-art VO and SLAM systems. Our analysis of these results is two-fold. First, we assessed the system performance against the key requirements of visually impaired navigation applications. Second, we analyzed the impact of specific environmental factors, such as crowd density, trajectory length, lighting conditions, and motion transition, on system performance. This analysis validates the effectiveness of InCrowd-VI as a benchmark, while also identifying critical areas that require further research to make SLAM viable for visually impaired navigation.

To effectively address the needs of BVI individuals during navigation, several key requirements must be satisfied, including real-time performance, high localization accuracy, and robustness. In addition, the system should maintain long-distance consistency with minimal drifts. For this study, we established a localization accuracy criterion of 0.5 meters for indoor environments, which we consider suitable for BVI navigation applications. Across the sequences, ORB-SLAM3 and DROID-SLAM generally met or approached this criterion in low-density crowds. However, under more challenging conditions, both classical and learning-based systems exhibit a higher ATE, which significantly surpasses the required accuracy. Although not directly comparable to the requirement for robustness to occlusions, dynamic objects, and other challenging conditions, pose estimation coverage (PEC) provides insights into system reliability. Robust pose estimation is essential for maintaining consistent localization. BVI navigation requires minimal interruption during the operation. Deep learning-based approaches consistently achieved high PEC values ($> 90\%$) across all sequences, whereas classical approaches failed to maintain consistent tracking, particularly in scenes with lighting variations. This indicates a limitation in handling the dynamics in the environment, which is critical for real-world BVI applications.

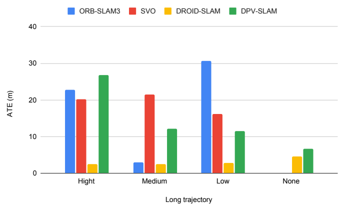
Real-time performance requires the system to match or exceed the walking speed of the user. Our processed distance rate (PDR) metric, when compared to the average walking speed (AWS), indicates that classical approaches generally achieve real-time or near real-time performance across almost all sequences. However, deep learning-based methods fall short of real-time processing, indicating delays that could hinder real-time feedback for users. Regarding long-distance consistency, the drift percentage (DP) metric revealed that both classical and deep learning-based approaches frequently exceeded the desired 1% drift threshold, particularly for longer trajectories. Classical approaches typically show higher drift percentages (often exceeding 5-10% of the trajectory length), whereas deep learning-based methods generally maintain lower drift rates but still struggle to consistently achieve the desired threshold. It is important to note that, although some systems may approach or meet certain requirements under controlled conditions, they all face significant challenges in maintaining consistent performance across diverse real-world scenarios represented in the InCrowd-VI dataset. This highlights the gap between the current VO and SLAM capabilities and requirements for visually impaired navigation in crowded indoor environments.

Beyond these requirements, our analysis reveals how specific environmental factors affect the system performance. The evaluation results highlight the significant impact of crowd density on the performance of VO and SLAM systems. In scenarios with high crowd density, systems often exhibit a higher ATE than sequences with low or no crowd density. The trajectory length also plays a crucial role. For instance, in the Entrance_checkin1 scene, despite the medium crowd density, the system reported a lower error owing to the short trajectory length. In addition to crowd density, other challenging conditions also affect system performance. The TS116 scene, which is a typical university building environment with artificial lighting and sunlight in some areas, exemplifies the impact of lighting variations. Despite the absence of pedestrians, ORB-SLAM3 encountered difficulties during initialization and experienced frequent tracking losses owing to lighting challenges. Similarly, in the TH_loop scene, the user's proximity to a wall causing occlusions, combined with challenging lighting conditions from large windows, caused tracking issues, even without the presence of pedestrians.

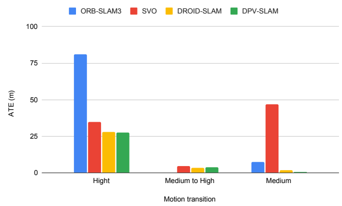
The results also demonstrated the influence of specific scenario characteristics. In scenes involving stairs, the short trajectory lengths allow the systems to handle elevation changes effectively, despite the inherent challenges. However, in scenes with motion transitions, particularly Getoff_pharmacy, where the user transitions from a moving train to a stationary platform, the systems struggle to differentiate between ego-motion and environmental motion, resulting in a poor performance. Figure 7 provides a visual representation of the impact of these challenging conditions on ATE across different systems. The charts reveal that deep learning-based methods generally maintain a lower ATE across most challenging conditions, whereas classical approaches tend to show more dramatic accuracy degradation with increasing environmental complexity.



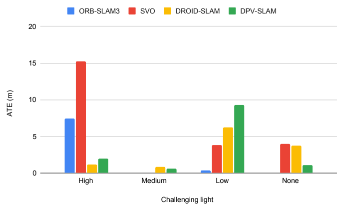
(a) Crowd



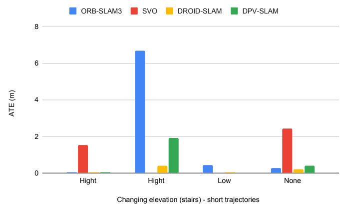
(b) Long Trajectory



(c) Motion Transition



(d) Challenging Light



(e) Changing Elevation (Stairs)

Figure 7. ATE comparison of evaluated SLAM systems under challenging conditions, with the x-axis depicting sequences categorized by crowd density: high, medium, low, and none.

In summary, our evaluation revealed three key findings: (1) deep-learning-based methods demonstrate superior robustness but struggle with real-time processing, whereas classical approaches offer better processing speed but lack consistency in challenging conditions; (2) environmental factors, such as crowd density and lighting variations, significantly impact all systems, with performance degradation increasing in proportion to crowd density; and (3) none of the evaluated systems fully satisfied the combined requirements for reliable BVI navigation in terms of accuracy, robustness, and real-time performance in complex indoor environments. These findings underscore the need for new approaches that can better balance these competing demands while maintaining reliability across diverse real-world conditions.

5. Conclusion

This paper introduces InCrowd-VI, a novel visual-inertial dataset designed to address the challenges of SLAM in indoor pedestrian-rich environments, particularly for visually impaired navigation. Our evaluation of the state-of-the-art VO and SLAM algorithms on InCrowd-VI demonstrated significant performance limitations across both classical and deep learning approaches,validating that the dataset effectively captures challenging real-world scenarios. These systems struggle with the complexities of crowded scenes, lighting variations, and motion transitions, highlighting the gap between the current capabilities and real-world indoor navigation demands. InCrowd-VI serves as a crucial benchmark for advancing SLAM research in complex, crowded indoor settings. It provides realistic, user-centric data that closely mirrors the challenges faced by visually impaired individuals navigating such environments. Future studies should focus on addressing the key challenges identified by the InCrowd-VI dataset. Although InCrowd-VI is a valuable resource for indoor SLAM research, it is important to acknowledge its limitations. The absence of depth information restricts its applicability for testing depth-based SLAM systems, and its focus on indoor environments limits its utility in outdoor and mixed-environment scenarios.

Author Contributions: Conceptualization, M.B.; methodology, M.B.; software, M.B.; validation, M.B.; formal analysis, M.B.; investigation, M.B.; resources, M.B.; data curation, M.B.; writing—original draft preparation, M.B.; writing—review and editing, M.B., H.-P.H.; visualization, M.B.; supervision, H.-P.H. and A.D.; funding acquisition, A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by SNSF (Swiss National Science Foundation) grant number 40B2-0_194677. The APC was funded by ZHAW Open Acess, Zurich University of Applied Sciences.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The InCrowd-VI dataset along with custom tools for data extraction and processing, will be made publicly available online upon publication.

Acknowledgments: The authors would like to thank the Robotics and Perception Group at the Department of Informatics, University of Zurich, for providing the Meta Aria glasses used in this study. The authors also extend their gratitude to Giovanni Cioffi for his valuable feedback.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Durrant-Whyte, H.; Bailey, T. Simultaneous localization and mapping: part I. *IEEE robotics & automation magazine* **2006**, *13*, 99–110.
2. Cadena, C.; Carlone, L.; Carrillo, H.; Latif, Y.; Scaramuzza, D.; Neira, J.; Reid, I.; Leonard, J.J. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics* **2016**, *32*, 1309–1332.
3. Pellerito, R.; Cannici, M.; Gehrig, D.; Belhadj, J.; Dubois-Matra, O.; Casasco, M.; Scaramuzza, D. Deep Visual Odometry with Events and Frames. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS. IEEE, 2024.
4. Bresson, G.; Alsayed, Z.; Yu, L.; Glaser, S. Simultaneous localization and mapping: A survey of current trends in autonomous driving. *IEEE Transactions on Intelligent Vehicles* **2017**, *2*, 194–220.
5. Hanover, D.; Loquercio, A.; Bauersfeld, L.; Romero, A.; Penicka, R.; Song, Y.; Cioffi, G.; Kaufmann, E.; Scaramuzza, D. Autonomous drone racing: A survey. *IEEE Transactions on Robotics* **2024**.
6. Alberico, I.; Delaune, J.; Cioffi, G.; Scaramuzza, D. Structure-Invariant Range-Visual-Inertial Odometry. *arXiv preprint arXiv:2409.04633* **2024**.
7. Somasundaram, K.; Dong, J.; Tang, H.; Straub, J.; Yan, M.; Goesele, M.; Engel, J.J.; De Nardi, R.; Newcombe, R. Project Aria: A new tool for egocentric multi-modal AI research. *arXiv preprint arXiv:2308.13561* **2023**.
8. Zhang, L.; Helmberger, M.; Fu, L.F.T.; Wisth, D.; Camurri, M.; Scaramuzza, D.; Fallon, M. Hilti-oxford dataset: A millimeter-accurate benchmark for simultaneous localization and mapping. *IEEE Robotics and Automation Letters* **2022**, *8*, 408–415.
9. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012, pp. 3354–3361.
10. Burri, M.; Nikolic, J.; Gohl, P.; Schneider, T.; Rehder, J.; Omari, S.; Achtelik, M.W.; Siegwart, R. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research* **2016**, *35*, 1157–1163.
11. Pfrommer, B.; Sanket, N.; Daniilidis, K.; Cleveland, J. Penncosyvio: A challenging visual inertial odometry benchmark. 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017, pp. 3847–3854.
12. Majdik, A.L.; Till, C.; Scaramuzza, D. The Zurich urban micro aerial vehicle dataset. *The International Journal of Robotics Research* **2017**, *36*, 269–273.
13. Li, W.; Saeedi, S.; McCormac, J.; Clark, R.; Tzoumanikas, D.; Ye, Q.; Huang, Y.; Tang, R.; Leutenegger, S. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. *arXiv preprint arXiv:1809.00716* **2018**.
14. Schubert, D.; Goll, T.; Demmel, N.; Usenko, V.; Stückler, J.; Cremers, D. The TUM VI benchmark for evaluating visual-inertial odometry. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018, pp. 1680–1687.

15. Delmerico, J.; Cieslewski, T.; Rebecq, H.; Faessler, M.; Scaramuzza, D. Are we ready for autonomous drone racing? the UZH-FPV drone racing dataset. 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 6713–6719.
16. Ramezani, M.; Wang, Y.; Camurri, M.; Wisth, D.; Mattamala, M.; Fallon, M. The newer college dataset: Handheld lidar, inertial and vision with ground truth. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 4353–4360.
17. Minoda, K.; Schilling, F.; Wüest, V.; Floreano, D.; Yairi, T. Viode: A simulated dataset to address the challenges of visual-inertial odometry in dynamic environments. *IEEE Robotics and Automation Letters* **2021**, *6*, 1343–1350.
18. Lee, D.; Ryu, S.; Yeon, S.; Lee, Y.; Kim, D.; Han, C.; Cabon, Y.; Weinzaepfel, P.; Guérin, N.; Csurka, G.; others. Large-scale localization datasets in crowded indoor spaces. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3227–3236.
19. Bojko, A.; Dupont, R.; Tamaazousti, M.; Borgne, H.L. Self-improving SLAM in dynamic environments: learning when to mask. *arXiv preprint arXiv:2210.08350* **2022**.
20. Zhang, Y.; An, N.; Shi, C.; Wang, S.; Wei, H.; Zhang, P.; Meng, X.; Sun, Z.; Wang, J.; Liang, W.; others. CID-SIMS: Complex indoor dataset with semantic information and multi-sensor data from a ground wheeled robot viewpoint. *The International Journal of Robotics Research* **2023**, p. 02783649231222507.
21. Recchiuto, C.T.; Scalmato, A.; Sgorbissa, A. A dataset for human localization and mapping with wearable sensors. *Robotics and Autonomous Systems* **2017**, *97*, 136–143.
22. Cortés, S.; Solin, A.; Rahtu, E.; Kannala, J. ADVIO: An authentic dataset for visual-inertial odometry. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 419–434.
23. Charatan, D.; Fan, H.; Kimia, B. Benchmarking Pedestrian Odometry: The Brown Pedestrian Odometry Dataset (BPOD). 2022 International Conference on 3D Vision (3DV). IEEE, 2022, pp. 1–11.
24. Liu, Y.; Fu, Y.; Chen, F.; Goossens, B.; Tao, W.; Zhao, H. Simultaneous localization and mapping related datasets: A comprehensive survey. *arXiv preprint arXiv:2102.04036* **2021**.
25. Tosi, F.; Zhang, Y.; Gong, Z.; Sandström, E.; Mattoccia, S.; Oswald, M.R.; Poggi, M. How NeRFs and 3D Gaussian Splatting are Reshaping SLAM: a Survey. *arXiv preprint arXiv:2402.13255* **2024**.
26. Meta AI Research. Project Aria Hardware Specifications: https://facebookresearch.github.io/projectaria_tools/docs/tech_spec/hardware_spec, 2024. [Online; accessed 2024-09-04].
27. Schops, T.; Sattler, T.; Pollefeys, M. Bad slam: Bundle adjusted direct rgb-d slam. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 134–144.
28. Bennett, H.J.; Valenzuela, K.A.; Fleenor, K.; Morrison, S.; Haegele, J.A. Walking biomechanics and energetics of individuals with a visual impairment: a preliminary report. *Human Movement* **2019**, *20*, 8–18.
29. Teed, Z.; Deng, J. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems* **2021**, *34*, 16558–16569.
30. Lipson, L.; Teed, Z.; Deng, J. Deep Patch Visual SLAM. *arXiv preprint arXiv:2408.01654* **2024**.
31. Campos, C.; Elvira, R.; Rodríguez, J.J.G.; Montiel, J.M.; Tardós, J.D. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics* **2021**, *37*, 1874–1890.
32. Forster, C.; Zhang, Z.; Gassner, M.; Werlberger, M.; Scaramuzza, D. SVO: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics* **2016**, *33*, 249–265.
33. Teed, Z.; Lipson, L.; Deng, J. Deep patch visual odometry. *Advances in Neural Information Processing Systems* **2024**, *36*.
34. Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; Cremers, D. A benchmark for the evaluation of RGB-D SLAM systems. 2012 IEEE/RSJ international conference on intelligent robots and systems. IEEE, 2012, pp. 573–580.
35. Computer Vision Group TUM School of Computation, I.; of Munich, T.T.U. Useful tools for the RGB-D benchmark. https://edit.paperpal.com/documents/a108dfc5-1231-4f54-a402-15d7c7b5f556.
36. Zhang, Z.; Scaramuzza, D. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018, pp. 7244–7251.
37. Grupp, M. evo: Python package for the evaluation of odometry and SLAM. https://github.com/MichaelGrupp/evo, 2017.

 doi:10.20944/preprints202411.1813.v1