Benchmark for Evaluating Long-Term Localization in Indoor Environments Under Substantial Static and Dynamic Scene Changes

Niklas Trekel Tiziano Guadagnino Thomas Läbe Louis Wiesmann
Perrine Aguiar Jens Behley Cyrill Stachniss

***Abstract*— Accurate localization is crucial for the autonomous operation of mobile robots. Specifically for indoor scenarios, localization algorithms typically rely on a previously generated map. However, many real-world sites like warehouses or healthcare environments violate the underlying assumption that the robot's surroundings are mainly static. In this paper, we introduce a new dataset plus a benchmark that enables evaluating and comparing indoor localization methods in complex and changing real-world scenarios. While several datasets for indoor scenes exist, only a few combine the long-term localization aspect of repeatedly revisiting the same environment under varying conditions with precise ground truth over multiple rooms. Our dataset comprises various sequences recorded with a wheeled robot covering an office environment. We provide data from two 2D LiDARs, multiple consumer-grade RGB-D cameras, and the robot's wheel odometry. By densely placing fiducial markers on every room ceiling, we can also provide accurate pose information within a single global frame for the whole environment, estimated through an additional upward-facing camera. We evaluate existing localization algorithms on our data and make the dataset together with a server-based benchmark evaluation publicly available. This facilitates an unbiased evaluation of localization approaches and enables further research on their application in challenging indoor scenarios.**

I. Introduction

Localization is a key component for autonomously operating mobile robots deployed in indoor sites like factories, warehouses, or healthcare facilities [9], [22] to enable safe navigation and interaction with objects therein. Due to limited access to external positioning systems such as GNSS, robots typically localize themselves within prebuilt maps of the environment. However, the aforementioned environments oftentimes contain short-term changes due to moving people or objects, as well as long-term changes like rearrangements of items in the scene. These can impede the localization quality over time, potentially posing a safety risk for robot localization. Rebuilding the map when changes occur may solve this issue, but it discards all previously collected information about the environment despite large parts of the scene potentially remaining unchanged. Localization algorithms that maintain their accuracy in the presence of commonly occurring changes, therefore, offer a more efficient solution. Public datasets and benchmarks play a relevant role in investigating the performance of localization algorithms under such conditions as they allow an unbiased and reproducible evaluation. Especially when working with robots, they also both reduce the research community's effort in data collection and circumvent the work and costs required to build a hardware platform.

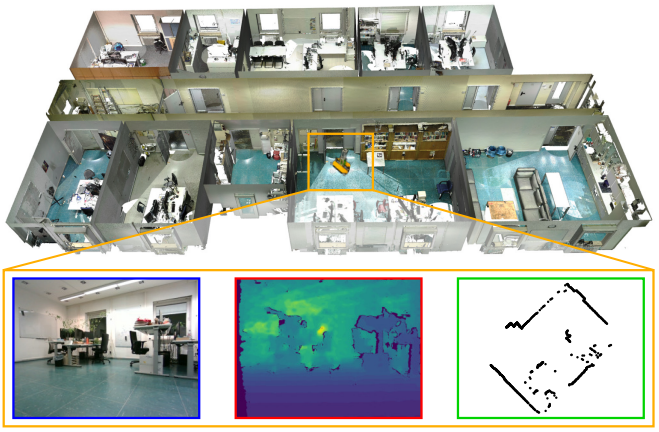


Fig. 1: A colored point cloud of the office environment where we recorded our long-term indoor localization dataset, with sample sensor data from RGB images, depth images, and 2D LiDAR scans.

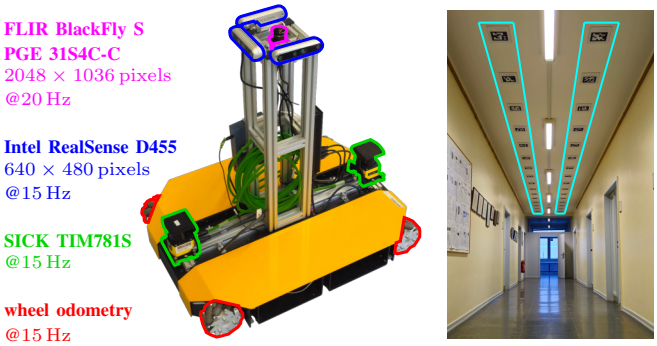
Although numerous indoor datasets with various sensor modalities exist, many aim to challenge simultaneous localization and mapping (SLAM) algorithms already during the mapping stage [18], [20], [21], [44], [45]. However, evaluating the effective usage of a previously generated map requires datasets in which the robot repeatedly traverses the same environment. In addition, generating accurate ground truth for indoor scenes with multiple rooms remains challenging. While being accurate, standard external tracking methods, such as motion capture systems or laser trackers, require a line-of-sight to the robot and, therefore, constrain the space with known pose information for evaluation of localization performance [30], [31], [35], [44].

The main contribution of this work is a new dataset plus a benchmark that extends existing work in these aspects, specifically designed for evaluating long-term indoor localization algorithms for ground robots. The dataset is recorded in an office environment with multiple rooms, as shown in Fig. 1. It contains sequences collected under various conditions, such as short-term dynamics through moving people or objects and long-term changes in the scene. We provide data from 2D LiDAR sensors, RGB-D

All authors are with the Center for Robotics, University of Bonn, Germany. Cyrill Stachniss is also with the Lamarr Institute for Machine Learning and Artificial Intelligence, Germany.

This work has partially been funded by the German Federal Ministry of Education and Research (BMBF) in the project "Robotics Institute Germany", grant No. 16ME0999 and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob.

cameras, and wheel encoders, as well as pose ground truth in a single global frame across all rooms. In sum, we contribute: (i) a dataset tailored to the evaluation of long-term localization algorithms with various challenging scenarios, including dynamic and long-term changes in the scene, that also provides accurate ground truth across the complete environment, (ii) the evaluation of existing algorithms on our dataset, providing insights into shortcomings of existing methods and thus incentives for possible future research directions, and (iii) a public benchmark that performs server-based evaluation on a test set with hidden ground truth to enable an objective comparison of submitted results. Our dataset website and a link to the benchmark challenge are available at: `https://www.ipb.uni-bonn.de/html/projects/localization_benchmark/`.



(a) Omnidirectional Clearpath Dingo robot. (b) AprilTag markers.

Fig. 2: The sensor platform used for data collection, with two 2D LiDAR scanners, three RGB-D cameras, and a fisheye camera used to detect AprilTags on the ceiling for ground truth generation.

II. Related Work

Localization for mobile robots is a well-explored research topic [38], [43]. To enable mobile robots to localize themselves in previously unseen environments using solely onboard sensors, a large body of work addresses the SLAM problem. Researchers have developed a wide range of approaches, encompassing various map representations, sensor modalities, and estimation algorithms [4], [6], [34]. Particularly in indoor settings, mobile robots typically operate within restricted environments, such that the problem can often be simplified to localization within a known map once the environment has been mapped. Monte-Carlo localization (MCL), as introduced by Dellaert et al. [10], is a widely established method for indoor localization of ground robots. Several extensions to the original MCL framework exist, including improvements to its efficiency [14] and robustness [26], [36], [37]. While MCL is most commonly used with LiDAR sensors [1], [10], [14], [23], [42], there also exist approaches utilizing cameras [2], [19], [41] and other sensor information, such as WiFi [19], [32]. Earlier studies have conducted extensive evaluations of the accuracy of existing localization methods [29], [33], underlining their achievable precision.

In this context, publicly available datasets and benchmarks avoid duplicating complete experimental setups and thus further simplify the development, testing, and comparison of newly designed methods. Therefore, public datasets and benchmarks recently gained importance for driving advances in robotics research, as demonstrated by seminal works in multiple domains [3], [16], [35]. Specifically for indoor ground vehicles, early contributions provided extensive datasets covering complete floors [7], multiple floors [12], or even different indoor sites [28]. Advances in sensor technologies stimulated a trend towards increasingly multimodal sensor setups in recent datasets, including combinations of 3D LiDARs, multiple RGB-D cameras [21], and event cameras [15], [20], [44]. Many existing datasets address the presence of degenerate conditions, e.g., dynamics and rearrangement of furniture [12], [20], [31]. Furthermore, impairments of the sensor's information are considered, e.g., varying lighting conditions [28], [44], lack of visual features [18], [21], and common issues with moving mobile platforms [20], [45], [46]. Few datasets also address collaborative localization or SLAM with multiple robots [13], [24].

Evaluation of the pose accuracy of localization and SLAM algorithms requires ground truth, i.e., time-stamped positions and orientations of the platform—if possible, obtained with high accuracy and using an orthogonal sensing modality. While motion capture systems and total stations provide accurate pose information, they require a line of sight from a fixed station to the robot. Their application to multi-room environments is hence cumbersome or restricts robot trajectories to confined spaces [18], [31], [44]. Other approaches register exteroceptive sensor data to high-fidelity scans of the environment [15], [20], [46], or to building floor plans [12], which requires to handle structural changes and dynamics in the sensor data. Furthermore, solely using the exteroceptive sensors of the robot together with existing SLAM or odometry estimation algorithms only provides pose accuracy in the order of the algorithms aimed to be evaluated [5], [28], [45].

We circumvent these shortcomings using fiducial markers instead, as depicted in Fig. 2b, similar to previous work [21], [27]. However, we place markers densely in all rooms and obtain accurate ground truth in a single global reference frame by jointly optimizing the marker positions, which we extract from highly accurate terrestrial laser scanner (TLS) data. We provide an overview of this process in Fig. 3.

Moreover, as most of the aforementioned datasets target SLAM algorithms, they aim to cover a wide variety of conditions between individual sequences and, therefore, perform recordings in distinct places. Instead, to facilitate the evaluation of localization algorithms, we follow the principle of repeatedly revisiting the same environment under changing conditions, similar to works of Shi et al. [31], Pronobis et al. [28], and Fallon et al. [12], which are the closest to our research. Our work combines several key elements: it provides sequences all recorded within the same environment under varying conditions, with a dedicated mapping run covering the whole scene for the initial map generation, comes with accurate ground truth in a single reference frame across a multi-room setting, and establishes

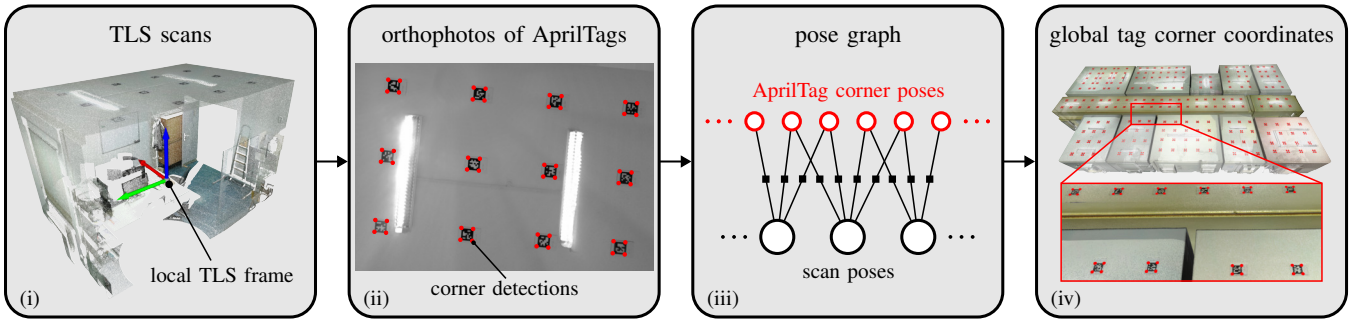


Fig. 3: Our method for obtaining all AprilTag corner coordinates in one global reference frame. (i) We collect TLS scans in all parts of the environment. (ii) We generate an orthophoto from each TLS scan capturing the ceiling-mounted AprilTags and detect their corner points. (iii) We assign poses to all corner points based on the tag orientation and jointly optimize the corner poses and the local TLS frame poses using a pose graph. (iv) The final result comprises globally consistent AprilTag corner positions on the complete floor.

a public benchmark challenge for an automatized assessment of methods on non-public ground truth data using server-side automatic evaluations. The previously mentioned datasets only partially cover the combination of these aspects. As a result, we present a new benchmark tailored explicitly to long-term localization.

III. A Long-Term Indoor Localization Dataset

We collected our dataset in an office environment consisting of multiple office rooms of different sizes and a kitchen. We provide 2D LiDAR, RGB-D, and wheel odometry data, which we, for convenience, make available both as ROS bag files and at the same time as individual data files, i.e., images in PNG format, time series in .txt files, and parameters in .yaml files.

A. Hardware Platform

Our recording platform is a Clearpath Dingo omnidirectional wheeled robot equipped with two 2D LiDAR scanners, three RGB-D cameras, and an upward-facing marker camera with a fisheye lens solely used to detect ceiling-mounted AprilTag markers for referencing, as depicted in Fig. 2. We jointly estimate and provide intrinsics and extrinsics of all exteroceptive sensors using the approach by Wiesmann et al. [40]. To estimate the transformation between these sensors and the wheel odometry's reference frame, we solve a least-squares problem using sample trajectories from the wheel odometry and the marker camera used in our ground truth system, see Fig. 2. All sensor time stamps are relative to the onboard PC. The marker camera is synchronized via the IEEE 1588 precise time protocol (PTP) [11], while the RGB-D cameras use the manufacturer's internal software synchronization mechanism.

B. Ground Truth System

Our ground truth system consists of fiducial markers with known 3D coordinates obtained from a TLS and that are densely attached to all ceilings. An upward-facing marker camera tracks these markers to estimate the robot's pose, providing ground truth poses with the camera's frequency of $20\,\mathrm{Hz}$. As shown in Fig. 2b, we mounted in total 213 AprilTags [25] of family 36h11 with a spacing between tags such that we obtain a density of approximately $1\,\mathrm{tag/m^2}$. Our method to obtain the poses of all AprilTags in one global frame is visualized in Fig. 3. First, we iteratively scan the complete floor using a Faro Focus3D-X130 TLS, which produces high-density point clouds with millimeter accuracy. We then generate orthophotos from each scanned point cloud using an orthogonal projection of the scan points along the local TLS frame's z-axis, i.e., the ceiling normal direction. The orthophotos then capture the AprilTags on the ceiling, enabling us to detect the AprilTag's corners to get their coordinates relative to the TLS's local reference frame. To combine all tag coordinates in one global reference frame, we first assign a full pose to each corner by aligning coordinate frame axes based on each tag's four corner positions. Then, we jointly optimize the TLS scan poses and all AprilTag corner poses. We, therefore, optimize a pose graph with relative poses between corners and scan poses as factors using least squares and obtain global position coordinates of the AprilTag corners as a final result. With our method, we obtain a root mean squared error (RMSE) of $4.7\,\mathrm{mm}$ over all scans between the resulting global AprilTag corner coordinates and the transformed local corner points relative to the optimized TLS scan poses.

To obtain the ground truth robot poses themselves, we detect AprilTag corners within each image captured by the upward-facing marker camera at a frequency of $20\,\mathrm{Hz}$. We can then estimate the camera's pose by solving a least-squares perspective-n-point (PnP) problem between the detected AprilTag corners and their known respective 3D coordinates from the TLS data. While the accuracy is difficult to measure over a complete floor and depends on the number of visible tags in the camera image, we compared our localization system against a commercial motion capture system available in a single room. We first measured an average accuracy of $1.2\,\mathrm{cm}$ and $1°$ for translation and rotation, respectively. The majority of the error, however, stems from the time synchronization between the robot and the motion capture system. We are able to reduce the offsets to $3\,\mathrm{mm}$ and $0.1°$ after compensating for a constant time offset between the systems. Since most rooms covered in the dataset contain fewer tags and with potential additional errors due to the TLS scan alignment, we overall expect a slightly lower accuracy than $3\,\mathrm{mm}$ and $0.1°$ for our dataset's ground truth.

TABLE I: Overview of our dataset's statistics, grouped by the characteristics of the sequences.

type	number of sequences	duration [s]	size [GB]	distance [m]
mapping	1	812.0	56.4	217.8
static	7	2857.8	198.5	1025.1
dynamics	5	1261.2	87.6	375.8
long-term changes	6	2688.0	186.7	979.0
long-term changes + dynamics	2	1206.0	83.8	437.2
overall	21	8824.9	612.9	3034.8

C. Collected Data

Our dataset consists of a total of 21 sequences that we recorded with our hardware platform. In Tab. I, we give an overview of the statistics of our collected data. While our dataset website provides a description for each individual sequence, we here classify the sequences by their characteristics as follows:

1) Mapping Sequence: To enable the generation of an initial map representation of the scene to be used by the localization system under evaluation, we provide a mapping sequence with ground truth poses in which the robot visits each room.

2) Static Conditions: To evaluate the localization performance under optimal conditions, we recorded multiple runs, preserving the state of the mapping sequence. These sequences also serve well for debugging purposes. In addition, we recorded sequences representing conditions during a regular day in the office.

3) Short-Term Dynamics: We collected data that contains varying degrees of short-term dynamics in the scene. In particular, we include multiple sequences that contain up to 10 people moving in the vicinity of the robot, a sequence where cardboard boxes are additionally pushed through the sensor's field of view, and a sequence where the robot follows a person carrying a large board.

4) Long-Term Scene Changes: We include sequences with long-term scene changes compared to the mapping sequence. The sequences differ in the severity of the changes, ranging from daily situations, e.g., closed doors that were previously open, to objects being newly added or moved. For the latter, we rearranged the furniture in some rooms, e.g., tables, chairs, and sofas, and added cardboard boxes and boards to the environment to obstruct previously visible geometry.

5) Combination of Long-Term Scene Changes and Dynamics: Finally, we combined the former aspects in sequences where people move through a scene that contains long-term changes compared to the mapping sequence.

IV. Baseline Results

The main focus of this work is to provide a benchmarking setup that enables the evaluation of localization algorithms, specifically in their long-term localization performance. We design our experiments to showcase the featured characteristics of our dataset by evaluating existing methods for LiDAR-based MCL on a representative subset of our dataset's sequences and highlighting remaining challenges in the research field.

Fig. 4: Occupancy grid map generated from the mapping sequence.

A. Experimental Setup

We evaluate algorithms on a pre-built map for localization, which we generate using the front 2D LiDAR's deskewed scans from the mapping sequence. We refine the laser scanner's poses using pose graph optimization, where the ground truth poses from the AprilTag system serve as unary factors, and we add edges between nodes with relative transformations obtained using scan matching. We thereby use the approach from Choi et al. [8] with the implementation provided in the Open3D library [47] and add edge candidates between nodes within a radius of $1\,\mathrm{m}$. These are tested for validity and filtered during the optimization stage. Fig. 4 shows an occupancy grid map generated from the laser scans. For our evaluation, we considered four different baselines. We use the AMCL ROS-package that was built based on the work of Fox [14], a MCL implementation from the Sapienza Robust Robotics Group (SRRG) [17], which we abbreviate with SRRG-Loc, LocNDF [39], and ENM-MCL [23]. While AMCL and SRRG-Loc utilize the pre-built occupancy map in Fig. 4 with a grid resolution of $5\,\mathrm{cm}$, LocNDF and ENM-MCL use a learning-based implicit map representation to query the Euclidean distance (LocNDF) or both the Euclidean and projective distance (ENM-MCL) to occupied space. We train the learning-based approaches on the same laser scans used to generate the grid map in Fig. 4. We limit the range of particles for AMCL between 1,000 and 10,000 and use a fixed particle size of 10,000 for all other algorithms. While SRRG-Loc runs in real-time using all laser beams, we use a subset of 100 beams for AMCL to achieve real-time performance with the maximum number of particles. LocNDF and ENM-MCL run offline, where LocNDF uses all beams, and ENM-MCL uses a reduced number of 128 beams due to memory limitations.

For the evaluation, we select a subset of four sequences from our data, each representing one of our dataset's different characteristics. Fig. 5 depicts the qualitative properties of the sequences.

An entirely static sequence S_s without any scene changes compared to the mapping sequence serves as a reference for the localization performance of the algorithms under optimal conditions. The other sequences contain inconsistencies with the mapping run, where sequence S_d contains dynamics in the form of moving people, sequence S_{lc} includes long-term

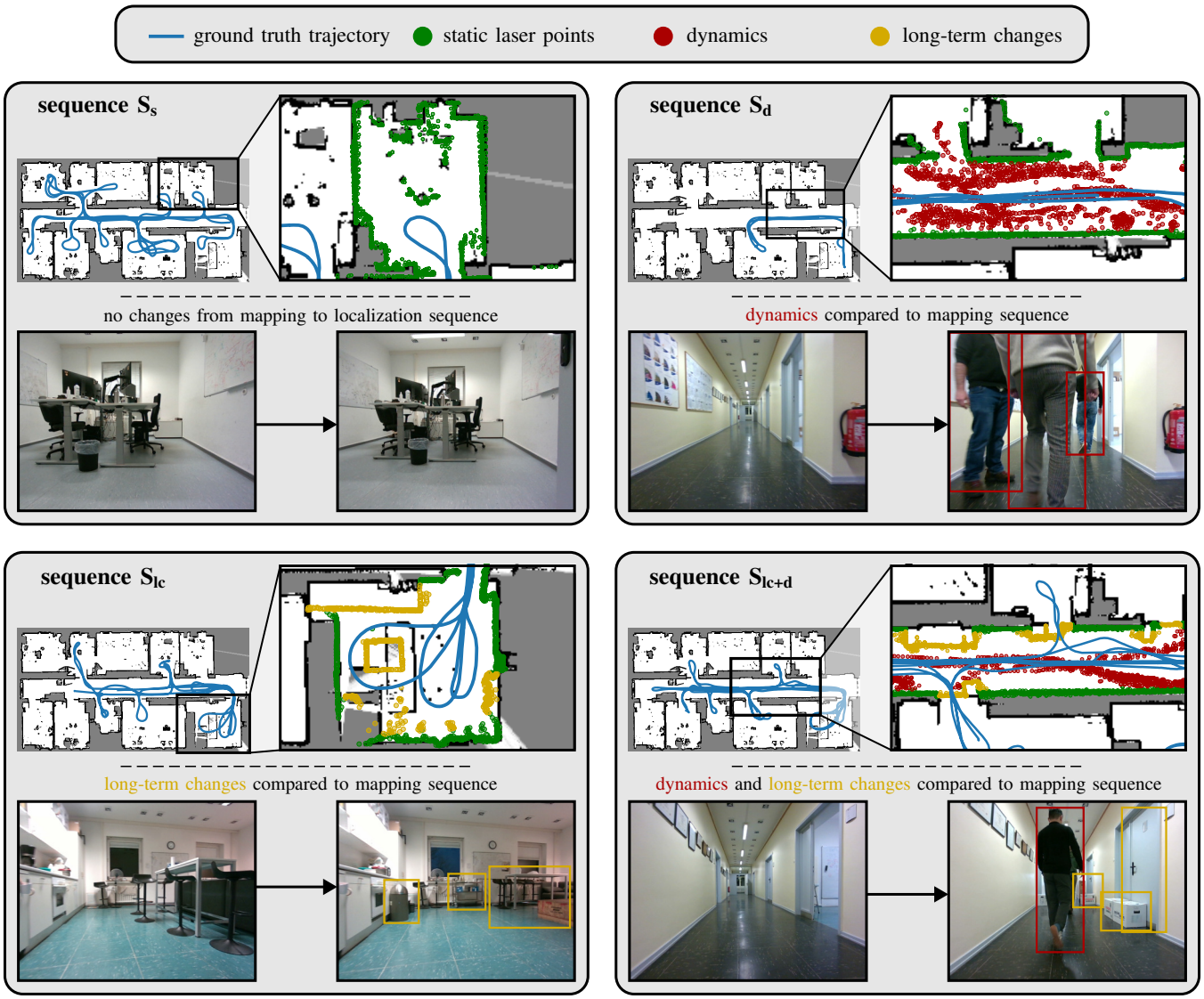


Fig. 5: Overview of the four localization sequences considered in our evaluation. For each sequence, we depict the ground truth trajectory together with potential changes in the environment compared to the mapping sequence. We highlight these changes both in exemplary camera images and the laser data in the enlarged map views.

changes through rearranged furniture or objects added to the scene, and sequence S_{lc+d} has combinations thereof. We divide our evaluation into two aspects: global localization and pose tracking performance. This enables us to isolate how map inconsistencies affect convergence to the correct pose and impede pose-tracking accuracy even after convergence.

B. Global Localization

The first experiment evaluates the global localization performance of the considered algorithms in each sequence. Therefore, we let the algorithms initialize particles uniformly in the free space of the map and test for convergence to the correct global pose. We define convergence as the first point in time when the standard deviation corresponding to the largest eigenvalue both of the translational and rotational part of the localizer's reported pose covariance matrix falls below a threshold of $0.2\,\mathrm{m}$ and $10°$, respectively. We run each algorithm 10 times and measure the success rate and convergence

TABLE II: Global localization success rate and convergence times for the four sample sequences from our dataset over 10 runs each.

sequence	method	success rate [%] ↑	convergence time mean ± std [s] ↓
S_s	AMCL	60	11.77 ± 5.42
	SRRG-Loc	**100**	9.33 ± 4.19
	LocNDF	**100**	13.93 ± 3.72
	ENM-MCL	**100**	$\mathbf{7.95 \pm 2.24}$
S_d	AMCL	30	19.18 ± 13.07
	SRRG-Loc	**90**	$\mathbf{11.46 \pm 2.02}$
	LocNDF	80	13.95 ± 5.3
	ENM-MCL	70	16.09 ± 5.23
S_{lc}	AMCL	10	$\mathbf{11.83 \pm 7.95}$
	SRRG-Loc	60	14.65 ± 2.95
	LocNDF	30	18.47 ± 12.74
	ENM-MCL	**90**	12.09 ± 4.47
S_{lc+d}	AMCL	60	12.29 ± 4.73
	SRRG-Loc	**100**	$\mathbf{11.18 \pm 3.26}$
	LocNDF	**100**	12.63 ± 3.93
	ENM-MCL	90	13.74 ± 2.84

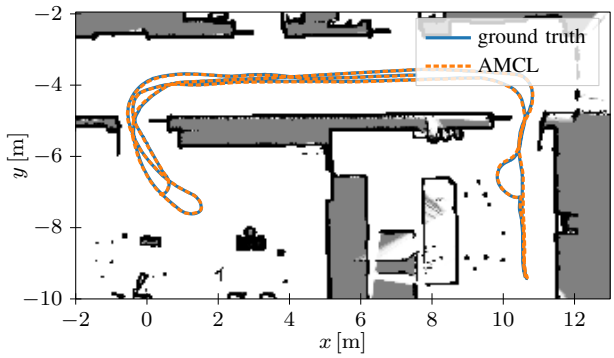
TABLE III: Absolute trajectory error (ATE) in position and orientation for pose tracking for the four considered sample sequences of our dataset. Each entry reports mean ± standard deviation over 10 runs.

seq.	method	position				orientation			
		RMSE ↓ [cm]	max. ↓ [cm]	$< 5\,\mathrm{cm}$ ↑ [%]	$< 10\,\mathrm{cm}$ ↑ [%]	RMSE ↓ [°]	max. ↓ [°]	$< 2°$ ↑ [%]	$< 4°$ ↑ [%]
S_s	AMCL	3.12 ± 0.05	$\mathbf{7.72 \pm 0.45}$	93.87 ± 0.73	$\mathbf{100 \pm 0}$	1.86 ± 0.01	5.39 ± 0.3	70.39 ± 0.44	96.78 ± 0.2
	SRRG-Loc	4.73 ± 0.01	10.29 ± 0.1	60.88 ± 0.37	99.89 ± 0.06	1.82 ± 0	5.47 ± 0.05	71.7 ± 0.17	96.28 ± 0.08
	LocNDF	$\mathbf{2.94 \pm 0}$	7.77 ± 0.07	$\mathbf{94.66 \pm 0.17}$	$\mathbf{100 \pm 0}$	$\mathbf{1.6 \pm 0}$	$\mathbf{5.33 \pm 0.04}$	$\mathbf{78.36 \pm 0.1}$	$\mathbf{98.28 \pm 0.07}$
	ENM-MCL	3.32 ± 0.01	15.16 ± 0.4	92.56 ± 0.2	99.83 ± 0.01	1.83 ± 0	6.11 ± 0.04	71.13 ± 0.08	97.13 ± 0.07
S_d	AMCL	$\mathbf{3.21 \pm 0.2}$	8.56 ± 0.51	89.56 ± 2.42	$\mathbf{100 \pm 0}$	1.87 ± 0.02	$\mathbf{4.83 \pm 0.21}$	73.57 ± 0.57	95.28 ± 1.18
	SRRG-Loc	4.87 ± 0.03	9.58 ± 0.14	58.3 ± 0.57	$\mathbf{100 \pm 0}$	1.83 ± 0.01	6.25 ± 0.09	76.38 ± 0.11	95.06 ± 0.07
	LocNDF	3.31 ± 0.03	$\mathbf{8.53 \pm 0.14}$	90.26 ± 0.37	$\mathbf{100 \pm 0}$	$\mathbf{1.48 \pm 0}$	4.86 ± 0.04	$\mathbf{83.02 \pm 0.11}$	$\mathbf{97.37 \pm 0.17}$
	ENM-MCL	3.4 ± 0.03	21.41 ± 0.48	$\mathbf{90.97 \pm 0.54}$	99.17 ± 0.02	1.87 ± 0	8.26 ± 0.06	79.89 ± 0.31	92.98 ± 0.11
S_{lc}	AMCL	13.51 ± 0.74	70.3 ± 5.23	61.41 ± 1.84	$\mathbf{82.89 \pm 0.79}$	4.49 ± 0.31	$\mathbf{25.31 \pm 2.91}$	61.56 ± 0.66	$\mathbf{84.01 \pm 0.85}$
	SRRG-Loc	258.2 ± 213.8	803.2 ± 372.3	24.18 ± 6.45	55.22 ± 14.94	40.97 ± 23.25	151.6 ± 41.49	45.9 ± 12.01	62.69 ± 15.08
	LocNDF	660.1 ± 217.3	1709 ± 532.1	22.35 ± 13.47	31.87 ± 18.03	74.2 ± 17.66	179.7 ± 0.59	32.59 ± 12.76	42.42 ± 16.09
	ENM-MCL	$\mathbf{9.51 \pm 0.08}$	$\mathbf{40.93 \pm 0.3}$	$\mathbf{62.02 \pm 0.75}$	82.24 ± 0.33	$\mathbf{3.69 \pm 0.02}$	25.5 ± 0.4	$\mathbf{63.69 \pm 0.25}$	83.76 ± 0.24
S_{lc+d}	AMCL	$\mathbf{11.13 \pm 1.68}$	$\mathbf{72.07 \pm 18.53}$	$\mathbf{64.18 \pm 1.33}$	$\mathbf{87.34 \pm 0.96}$	$\mathbf{2.75 \pm 0.22}$	$\mathbf{20.34 \pm 2.85}$	$\mathbf{68.54 \pm 0.62}$	$\mathbf{90.09 \pm 0.61}$
	SRRG-Loc	828.1 ± 160.5	1999 ± 252.6	10.76 ± 2.65	18.2 ± 4.89	79.59 ± 22.01	173.9 ± 14.91	19.43 ± 6.46	27.08 ± 8.15
	LocNDF	972.2 ± 255	2522 ± 431.2	18.62 ± 5.23	21.41 ± 7.2	106.7 ± 25.22	179.9 ± 0.14	21.69 ± 10.55	26.96 ± 13.06
	ENM-MCL	898.5 ± 397.5	2149 ± 1256	14.14 ± 1.54	19.19 ± 2.37	91.41 ± 41.24	166.6 ± 24.81	20.87 ± 10.65	27.84 ± 15.56

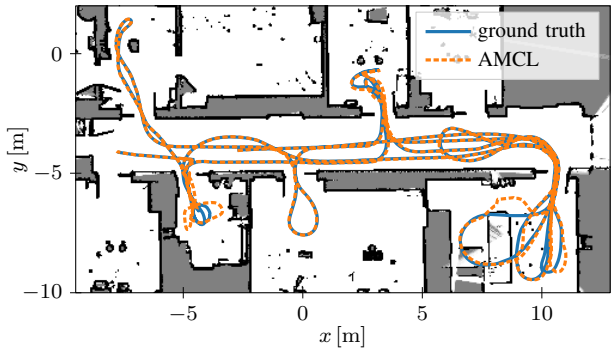
times over all runs in each sequence. We consider a run successful if, at the point of convergence, the error both in translation and orientation is below three times the thresholds for the standard deviations used in the convergence criterion defined previously. We report the results thereof in Tab. II. Overall, the results indicate that AMCL has the lowest localization performance on our sequences and does not always converge to the correct solution, even for a completely static environment. In comparison between the sequences, the success rate notably decreases for sequences S_d and S_{lc} due to observations from the laser scans that are inconsistent with the map. While sequence S_{lc+d} holds the same environmental changes as S_{lc}, the localization performance remains close to the static case since the sequence starts in a part of the environment without drastic scene changes. In terms of convergence times, the algorithms mostly show results of a similar magnitude. However, in many cases, convergence times increase in the presence of environmental changes.

C. Pose Tracking

The second evaluation analyzes the tracking performance of the methods after convergence. Therefore, we ignore the aspect of global localization and initialize the particles around the initial ground truth pose. For this, we sample particles from a Gaussian with standard deviations $[0.2\,\mathrm{m}, 0.2\,\mathrm{m}, 10°]$ along global x and y coordinates and the yaw angle, respectively. To measure pose accuracy, we independently consider the absolute trajectory error (ATE) for translation and rotation and run each baseline again 10 times per sequence. We show the obtained results in Tab. III. We report RMSE over the complete trajectory and maximum of the pose error as well as proportions of the pose errors below a selection of thresholds. The results show that the static sequence S_s and sequence with dynamics S_d in general report similar and low errors, where most of the localizers show errors over nearly the entire trajectory below $10\,\mathrm{cm}$ in translation and $5°$ in orientation. The robustness of the algorithms against dynamics can be explained by the fact



(a) Trajectory of the sequence with dynamics S_d.



(b) Trajectory of the sequence with long-term changes S_{lc}.

Fig. 6: Qualitative results of AMCL on sequences S_d and S_{lc}.

that the algorithms strongly rely on odometry information, such that, even with short-term inconsistencies between the LiDAR sensor information and the map, the prediction of the robot's movement remains relatively accurate. In contrast, the tracking performance in the sequences S_{lc} and S_{lc+d} drastically reduces, where the scene changes in some rooms are so significant that large parts of the laser scans are inconsistent with the map over a longer time, as shown in Fig. 5. While ENM-MCL and AMCL report the lowest errors in sequence S_{lc} and S_{lc+d}, respectively, their maximum errors of $40\,\mathrm{cm}$ and above remain potentially unsafe for

indoor navigation. The other algorithms, as well as ENM-MCL for sequence S_{lc+d}, sometimes entirely diverge and do not recover to the correct pose, hence reporting RMSEs in meter range. Besides the inconsistencies between the sensor data and the map, SRRG-Loc additionally suffers from its assumptions for the movement of particles in the map, as it removes particles when they enter occupied space. While this is advantageous in a static environment as it filters out implausible particles, it becomes particularly detrimental in situations as depicted in Fig. 5 for sequence S_{lc}. Since the robot traverses a part of the environment covered by furniture during the mapping run, the assumption prevents particles from traversing the newly freed space. We depict exemplary qualitative results of AMCL on the sequences S_d and S_{lc} in Fig. 6. While the localization accuracy in the dynamic case is high enough to make the difference to the ground truth nearly unnoticeable, the method's trajectory shows clear deviations for the sequence with long-term changes S_{lc}. Here, it is evident that the localization performance mainly decreases in certain rooms that contain the most drastic scene changes.

In summary, our evaluation suggests that existing localization methods achieve good pose estimation accuracy under static conditions and demonstrate robustness against short-term dynamics. However, large-scale and long-term scene changes pose a considerable challenge to the investigated methods, both during global localization and pose tracking. How to consider environmental changes in the localization algorithm, potentially by integrating additional information provided by the cameras, can be subject to further research. This motivated us to release, together with our data, a public benchmark with non-public ground truth for the remaining sequences of our dataset. Our benchmark website comes with an automated tool to evaluate and rank the performance of participating algorithms on these sequences. Therefore, we evaluate submitted localization pose results against our ground truth poses and provide a set of standardized evaluation metrics. In this way, we enable a fair comparison of the algorithms on our sequences that cover diverse and ongoing challenges and thus aim to stimulate further research.

V. Conclusion

In this paper, we presented a new benchmark and dataset for indoor localization in a multi-room office scenario, including ground truth poses. Our dataset is collected with a mobile robot platform and contains recorded data from RGB-D cameras, laser scanners, and the robot's wheel odometry. It covers diverse scenarios, including short-term dynamics and long-term scene changes, and comes with accurate ground truth in a single global frame. Furthermore, we evaluated existing localization algorithms on a selection of representative sequences from our dataset and pointed out challenging scenarios. Here, especially long-term scene changes revealed themselves as remaining difficult to handle. To enable the research community to approach these challenges and motivate further research in this field, we release a public benchmark challenge with non-public ground truth and server-side evaluations together with our data.

References

[1] N. Akai. Reliable Monte Carlo localization for mobile robots. *Journal of Field Robotics (JFR)*, 40(3):595–613, 2023.

[2] M. Bennewitz, C. Stachniss, W. Burgard, and S. Behnke. Metric Localization with Scale-Invariant Visual Features using a Single Perspective Camera. In *Proc. of the European Robotics Symposium*, 2006.

[3] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. Achtelik, and R. Siegwart. The EuRoC micro aerial vehicle datasets. *Intl. Journal of Robotics Research (IJRR)*, 35(10):1157–1163, 2016.

[4] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. Leonard. Past, Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age. *IEEE Trans. on Robotics (TRO)*, 32(6):1309–1332, 2016.

[5] N. Carlevaris-Bianco, A. Ushani, and R. Eustice. University of Michigan North Campus long-term vision and lidar dataset. *Intl. Journal of Robotics Research (IJRR)*, 35(9):1023–1035, 2016.

[6] L. Carlone, A. Kim, F. Dellaert, T. Barfoot, and D. Cremers. *SLAM Handbook. From Localization and Mapping to Spatial Intelligence*. Cambridge University Press.

[7] S. Ceriani, G. Fontana, A. Giusti, D. Marzorati, M. Matteucci, D. Migliore, D. Rizzi, D. Sorrenti, and P. Taddei. Rawseeds ground truth collection systems for indoor self-localization and mapping. *Autonomous Robots*, 27(4):353–371, 2009.

[8] S. Choi, Q. Zhou, and V. Koltun. Robust Reconstruction of Indoor Scenes. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] E.M.G.N.V. Cruz, S. Oliveira, and A. Correia. Robotics Applications in the Hospital Domain: A Literature Review. *Applied System Innovation*, 7(6), 2024.

[10] F. Dellaert, D. Fox, W. Burgard, and S. Thrun. Monte carlo localization for mobile robots. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 1999.

[11] J.C. Eidson, M. Fischer, and J. White. IEEE-1588™ Standard for a precision clock synchronization protocol for networked measurement and control systems. In *Proc. of the Annual Precise Time and Time Interval Systems and Applications Meeting*, 2002.

[12] M.F. Fallon, H. Johannsson, M. Kaess, and J.J. Leonard. The MIT Stata Center dataset. *Intl. Journal of Robotics Research (IJRR)*, 32(14):1695–1699, 2013.

[13] D. Feng, Y. Qi, S. Zhong, Z. Chen, Q. Chen, H. Chen, J. Wu, and J. Ma. S3E: A Multi-Robot Multimodal Dataset for Collaborative SLAM. *IEEE Robotics and Automation Letters (RA-L)*, 9(12):11401–11408, 2024.

[14] D. Fox. KLD-sampling: Adaptive particle filters. In *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2001.

[15] L. Gao, Y. Liang, J. Yang, S. Wu, C. Wang, J. Chen, and L. Kneip. VECtor: A Versatile Event-Centric Benchmark for Multi-Sensor SLAM. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):8217–8224, 2022.

[16] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[17] G. Grisetti. srrg-localizer2d (1.6.0). `https://gitlab.com/srrg-software/srrg_localizer2d`, 2018.

[18] M. Helmberger, K. Morin, B. Berner, N. Kumar, G. Cioffi, and D. Scaramuzza. The Hilti SLAM Challenge Dataset. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7518–7525, 2022.

[19] S. Ito, F. Endres, M. Kuderer, G. Tipaldi, C. Stachniss, and W. Burgard. W-RGB-D: Floor-Plan-Based Indoor Global Localization Using a Depth Camera and WiFi. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2014.

[20] J. Jiao, H. Wei, T. Hu, X. Hu, Y. Zhu, Z. He, J. Wu, J. Yu, X. Xie, H. Huang, R. Geng, L. Wang, and M. Liu. FusionPortable: A Multi-Sensor Campus-Scene Dataset for Evaluation of Localization and Mapping Accuracy on Diverse Platforms. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2022.

[21] P. Kaveti, A. Gupta, D. Giaya, M. Karp, C. Keil, J. Nir, Z. Zhang, and H. Singh. Challenges of Indoor SLAM: A Multi-Modal Multi-Floor Dataset for SLAM Evaluation. In *Proc. of the Intl. Conf. on Automation Science and Engineering (CASE)*, 2023.

[22] R. Keith and H.M. La. Review of Autonomous Mobile Robots for the Warehouse Environment. *arXiv preprint*, arXiv:2406.08333, 2024.

[23] H. Kuang, Y. Pan, X. Zhong, L. Wiesmann, J. Behley, and C. Stachniss. Improving Indoor Localization Accuracy by Using an Efficient Implicit Neural Map Representation. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2025.

[24] K. Leung, Y. Halpern, T. Barfoot, and H. Liu. The UTIAS multi-robot cooperative localization and mapping dataset. *Intl. Journal of Robotics Research (IJRR)*, 30(8):969–974, 2011.

[25] E. Olson. AprilTag: A Robust and Flexible Visual Fiducial System. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2011.

[26] P. Pfaff, W. Burgard, and D. Fox. Robust Monte-Carlo Localization Using Adaptive Likelihood Models. In *Proc. of the European Robotics Symposium*, 2006.

[27] B. Pfrommer, N. Sanket, K. Daniilidis, and J. Cleveland. PennCOSYVIO: A Challenging Visual Inertial Odometry Benchmark. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2017.

[28] A. Pronobis and B. Caputo. COLD: The CoSy Localization Database. *Intl. Journal of Robotics Research (IJRR)*, 28(5):588–594, 2009.

[29] J. Roewekaemper, C. Sprunk, G. Tipaldi, C. Stachniss, P. Pfaff, and W. Burgard. On the Position Accuracy of Mobile Robot Localization based on Particle Filters combined with Scan Matching. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2012.

[30] M. Sewtz, Y. Fanger, X. Luo, T. Bodenmüller, and R. Triebel. IndoorMCD: A Benchmark for Low-Cost Multi-Camera SLAM in Indoor Environments. *IEEE Robotics and Automation Letters (RA-L)*, 8(3):1707–1714, 2023.

[31] X. Shi, D. Li, P. Zhao, Q. Tian, Y. Tian, Q. Long, C. Zhu, J. Song, F. Qiao, L. Song, Y. Guo, Z. Wang, Y. Zhang, B. Qin, W. Yang, F. Wang, R.H.M. Chan, and Q. She. Are We Ready for Service Robots? The OpenLORIS-Scene Datasets for Lifelong SLAM. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2020.

[32] S. Siddiqi, G.S. Sukhatme, and A. Howard. Experiments in monte-carlo localization using wifi signal strength. In *Proc. of the Intl. Conf. on Advanced Robotics (ICAR)*, 2003.

[33] C. Sprunk, J. Röwekämper, G. Parent, L. Spinello, G.D. Tipaldi, W. Burgard, and M. Jalobeanu. An Experimental Protocol for Benchmarking Robotic Indoor Navigation. In *Proc. of the Intl. Symp. on Experimental Robotics (ISER)*, 2014.

[34] C. Stachniss, J. Leonard, and S. Thrun. *Springer Handbook of Robotics, 2nd edition*, chapter 46: Simultaneous Localization and Mapping. Springer Verlag, 2016.

[35] 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 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2012.

[36] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[37] S. Thrun, D. Fox, W. Burgard, and F. Dellaert. Robust Monte Carlo Localization for Mobile Robots. *Artificial Intelligence*, 128(1-2):99–141, 2001.

[38] I. Ullah, D. Adhikari, H. Khan, M.S. Anwar, S. Ahmad, and X. Bai. Mobile robot localization: Current challenges and future prospective. *Computer Science Review*, 53:100651, 2024.

[39] L. Wiesmann, T. Guadagnino, I. Vizzo, N. Zimmerman, Y. Pan, H. Kuang, J. Behley, and C. Stachniss. LocNDF: Neural Distance Field Mapping for Robot Localization. *IEEE Robotics and Automation Letters (RA-L)*, 8(8):4999–5006, 2023.

[40] L. Wiesmann, T. Läbe, L. Nunes, J. Behley, and C. Stachniss. Joint Intrinsic and Extrinsic Calibration of Perception Systems Utilizing a Calibration Environment. *IEEE Robotics and Automation Letters (RA-L)*, 9(10):9103–9110, 2024.

[41] J. Wolf, W. Burgard, and H. Burkhardt. Robust vision-based localization by combining an image-retrieval system with Monte Carlo localization. *IEEE Trans. on Robotics (TRO)*, 21(2):208–216, 2005.

[42] A. Yilmaz and H. Temeltas. Self-adaptive Monte Carlo method for indoor localization of smart AGVs using LIDAR data. *Journal on Robotics and Autonomous Systems (RAS)*, 122:103285, 2019.

[43] H. Yin, X. Xu, S. Lu, X. Chen, R. Xiong, S. Shen, C. Stachniss, and Y. Wang. A Survey on Global LiDAR Localization: Challenges, Advances and Open Problems. *Intl. Journal of Computer Vision (IJCV)*, 132(8):3139–3171, 2024.

[44] J. Yin, A. Li, T. Li, W. Yu, and D. Zou. M2DGR: A Multi-Sensor and Multi-Scenario SLAM Dataset for Ground Robots. *IEEE Robotics and Automation Letters (RA-L)*, 7(2):2266–2273, 2021.

[45] J. Yin, H. Yin, C. Liang, and Z. Zhang. Ground-Challenge: A Multi-sensor SLAM Dataset Focusing on Corner Cases for Ground Robots. In *Proc. of the IEEE Intl. Conf. on Robotics and Biomimetics (ROBIO)*, 2023.

[46] L. Zhang, M. Helmberger, L. Fu, D. Wisth, M. Camurri, D. Scaramuzza, and M. Fallon. Hilti-Oxford Dataset: A Millimeter-Accurate Benchmark for Simultaneous Localization and Mapping. *IEEE Robotics and Automation Letters (RA-L)*, 8(1):408–415, 2022.

[47] Q. Zhou, J. Park, and V. Koltun. Open3D: A modern library for 3D data processing. *arXiv preprint*, arXiv:1801.09847, 2018.