Challenges of Indoor SLAM: A multi-modal multi-floor dataset for SLAM evaluation

Pushyami Kaveti, Aniket Gupta, Dennis Giaya, Madeline Karp,
Colin Keil, Jagatpreet Nir, Zhiyong Zhang, Hanumant Singh∗

Abstract— **Robustness in Simultaneous Localization and Mapping (SLAM) remains one of the key challenges for the real-world deployment of autonomous systems. SLAM research has seen significant progress in the last two and a half decades, yet many state-of-the-art (SOTA) algorithms still struggle to perform reliably in real-world environments. There is a general consensus in the research community that we need challenging real-world scenarios which bring out different failure modes in sensing modalities. In this paper, we present a novel multi-modal indoor SLAM dataset covering challenging common scenarios that a robot will encounter and should be robust to. Our data was collected with a mobile robotics platform across multiple floors at Northeastern University's ISEC building. Such a multi-floor sequence is typical of commercial office spaces characterized by symmetry across floors and, thus, is prone to perceptual aliasing due to similar floor layouts. The sensor suite comprises seven global shutter cameras, a high-grade MEMS inertial measurement unit (IMU), a ZED stereo camera, and a 128-channel high-resolution lidar. Along with the dataset, we benchmark several SLAM algorithms and highlight the problems faced during the runs, such as perceptual aliasing, visual degradation, and trajectory drift. The benchmarking results indicate that parts of the dataset work well with some algorithms, while other data sections are challenging for even the best SOTA algorithms. The dataset is available at https://github.com/neufieldrobotics/NUFR-M3F.**

Index Terms— **Multi-modal datasets, Simultaneous Localization and Mapping, Indoor SLAM, lidar mapping, perceptual aliasing**

I. INTRODUCTION

This paper presents a multi-modal SLAM dataset of several real-world sequences captured in a large-scale indoor environment. Simultaneous Localization and Mapping is an extensively researched topic in robotics that has seen major advances in recent decades [1]. It is a hardware and software co-design problem, and the performance of the solution is a function of the right choice of complementary sensors, their proper configuration and calibration, vehicle motions, and, finally, the uncertainties in the real-world mapping environment. Often, methods that work well in certain scenarios fail in the real world due to various factors. These may include environmental uncertainties, dynamic objects, illumination artifacts, and issues associated with robotic motion and trajectories.

The performance of state-of-the-art (SOTA) SLAM algorithms is limited by the lack of publicly available datasets for testing: KITTI[2], TUM RGB-D[3], TUM Mono[4], and Euroc MAV[5]. These datasets have many strengths

∗All authors are affiliated with Northeastern University, Boston. MA, USA

Fig. 1: (a) The data collection rig mounts to an omnidirectional base with the sensors approximately 1.2m above the ground. (b) The data collection site, Northeastern University's Interdisciplinary Science and Engineering Complex (ISEC), which has an open atrium and several floors with a high degree of symmetry in their layout and overall design.(c) A composite rendition of the lidar point cloud depicting all the floors from top view.

and have positively impacted the SLAM algorithm design and evaluation using different sensor modalities, including monocular vision, stereo vision, visual-inertial odometry (VIO), RGBD cameras, and 3D lidars. However, new large-scale public datasets with multiple sensing modalities are essential. Recent work [6][7] has also shown that fusing data from multiple sensors improves the robustness and accuracy of SLAM estimates in challenging scenarios often encountered in the real-world.

Our dataset described in table I consists of visual, inertial, and lidar sensor data, allowing for multi-modal SLAM

TABLE I: An overview of the indoor multi-modal SLAM datasets

	Sensors	Arrangement	Frame Rate(Hz)	No. of Sequences	Platform	Ground truth	Sync	Environment	
Newer College	2 GS cameras 1 LiDAR 1 IMU	2 stereo, 2 Non-overlapping	30	3	Handheld	Survey grade 3D Imaging Laser	HW + SW	Campus	HW + SW
PennCosy-VIO	3 RGB GS cameras 2 Gray RS cameras 2 IMUs	3 min overlapping, 2 stereo	20	4	Handheld	Fiducial Markers	HW + SW	Campus	
HILTI	5 GS Gray cameras 2 LiDARS 3 IMU	2 stereo, 3 Non-overlapping	10	12	Handheld	MoCAP	HW +SW	Construction site	
HILTI-Oxford	5 GS IR cameras 1 LiDAR 1 IMU	2 stereo, 3 Non-overlapping	40	16	Handheld	Survey grade 3D Imaging Laser	HW + SW	Construction Site	
Ours	7 GS RGB cameras 1 Zed 2i 1 LiDAR 1 IMU	5 Fronto-Parallel 2 Side-ways	20	14	Mobile Robot	Fiducial markers +3D LiDAR alignment	HW + SW	Campus office	

evaluations. The specifications of each sensor are detailed in table II. The entire sensor suite shown in figure 1a is time synchronized and spatially calibrated across all sensors for accurate data capture and analysis as shown in figure 2.

To the best of our knowledge, this is the first dataset that has continuous multi-floor data for SLAM, and we know of no algorithm that is capable of processing the uninterrupted data across multiple floors into an accurate map of the entire building in an autonomous manner. The dataset presents new fundamental challenges to further the research on informing design decisions and algorithmic choices in performing SLAM with higher reliability. Even if (or when) this becomes possible, the dataset poses interesting questions related to localization due to symmetry across floors. This dataset serves to complement the recent multi-modal benchmarking datasets [8][9][10]. The contributions of this paper can be summarized as:

- It outlines challenging multi-modal indoor datasets covering a variety of scenarios including featureless spaces, reflective surfaces, and multi-storeyed sequences.
- The multi-storeyed sequences, as is typical with modern architecture, features floors that are essentially identical in design and layout which leads to perceptual aliasing scenarios. These scenarios trip up state-of-the-art SLAM algorithms, the vast majority of which rely on bag of words models for relocalization and loop closure.
- It features an extensive set of sensors consisting of seven cameras, a high-resolution lidar, and an IMU. All the sensors are hardware synchronized and calibrated across the entire sensor suite.
- We have benchmarked several state-of-the-art algorithms across the visual, visual-inertial, and lidar SLAM methodologies and present a comparison among these different algorithms and sensor modalities that highlights their individual strengths and areas where there are engineering or fundamental theoretical issues that the community may need to focus on.

II. RELATED WORK

Several SLAM datasets exist in the literature which vary in regards to the data acquisition environment, varying sensing modalities, type of motions, degree of difficulty, number of sensors, and synchronization of the data capture. The table I summarizes several multi-modal datasets closely related to us.

KITTI[2] is one of the first and most popular benchmarking multi-modal datasets motivated by self-driving cars. It has a linear array of four cameras consisting of two stereo pairs - One RGB and one grayscale, a lidar, an IMU, and a GPS. Following this, many outdoor urban datasets emerged in the domain of autonomous driving, such as [11][12][13], which allowed the evaluation of various odometry and SLAM algorithms. Many earlier indoor SLAM datasets targeted visual odometry(VO) and visual-inertial odometry (VIO) tasks for monocular and stereo systems. The TUM[14] and EUROC[5] datasets are extensively used for benchmarking VO and VIO solutions. These datasets have global shutter stereo cameras, hardware synchronized with the IMU, and millimeter-accurate ground truth from motion capture systems.

A few recent efforts [8][15] gathered multi-sensor (beyond stereo) and multi-modal data in urban indoor environments. PennCOSYVIO[15], is collected at Upenn's campus area with a stereo VI sensor, two project Tango devices, and three GoPro cameras arranged in a minimally overlapping configuration. The sensors are mounted on a handheld platform and carried across indoor and outdoor areas. The ground truth is provided using fiducial markers placed along the trajectories. The Newer College Dataset[8] and its extension contain synchronized image data from the Alphasense sensor with four cameras - two facing forward and two on the side as well as a lidar mounted on a handheld device.

More recently, the Hilti[9] and Hilti-Oxford[16] datasets attracted a lot of attention through their SLAM challenge, where multiple teams from both academia and industry participated. The main objective of this dataset is to push the limits of the state-of-the-art multi-sensor SLAM algorithms to aid real-world applications. There are indoor and

Fig. 2: Top view of the sensor rig showing sensor frames for the front-facing camera array (red), the non-overlapping side cameras (orange), the ZED camera (purple), the IMU (green) and the lidar (blue). Note the above image follows the convention that $\otimes$ indicates an axis into the plane of the image, and $\bullet$ indicates an axis out of the plane of the image. All of the cameras are z-axis forward, y-axis down.

TABLE II: Description of various sensors and their settings used to collect our dataset. Note that Zed2i sensor is available only in the Snell dataset.

Sensor	No	Type	Description
Camera	7	FLIR Blackfly S USB3	1.3 megapixel color cameras with a resolution of 720 x 540 and FoV of 57 ° at 20 hz.
Lidar	1	Ouster OS-128	128 channel LiDAR with vertical FoV of 45° at 10 Hz
RGB-D camera	1	Zed 2i	stereo cameras with resolution of 1280 x 720 at 15 Hz, IMU at 200 Hz
IMU	1	Vectornav 100	9-DOF IMU running at 200 Hz.

outdoor sequences of construction sites and parking areas that contain some challenging scenarios of abrupt and fast motions and featureless areas. These datasets are collected with an Alphasense five-camera module with a stereo pair, three non-overlapping wide-angle cameras, an IMU, and two laser scanners.

All these datasets contain challenging sequences with changing lighting and texture, challenging structures such as staircases, and featureless spaces. Our dataset also consists of multi-modal data with cameras, lidar, and inertial measurements. In addition to featuring the challenging scenarios mentioned above, our dataset showcases symmetrical structures located on multiple floors which present unique challenges due to perceptual aliasing.

III. DATA COLLECTION SYSTEM

A. Hardware Setup

We built a rigid multi-sensor rig consisting of seven cameras, five facing forward and two facing sideways, an inertial measurement unit (IMU), a zed 2i sensor, and a lidar. The description and configuration of the sensors is shown in table II. The sensors' placement and coordinate frames are shown in the schematic figure 2. The cameras are arranged to accommodate overlapping and non-overlapping configurations. The front-facing multi-stereo camera array, together with the left and right cameras, collectively yields a 171-degree field of view. We use Ouster's 128-beam high-resolution lidar, which gives high-density point clouds with 130,000 points. All the cameras except the ZED stereo cameras are hardware synchronized with IMU at 20 frames per second. We built a buffer circuit where the IMU sends a signal to trigger all the cameras simultaneously. The lidar and zed sensor are software/network time synchronized with the other sensor streams. All the sensor timestamps are assigned based on the hardware trigger in combination with the computer's system clock. The multi-sensor rig and a Dell XPS laptop with 32GB RAM were mounted on a Clearpath's Ridgeback robotic platform and driven using a joystick across multiple floors of two of the Northeastern University's buildings for data collection. The Zed 2i sensor

is mounted for data collection in the Snell library building and is unavailable for the ISEC building.

B. Calibration

We obtain both intrinsic and extrinsic calibration for cameras, IMU, and lidar by applying different methods. We used Kalibr[17] to obtain the intrinsic and extrinsic parameters of the overlapping set of cameras and zed stereo cameras using a checkerboard target. For the side-facing cameras, it is not possible to use the same multi-camera calibration methods as we need the cameras to observe a single stationary target at the same time to find the correspondences and solve for the relative transformation. Instead, we perform IMU-camera calibration independently for the two side-facing cameras and the center front-facing camera to obtain $T_{c_i}^{IMU}$ and chain the camera-IMU transformations to get the inter-camera relative transforms using $T_{c_j}^{c_i} = (T_{c_i}^{IMU})^{-1} T_{c_j}^{IMU}$. We used target-based open source software packages [18] and [19] to obtain the lidar-camera extrinsic calibration parameters but noticed some misalignment of the point cloud with images which amplifies with range. We adjust the error by manually aligning the lidar point cloud with camera data.

C. Ground truth

Ground truth poses are essential to test and evaluate the accuracy of the SLAM algorithms. However, generating ground truth trajectories in indoor environments is a challenging task due to the lack of GPS signals and range limitations of popular indoor ground truthing mechanisms like MOCAP. There are additional challenges particular to our dataset, where the robot moves across multiple floors which makes it impossible to deploy a MOCAP system to track the robot. Given the necessity of ground-truth data for benchmarking novel algorithms, we used fiducial markers-based ground truth. These markers were used as stationary targets to localize the robot when they came into the cameras' field of view. We carefully placed multiple fiducial markers made of April tags[20] near the elevators on each floor. The location is chosen so as to allow the April tags to be visible at the start and end of trajectories on each floor as we drive the robot in loops, as well as at the transits across floors when we enter and exit the elevators. We explain how we compute the error metrics in detail in section (V-A).

Fig. 3: This figure shows a sample of the various available data streams, showing (a) the left facing side camera (Cam5), (b) and (c) a stereo pair from the front facing array (Cam1 & Cam3), (d) the right facing side camera (Cam6), and (e) the lidar point cloud.

Fig. 4: The full dataset has several points where the robot enters an elevator. The vision-only and lidar SLAM algorithms are not able to handle a scenario where significant movement is not rendered in the data. This figure shows the z-axis IMU acceleration as the robot ascends in the elevator from the first to the second floor. The spikes as the robot enters and exits the elevator correspond to the robot wheels rolling over the gap between the elevator and the hallway.

IV. DATASET

We collected two large datasets of indoor office environments. The datasets were generated by driving in a loop through different floors of Northeastern University's campus buildings and traveling by elevator between floors. The trajectories include several challenging scenarios that occur on a day-to-day basis, including narrow corridors, featureless spaces, jerky and fast motions, sudden turns, and dynamic objects, which are commonly encountered by a mobile robot in urban environments. All the trajectories have loops to allow the SLAM systems to perform loop closure and compute drift when continuous ground truth is unavailable. All the data is collected using ROS drivers for the respective sensors. The dataset details, including location, length of the trajectory, and ground truth, are consolidated in the table III.

A. ISEC Dataset:

The multiple-floor trajectory was collected in Northeastern University's Interdisciplinary Science and Engineering Complex (ISEC) building. There are four complete floor sequences in the dataset and multiple transit sequences, which include five elevator rides between floors. We start on the 5th floor and drive through the space such that it contains two loops, where the second part is a trajectory down and back a long corridor with a loop closure. We then take an elevator ride to the 1st floor, where we acquire another loop. The 1st floor sequence contains more dynamic objects, glass, and distinct architecture when compared to the other floors. From the 1st floor, we transit through a long corridor with white walls, take an elevator to 3rd floor, and then another one to the 4th floor. We cover the 4th floor and then proceed to the 2nd floor before taking the final elevator ride to the 5th floor, where we started. The first loop of 5th floor, 4th, and 2nd floor sequences are nearly identical with similar-looking office spaces. Thus, these indoor sequences cover areas with good and bad natural lighting, a mix of artificial and natural light, reflections, and dynamic content, such as students. The indoor data snapshots are shown in the figure 3.

B. Snell Library Dataset

This dataset is collected across multiple floors of Northeastern University's library building by taking elevator rides similar to the ISEC dataset. In general, the Snell dataset has better visual features but has longer trajectories with more dynamic content than the ISEC dataset, which can be a failure point for SLAM algorithms. This sequence poses a challenge to SLAM algorithms to map highly dynamic environments. We travel through 3 floors of the building with loop closures on each floor and where the 1st floor's appearance differs from the other two floors.

V. BENCHMARKING THE SOTA

To demonstrate the quality and usefulness of the dataset, we benchmark across a set of well-known state-of-the-art SLAM algorithms. The investigated algorithms are selected so as to have a broad coverage of the field, including visual SLAM, visual-inertial, and lidar-based solutions. The complete list of algorithms can be seen in table IV. We also provide the configuration settings we use to run each algorithm.

A. Evaluation

We run the visual and visual-inertial methods in stereo mode to use the metric scale in the evaluation. We use front-facing cameras 1 and 3 as the stereo pair for each algorithm (see figure 2 for camera placement), except for MCSLAM, which uses the full array of front-facing cameras. This pair was selected as a compromise between a wider stereo baseline and camera proximity to the IMU. We evaluate visual SLAM algorithms on trajectories collected

TABLE III: A comprehensive list of all the sequences in our dataset and their description. Trajectory lengths are approximate and should only be used for qualitative comparison. They were derived from the best available trajectory estimate for each segment. This was typically lidar odometry (LegoLOAM) for the loop sequences and VIO (Basalt or VINS Fusion) for sequences inside elevators. See section V for more details on trajectory estimates.

Datasets				
Label	Size (GB)	Duration (s)	Appx. Length (m)	Description
ISEC				
full_sequence	515.0	1539.70	782	reflective surfaces, minimal dynamic content, daylight, symmetric floors, elevators, open atrium
5th_floor	145.8	437.86	187	one loop, one out and back
transit_5_to_1	36.8	109.00	*	transit from 5th to 1st floor in middle elevator
1st_floor	43.0	125.58	65	one loop, open layout different from other floors, many exterior windows
transit_1_to_4	112.4	337.40	144	transit across 1st floor, up to 3rd floor in freight elevator, across 3rd floor, up to 4th floor in right elevator
4th_floor	43.2	131.00	66	one loop, some dynamic content towards end
transit_4_to_2	21.9	65.00	22	transit from 4th floor to second floor in right elevator,
2nd_floor	89.7	266.00	128	two loops in a figure eight
transit_2_to_5	22.2	65.86	128	transit from 2nd floor to fifth floor in right elevator
SNELL LIBRARY				
full_sequence	573.5	1,700.6	699	feature rich rooms, featureless hallways, many obstacles, stationary and dynamic people in scene
1st_floor	144.6	428.70	221	two loops with shared segment, some dynamic content
transit_1_to_3	28.3	84.00	*	transit from 1st floor to 3rd floor in left elevator
3rd_floor	213.7	633.59	345	two concentric loops with two shared segments, narrow corridor with dynamic content, near field obstructions
transit_3_to_2	27.8	82.41	*	transit from 3rd floor to 2nd floor in right elevator
2nd_floor	126.1	374.00	186	one loop, out and back in featureless corridor
transit_2_to_1	33.0	97.90	*	transit from 2nd floor to 1st floor in right elevator, dynamic objects cover FOV near end

on each floor, whereas visual-inertial algorithms are also evaluated during the transit sequences in the elevators. The elevator sequences are particularly valuable as they give us an insight into the utility of the inertial sensors when vision is ineffective, which is discussed in section V-B. We conducted quantitative analysis on the ISEC dataset by computing error metrics and limited the Snell Library dataset to qualitative results.

In most portions of the dataset, lidar odometry computed using Lego-LOAM can be used as a reasonable ground truth, but it does fail in some portions, and while the results are very good qualitatively, it is non-trivial to compute an upper bound on trajectory errors in the resulting pseudo ground truth. To avoid this kind of analysis, we provide a more limited ground truth evaluation for the dataset using fiducial markers, with a separate evaluation for each floor. We mount an AprilTag [20] tracking target on walls that are visible at the beginning and end of the trajectory at each floor, giving a fixed reference point from which to compute the drift accumulated by each algorithm. For the initial and final portions, when the target is visible, we compute the ground truth poses of the robot $\mathbf{T_{rig}^{target}}$ by localizing it with respect to the target using PnP ransac[21] followed least squares optimization. To align the trajectories, we estimate the rigid body transformation $\mathbf{T_O^{target}} \in \mathbf{se(3)}$ using the positions $\mathbf{t_{rig}^{O^{(i)}}}$ and $\mathbf{t_{rig}^{target^{(i)}}}$ of the tracked and ground truth poses belonging to the starting segment of the trajectory such that

$$T_O^{target} = \underset{T_O^{target} t_{rig}}{argmin} \sum_i \|T_O^{target} t_{rig}^{O^{(i)}} - t_{rig}^{target^{(i)}}\|^2 \quad (1)$$

Once we have the transformation $\mathbf{T_O^{target}}$, we compute the total translational error or the drift at the end of the trajectory between the investigated algorithm's reported pose and the ground truth pose computed using the fixed markers. We report this final drift error for each investigated algorithm as the Absolute Translational Error(ATE), and as a percentage of the approximate total length of the trajectory in table IV. We compute this drift for each floor individually for all the algorithms. We compute the average of the ATEs

TABLE IV: This table outlines the performance of various algorithms on the ISEC dataset. We evaluate each algorithm on loops on the 5th, 1st, 4th, and 2nd floors, in the order they appear in the continuous dataset. Inertial algorithms are also evaluated on the full dataset, which includes elevator transits between floors. Results are reported as the absolute transnational error at the final position in meters, and as a percentage of the estimated trajectory length.We run each algorithm with loop closure disabled, because most algorithms can use the AprilTag markers to form a loop closure, bringing the error close to zero, which does not produce a useful performance metric. While testing the 2nd and 4th floors individually, vins-Fusion resulted in unusually high drift and was left out of this analysis. All vins algorithms surprisingly display higher drift than the visual counterparts due to issues with initialization. We perform a loop closure analysis in Discussion subsection A.

Algorithm	5th Floor		1st Floor		4th Floor		2nd Floor		Full Dataset	
	ATE(m)	%	ATE(m)	%	ATE(m)	%	ATE(m)	%	ATE(m)	%
Visual SLAM										
ORB-SLAM3[22]	0.516	0.28%	0.949	1.46%	0.483	0.73%	0.310	0.24%	–	–
SVO[23]	0.626	0.33%	0.720	1.11%	0.482	0.73%	0.371	0.29%	–	–
MCSLAM	0.778	0.42%	1.085	1.67%	0.484	0.73%	0.458	0.36%	–	–
Visual-Inertial										
vins-Fusion[24]	1.120	0.60%	2.265	3.48%	-	-	-	-	15.844	2.03%
Basalt[25]	1.214	0.65%	4.043	6.22%	1.809	2.74%	3.054	2.39%	1.753	0.22%
SVO-inertial	0.649	0.35%	2.447	3.76%	0.558	0.85%	0.621	0.48%	16.202	2.07%
Deep Learning										
Droid SLAM[26]	0.441	0.24%	0.666	1.02%	0.112	0.17%	0.214	0.17%	–	–
LIDAR										
LEGO LOAM[27]	0.395	0.21%	0.256	0.39%	0.789	1.20%	0.286	0.22%	–	–

(a) Glass Surfaces   (b) Elevator Areas

Fig. 5: A sample of some challenging points in the dataset. Image (a) shows a glass wall with reflections that can introduce spurious features. Image (b) shows one of the elevator areas, where once the robot enters, the exteroceptive sensors such as LiDARs and cameras are fundamentally limited to track motion.

accumulated at the April tags on different floors for inertial algorithms. The April tags are placed at exactly known locations on each floor so that they are displaced vertically by a fixed distance, which is verified from the building floor plan.

B. Discussion

We want to point out that the accuracy metric does not fully describe the performance of a SLAM system. Evaluating drift from the beginning to the end of a trajectory can overlook essential details but is somewhat reflective of the pass-fail nature of real-world scenarios. A more comprehensive evaluation should look at features, like loop detection and closure, tracking failures, and map correction while considering reliability and robustness. In this section, we provide some qualitative assessments of the tested algorithms.

1) Perceptual Aliasing: Our dataset targets this primary challenge by showcasing multi-floor trajectories with similar-looking areas. Most of the vision-based SLAM frontends use a bag of words model [28] to compute the appearance-based similarity between images for loop detection. In addition, vision-only SLAM methods inherently lack the ability to

recognize elevator motion. Based on the end-to-end runs of the algorithms, we observed that all the evaluated VO and VIO algorithms are prone to wrong loop closures, confusing one floor with another. This happens with the 5th, 4th, and 2nd floors, which are symmetrical in structure, color, and layout. This leads to incorrect loop constraints between poses belonging to different floors causing the entire trajectory of one floor to shift in space. Figure 6 shows the constraints as edges between the $2^{nd}, 3^{rd}, 4^{th}$, and 5th floors even though there is no direct visibility across them whereas the first floor remains disconnected since it is unique in appearance. As a result, the trajectories appear to be on the same floor, and the possibility of wrong loop detections is high. In the case of VIO, despite having a good sense that we are not on the same floor, incorrect loop detections still happen.

2) Visual Degradation: Visual degradation occurs at multiple places along the trajectories when we encounter featureless spaces, reflective surfaces, and dynamic content, as shown in figure 5. All the algorithms run without tracking loss on the 1st, 2nd, and 4th floors with minimal drift. In the presence of dynamic objects such as moving people, stereo visual slam algorithms cause jagged artifacts due to corrupted relative motion estimates, as shown in figure 7. Visual-inertial and multi-camera SLAM systems do not display these problems. The vision-only algorithms do not always provide accurate estimates when we run into plain walls, glass surfaces, and during elevator rides. Among vision-only methods, feature-based methods such as ORBSLAM3 are more prone to tracking failures when featureless walls are encountered. Direct methods like SVO can still track due to optical flow but result in incorrect pose estimates causing drift in the subsequent poses. DroidSLAM, which is a learning-based stereo method, also copes in featureless scenarios; however, it lacks scale. These problems are highlighted in figure 7. Visual-inertial algorithms perform well in these scenarios due to the presence of a proprioceptive inertial sensing component, which can detect the physical motion

Fig. 6: This shows the perceptual aliasing problem that is typical of modern buildings. (a) It shows the estimated trajectory of Basalt, a visual-inertial SLAM system for the full multi-floor sequence of the ISEC building without the loop detection, and (b) shows the same sequence run after the loop closure detection. Here the green line segments connecting floors are the incorrectly identified loop closure constraints between poses due to the similarity in appearance.

Fig. 7: This figure shows the 5th-floor trajectories calculated by the various algorithms with two highlighted areas, **(A)** and **(B)**. **(A)** shows a portion of the sequence with dynamic content and its impact on the trajectory estimates, resulting in jagged artifacts for the vision-only algorithms. **(B)** highlights a featureless environment during a tight turn, which caused incorrect trajectory estimates or failure in the vision-only algorithms.

Fig. 8: This figure shows the difference in performance when we run VINS-Fusion on the Library dataset with the VN 100 IMU and ZED IMU. The figure compares x, y, z positions estimated by VINS fusion in both configurations. The red dotted lines show when we enter the elevator. On every floor, we come back to the starting position, and after 2nd floor, we come back to the first floor starting position again which was our origin. The figure clearly shows that VINS Fusion accumulates more drift with ZED sensor setup in all three axes.

of the vehicle. However, we observed that inertial sensing is not always effective when vision fails. For instance, when we ride in the elevator, the visual features detected on the elevator walls interfere with the inertial sensing leading to erroneous poses.

3) Other Issues: We have noticed that the performance of VIO algorithms heavily relies on the initial conditions and parameter tuning. In some sequences, the algorithms perform poorly when we start from specific points. Even starting the data with a time difference of +/- 2 seconds shifts the final drift by about 5 meters. The current SLAM algorithms have a massive list of different parameters that need to be tuned specifically to the dataset. These parameters are generally not standardized or consistent across different algorithms, and tuning them can be arduous. Learning-based methods have an edge in this regard since they do not need as much manual intervention. Additionally, the type and configuration of sensors also impact the performance of the algorithms, which is essential but is unfortunately one of the less researched topics. To demonstrate this, we compare the

estimated trajectories of one of the visual-inertial algorithms (VINS-Fusion) executed on the ZED's stereo inertial system and the stereo configuration with VN100 IMU used in earlier evaluations on the complete multi-floor sequence of the Snell Library dataset. We observe that the two runs differ significantly, as shown in figure 8.

4) Potential usage: The previous discussion clearly shows that the current algorithms fall short in performing large-scale indoor SLAM. There is much room for improvement in the various real-world scenarios discussed above. An upcoming research direction in this regard is to incorporate semantic information from vision into the SLAM framework, as explored in many recent works. One possible solution to improve loop closure detection would be to use contextual information specific to the location, structure, and objects to

distinguish between the floors. There is also a need for better modeling of IMU data that captures the noise properties better [29], contributing to better SLAM back-ends.

VI. CONCLUSION

We have presented a novel multi-modal SLAM dataset that contains visual, inertial, and lidar. The dataset contains several challenging sequences collected by driving a mobile robot across multiple floors of an open-concept office space with narrow corridors, featureless spaces, glass surfaces, and dynamic objects, which challenge the SLAM algorithms. One of the exciting features of our dataset is the symmetric and visually similar locations across different floors that cause perceptual aliasing. We evaluated several SLAM and visual odometry methods across different sensor modalities. The results demonstrate the limitations and areas of improvement in the current SOTA. The main goal of this dataset is to enable the development and testing of novel algorithms for indoor SLAM to address the various challenges discussed. We intend to expand the dataset to outdoors and add more challenging sequences in the future.

REFERENCES

[1] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[3] Jürgen Sturm, Wolfram Burgard, and Daniel Cremers, "Evaluating egomotion and structure-from-motion approaches using the TUM RGB-D benchmark," in *Proc. of the Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RJS International Conference on Intelligent Robot Systems (IROS)*, 2012, vol. 13.

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

[5] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart, "The EuRoC micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[6] Lintong Zhang, David Wisth, Marco Camurri, and Maurice Fallon, "Balancing the budget: Feature selection and tracking for multi-camera visual-inertial odometry," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1182–1189, 2021.

[7] Mohammed Chghaf, Sergio Rodriguez, and Abdelhafid El Ouardi, "Camera, lidar and multi-modal slam systems for autonomous ground vehicles: a survey," *Journal of Intelligent & Robotic Systems*, vol. 105, no. 1, pp. 2, 2022.

[8] Milad Ramezani, Yiduo Wang, Marco Camurri, David Wisth, Matias Mattamala, and Maurice Fallon, "The newer college dataset: Handheld lidar, inertial and vision with ground truth," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4353–4360.

[9] Michael Helmberger, Kristian Morin, Beda Berner, Nitish Kumar, Giovanni Cioffi, and Davide Scaramuzza, "The hilti slam challenge dataset," *IEEE Robotics and Automation Letters*, 2022.

[10] Peng Jiang, Philip Osteen, Maggie Wigness, and Srikanth Saripalli, "Rellis-3d dataset: Data, benchmarks and analysis," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 1110–1116.

[11] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[12] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al., "Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9308–9318.

[13] Siddharth Agarwal, Ankit Vora, Gaurav Pandey, Wayne Williams, Helen Kourous, and James McBride, "Ford multi-AV seasonal dataset," *The International Journal of Robotics Research*, vol. 39, no. 12, pp. 1367–1376, 2020.

[14] David Schubert, Thore Goll, Nikolaus Demmel, Vladyslav Usenko, Jörg Stückler, and Daniel Cremers, "The tum vi benchmark for evaluating visual-inertial odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1680–1687.

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

[16] Lintong Zhang, Michael Helmberger, Lanke Frank Tarimo Fu, David Wisth, Marco Camurri, Davide Scaramuzza, and Maurice Fallon, "Hilti-oxford dataset: A millimeter-accurate benchmark for simultaneous localization and mapping," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 408–415, 2022.

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

[18] Ankit Dhall, Kunal Chelani, Vishnu Radhakrishnan, and K. Madhava Krishna, "Lidar-camera calibration using 3d-3d point correspondences," *ArXiv*, vol. abs/1705.09785, 2017.

[19] Lipu Zhou, Zimo Li, and Michael Kaess, "Automatic extrinsic calibration of a camera and a 3d lidar using line and plane correspondences," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5562–5569, 2018.

[20] John Wang and Edwin Olson, "AprilTag 2: Efficient and robust fiducial detection," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2016.

[21] Richard Hartley and Andrew Zisserman, *Multiple view geometry in computer vision*, Cambridge university press, 2003.

[22] Carlos Campos, Richard Elvira, Juan J. Gomez, José M. M. Montiel, and Juan D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[23] Christian Forster, Zichao Zhang, Michael Gassner, Manuel Werlberger, and Davide Scaramuzza, "SVO: Semidirect visual odometry for monocular and multicamera systems," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 249–265, 2017.

[24] Tong Qin and Shaojie Shen, "Online temporal calibration for monocular visual-inertial systems," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3662–3669.

[25] V. Usenko, N. Demmel, D. Schubert, J. Stueckler, and D. Cremers, "Visual-inertial mapping with non-linear factor recovery," *IEEE Robotics and Automation Letters (RA-L) & Int. Conference on Intelligent Robotics and Automation (ICRA)*, vol. 5, no. 2, pp. 422–429, 2020.

[26] Zachary Teed and Jia Deng, "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras," *Advances in neural information processing systems*, 2021.

[27] Tixiao Shan and Brendan Englot, "Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4758–4765.

[28] Dorian Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, October 2012.

[29] Jagatpreet Nir, Benjamin Deming, and Hanumant Singh, "High fidelity inertial measurement unit (imu) modeling for underwater visual inertial navigation," in *OCEANS 2021: San Diego–Porto*. IEEE, 2021, pp. 1–8.