Kalpa Publications in Computing

Volume 22, 2025, Pages 368–379

Proceedings of The Sixth International Conference on Civil and Building Engineering Informatics

A Real World Visual SLAM Dataset for Indoor Construction Sites

Wenyu Li[1], Xinyu Chen[1], and Yantao Yu[12]*

[1] Department of Civil and Environmental Engineering,
The Hong Kong University of Science and Technology,
Hong Kong 999077, China.
wlief@connect.ust.hk, xchengl@connect.ust.hk, ceyantao@ust.hk

[2] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute,
Shenzhen 518045, China

Abstract

This paper presents a RGBD slam construction dataset with a mounted platform, designed to collect the unique challenges encountered in construction sites. An Ouster OS0-128 LiDAR is utilized as the sensor of LiDAR SLAM, working as the ground truth for localization. Our dataset records various construction settings with different stages of building materials and structures, such as concrete, brick, plaster, and putty, providing a comprehensive benchmark for training and evaluating SLAM algorithms. Through testing on current SLAM algorithms, we demonstrate the limitations of traditional approaches in these environments and provide a VINS based algorithm as the benchmark. This dataset serves as a valuable resource for researchers aiming to enhance SLAM performance in the real construction environments. The detailed information of the dataset is available at https://github.com/WenyuLWY/HCIC-Construction-VSLAM-Dataset.git

1 Introduction

In the field of automated construction, the deployment of construction robots has gained significant traction for the automation of tasks such as surveying, inspection, and material handling. Compared to human workers, construction robots can work continuously without fatigue, reducing project costs, while also improving the overall quality and consistency of different construction tasks. A key factor in enabling these robots to operate effectively in complex construction environments is accurate localization. Without precise positioning, robots cannot operate reliably across different sites, which is critical for ensuring safe and efficient automation.

Indoor localization presents greater challenges compared to outdoor environments due to the absence of GPS and other external positioning systems. To address these challenges, Simultaneous Localization and Mapping (SLAM) has developed as a foundational technology in the field of construction robotics, playing an essential role in enabling robots to autonomously navigate and localize within complex construction environments.

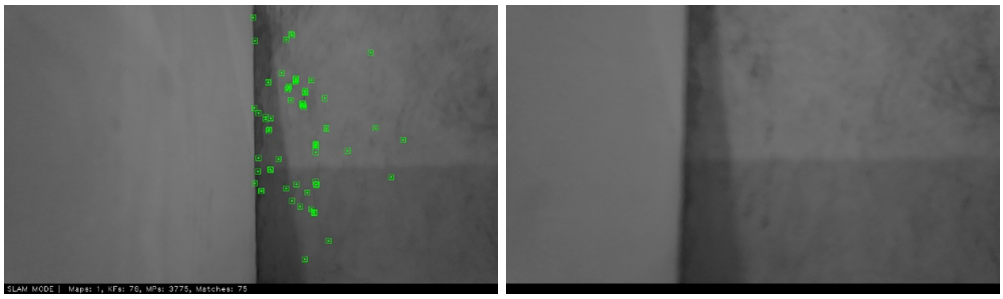
*Corresponding Author.

Although LiDAR-based SLAM is commonly used in construction, it faces limitations such as high costs and sensor weight. Visual SLAM (V-SLAM), by contrast, has gained prominence for its advantages. It uses cameras, which are more cost-effective and lightweight, and can capture rich environmental details such as texture, color, and structure, providing comprehensive information for mapping and localization. With recent advancements in computer vision and image processing, visual SLAM has also become more robust, offering greater flexibility in challenging construction environments.

Figure 1: Typical indoor working conditions for construction robots (Figures are from Bright Dream Robotics(BDR) company in Foshan, China, which provides us the building sites for data collecting.)

Using V-SLAM datasets is an effective method for testing and evaluating the performance of localization algorithms. However, most existing SLAM benchmark have been collected in finished indoor environments, such as offices or homes [1], or in outdoor urban and autonomous driving scenarios [2]. Recently, more challenging general indoor scene datasets have proposed [3, 4]. These popular datasets are typically ideal environments for SLAM experiments. Traditional visual SLAM systems built on these datasets perform well in such scenarios, where stable features such as corners, edges are reliable for feature extraction and matching. However, they will experience a series of challenges when applied to construction environments. A construction site is usually defined as low textural, structural, and with large surfaces. Figure 1 provides real examples for illustrating the working conditions of construction robots. The extreme setting results in poor performance and inaccurate localization of traditional feature point based SLAM algorithms. We directly evaluate ORBSLAM [5], a widely used SLAM algorithm known for its effectiveness in general localization tasks, as shown in Figure 2. In Figure 2a, although the system detects some feature points only on a relatively rough wall, it still estimates its pose. However in Figure 2b, it can not detect any feature from the image, therefore the system drifts with imu and fails soon.

This indicates that SLAM algorithms developed based on general datasets are unsuitable for real construction sites. The primary reason for this limitation is the lack of specialized SLAM datasets tailored to construction environments, which hinders the advancement of SLAM algorithms designed for construction robots. To address these issues, this paper introduces a novel dataset specifically created to capture the complexities of construction settings. This dataset serves as a comprehensive benchmark for training and evaluating SLAM algorithms, enabling them to better tackle the challenges inherent in real-world construction environments.



(a) Previous frame (b) Current frame

Figure 2: ORB feature extraction in our construction dataset(From ORBSLAM3).

2 Dataset and Methodology

2.1 Overview

The dataset was collected in indoor construction sites with various types of walls, ceilings, and other architectural characteristics typical of construction environments, such as unfinished structures, exposed pipes, and scattered materials. These scenes are complex and diverse, containing numerous geometric features and dynamic elements such as construction workers and moving equipment. This setting provides a unique challenge for evaluating visual SLAM algorithms.

For the core visual images in our dataset, we chose to capture RGB-D images instead of monocular or stereo images, which could directly provide depth information for each pixel. We have compared these cameras and finally make the decision for some reasons: First, in real-world navigation tasks, construction robots normally require absolute scales for accurate positioning and mapping, while the monocular slam cannot provide this absolute scale, making it unsuitable for reliable navigation in practical environment. Second, visual-inertial systems heavily depend on the Inertial Measurement Unit (IMU) to obtain scale. However, for ground robots typically move at a constant velocity on a 2D plane, the IMU will not be fully excited in each direction, therefore there will be barely no useful measurements. This can lead to divergence, large drift error and scale drift. Third, though stereo vision estimates the depth by matching features between the two camera images, it faces challenges in low-texture or changing lighting conditions commonly found in construction environments, making it difficult for stereo cameras to recover reliable scale and depth.

Finally, our robust solution is to provide an RGB-D camera for the visual SLAM benchmark. The usage of this kind of cameras could help reducing computational load and improving data processing efficiency, and focusing on the development of new visual SLAM algorithms. We have compared our dataset with several current construction slam datasets, the comparison is shown in Table 1.

Dataset	Motion	Visual Sensor Type	LiDAR Sensor
Hilti (Helmberger et al., 2022)	Handheld	Stereo/Grayscale	OS0-64/Livox MID70
ConSLAM (Trzeciak et al., 2023)	Handheld	RGB/NIR	Velodyne VIP-16
Ours	Mounted	RGB-D	OS0-128

Table 1: Comparison of current construction SLAM datasets

We noticed that we are the first that provide RGBD images. The Hilti dataset [6] might be the most well-known construction SLAM dataset, but it only collects 10hz stereo images using Alphasense cameras, due to the limitation by the image resolution. The ConSLAM dataset [7] is primarily a lidar slam dataset, so it only uses a simple RGB camera. Our dataset will be an ideal choice to test and develop visual slam in real construction sites.

2.2 Ground Truth

It is challenging to generate accurate reference positions for SLAM performance evaluation in indoor construction sites. GPS is often unreliable indoors due to weak signals and multipath effects, failing to provide the required precision. Some research employs Total Stations or Terrestrial Laser Scanners (TLS) to generate highly accurate ground truth data [6, 7]. However, these methods are frequently constrained by the complexity and slowness of operation, making them less suitable for real-time and changing indoor settings.

While some research has tried to use advanced lidar slam as the ground truth. TIERS [8] provides a reference location from lidar slam when motion capture system is unavailable in largescale environment. ConPR [9] fuses lidar slam and GPS measurement to generate global absolute trajectories for place recognition. Therefore, high-precision lidar systems can generate stable and accurate positioning data. In our study, we adopt a similar approach to directly obtain ground truth. An advanced lidar-inertial odometry Fast-LIO2 [10] is selected, with some modifications to ensure compatibility with our lidar. But users are allowed to use their own lidar slam algorithm as reference. Figure 3 shows the mapping and localization result after running the lidar slam in our lab, where the colored lines in the pictures represent the ground truth trajectories.

2.3 Hardware

Data collection was conducted using an Intel RealSense L515 camera, an Ouster OS0-128 LiDAR sensor, and both equipped with their built-in imu modules. The Ouster OS0-128 is a spinning high precision LiDAR sensor that offers a full 360-degree horizontal field of view and a 90-degree vertical range, making it ideal for comprehensive scene capture. With a range of up to 50 meters and a 10 Hz operating rate, this sensor effectively detects distant objects and provides accurate data for SLAM applications, aided by its integrated IMU for real-time positioning. The Intel RealSense L515 detects highly precise depth measurements up to 9 meters with a field of view of $70^\circ \times 55^\circ$, tailored for detailed indoor applications. Its higher capture rate of 30 Hz enables it to track rapid changes in dynamic environments. Additionally, the RGBD camera is also regarded as a solid-state lidar, due to its extremely accurate depth measurement. However, there is some debate about whether its use falls under the category of visual SLAM [8, 11]. The general consensus is that if the SLAM algorithm primarily relies on RGB images for visual feature extraction and uses depth images as supplementary input, it is still considered as visual RGBD SLAM [12]. On the other hand, methods that directly use point clouds generated from the depth images for localization are typically regarded as LiDAR SLAM [13]. Since the proposed dataset primarily utilizes RGB images and depth images, it aligns with the definition of visual SLAM and our dataset is categorized as a visual SLAM dataset. The specific configurations of the sensors are listed in Table 2.

The data is recording through a laptop, using Robot Operating System (ROS) with Ubuntu 20.04. The camera is placed on a tripod for stability(Figure 4a). The lidar, as shown in Figure 4b, is powered by an external power source and is connected to the laptop via an Ethernet cable.

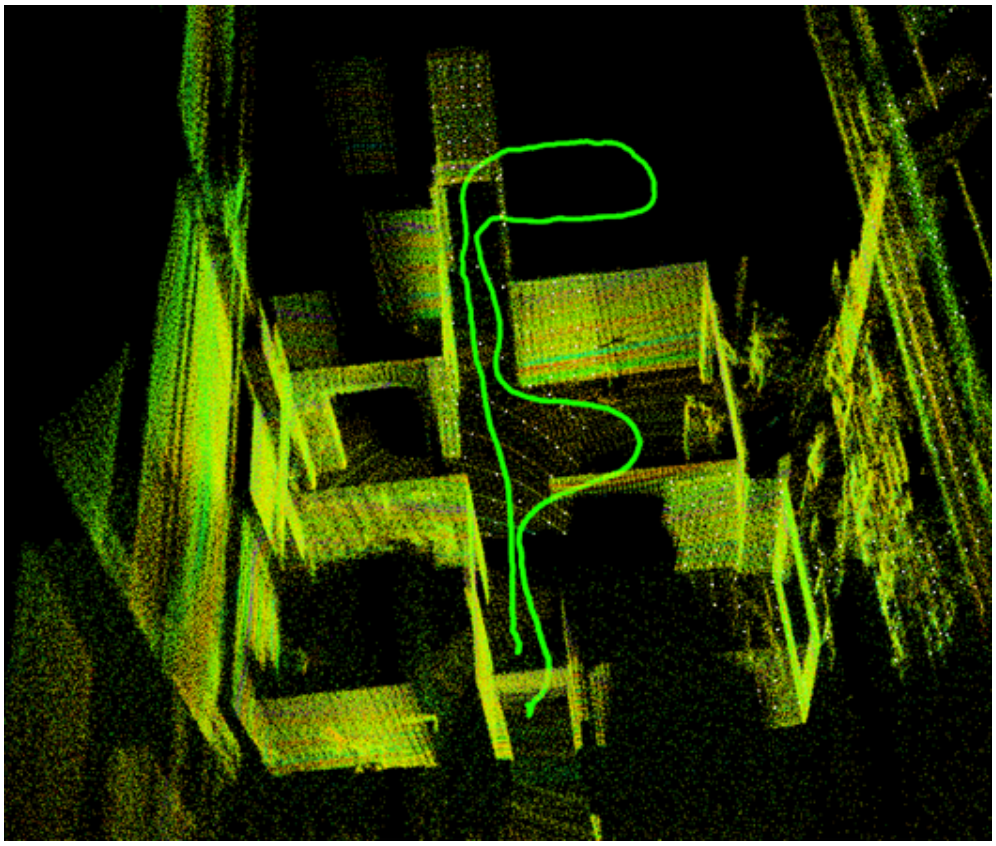


Figure 3: Results of the lidar slam on a test sequence

Sensor	Type	IMU	FoV	Range	Rate
OS0-128	spinning	built-in	$360° \times 90°$	50 m	10hz
RealSense L515	solid-state/RGBD	built-in	$70° \times 55°$	9 m	30hz

Table 2: Sensor configurations on the platform

All devices are mounted on a manually operated platform, which allows for flexible movement and data collection.

2.4 Sensor Data and Software

Figure 5 is an example of sensor data in one scanning. The above two images are RGB and depth image respectively. Each pixel in the depth image indicates the distance of the corresponding pixel in the RGB image. The red pointcloud is from OS0-128 lidar. It is used to register current scan to the pointcloud map. The colored pointclouds are generated by the L515 camera. Benefitting from the RGB-D mode, it is convenient to obtain the rgb pointclouds or receive depth images aligned with the RGB images. Since RGB pointclouds can be recovered from depth images, and the transmission rate of images is also much faster than that of pointclouds, we chose to use depth images rather than directly providing colored pointclouds in our dataset.

(a) RGBD camera (b) LiDAR sensor

Figure 4: Sensor configuration and data collection platform.

In Figure 5 the colored pointclouds are provided by some third-party utilities for visualization.

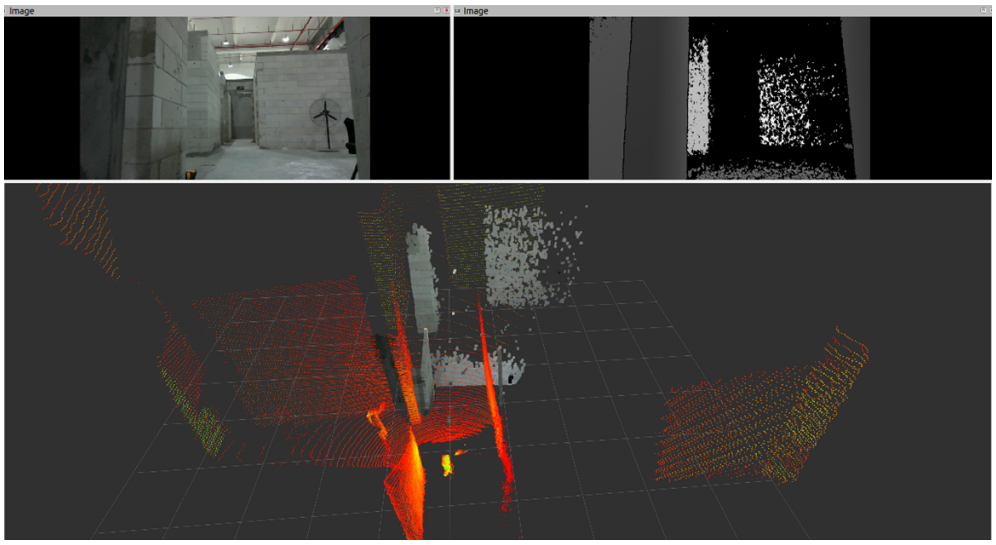


Figure 5: Sensor data of lidar and RGBD camera.

To transform the depth image to pointcloud, we can use the camera intrinsics(given in /camera/color/camera_info topic, as listed in Figure 5) by Equation 1:

$$\begin{aligned} x &= \frac{(j - cx)z}{fx} \\ y &= \frac{(i - cy)z}{fy} \\ z &= depth\_value * depth\_factor \end{aligned} \tag{1}$$

where the $depth\_value$ is from the depth image at (i, j); the $depth\_factor$ is set to 0.001 for this camera; fx, fy, cx, cy are the camera intrinsics provided by the realsense driver; x, y, z are the 3D coordinates respectively.

Some RGBD SLAM directly use pointcloud for Iterative Closest Point (ICP) matching rather than operate on the depth image [12, 13, 14], and shows impressive performances. However, due to limitations in on board computation, they need to downsample the depth maps multi threading parallelization before further proceeding.

In this dataset we provide a fast implementation for converting the depth image, which is mainly realized through the utilization of OpenCV cv::Mat class and continuous memory read/write operations. Compared with OpenCV official implementation, our approach allows additional channels like RGB or inverse depth, and supports the conversion from cv::Mat to the PointCloud2 message format in ROS. Additionally, it can be convert to PCL point clouds with further customization. This enhanced flexibility and efficiency make our dataset well-suited for real-time SLAM applications.

The sensor data is recorded in ROS bag format using the official ROS drivers with their default synchronization schemes enabled. For the RealSense camera, IMU data is first aligned to the image messages by matching the closest timestamps. Since the point cloud and images are captured by two different sensors, their timestamps are not perfectly matched, resulting in some misalignment, as shown in Figure 6. However, the RGB and depth images are synchronized in practice when parsing the data. It is just caused by timestamp precision. The Ouster LiDAR is set to the same time settings as the camera and is synchronized with its built-in IMU. As the LiDAR data is used as the ground truth, we did not align the LiDAR points with the image data. After processing the entire ROS bag and saving the estimated poses, the trajectories can be matched and evaluated for localization performance using public tools like EVO [15].

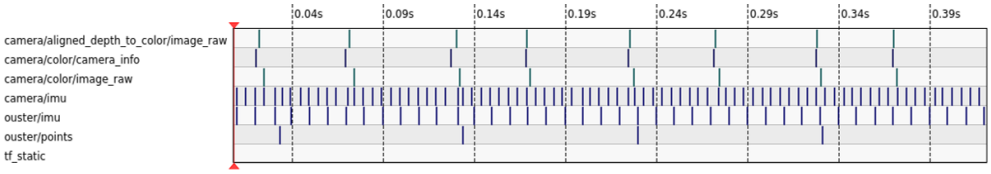
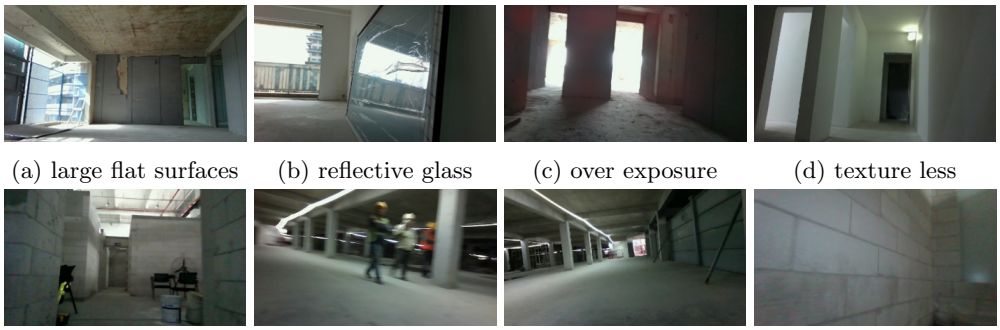


Figure 6: Rosbag information in rqt bag.

Figure 7 are some selected representative pictures from the dataset, which reflect real-world ongoing construction conditions. The motivation for collecting these scenes comes from recent SLAM research specifically focused for construction environments[16, 17, 18, 19]. We capture the diversity of construction settings, including various stages of building materials and structures, such as concrete, brick, plaster, and putty, which often have low or nearly no texture. Some sequence may also involve other auxiliary materials, equipment, workers and machinery. These factors contribute to a more realistic and challenging testing environment for SLAM algorithms.

Table 3 provides a list of the selected sequences from the raw collected data. Most sequences are captured in static environments to simulate a construction robot operating independently. We have also included a dynamic sequence (Floor1) to reflect the presence of workers around the robot. Additionally, our dataset includes a closed-loop sequence(Building B1). However, it should be noted that the robot's localization must rely on real-time odometry, and the impact of loop closure correction is limited.



(a) large flat surfaces (b) reflective glass (c) over exposure (d) texture less

(e) brick wall (f) motion blur (g) open and empty (h) repetitive texture

Figure 7: Representative example frames in our construction dataset.

Sequence	Duration	Description
Building A1, static	85s	Texture-less, multiple rooms
Building B1, static	78s	Repetitive texture, looped trajectory
Floor14, static2	56s	Over-exposure
Floor4, static1	35s	Glass, static workers
Floor4, static3	24s	Texture-less, single room
Floor1, dynamic	55s	Motion blur, dynamic workers

Table 3: Selected sequences in our dataset.

3 Benchmark Results and Discussion

We evaluated several visual SLAM systems with different frontends on our dataset, as illustrated in Figure 8a. The "Fast" frontend refers to the fast corner detection combined with optical flow tracking, implemented by DVINS [20]. VINS-MONO [21] utilizes Shi-Tomasi (S-T) corner points with optical flow tracking, but for this experiment, we selected the RGB-D version of VINS [22]. ORBSLAM [5] was also tested but failed to complete the entire trajectory, as shown in Figure 2. It struggled to detect features when the robot approached a gray wall (Figure 8b) closely, leading to totally drift. This demonstrates the difficulties feature matching based methods in construction environments with low texture or flat surfaces.

The mentioned visual SLAM algorithms are highly representative methods and are widely regarded as benchmarks by most researchers. However, our simple tests revealed that VINS-based improved methods demonstrate greater robustness. Therefore, we recommend using the VINS-based visual SLAM algorithm as a benchmark for our dataset. For visualization, we utilize RTABMAP [23] as the dense mapping backend, replacing the default visual odometry with VINS RGB-D [22]. RTABMAP is particularly known for its versatility and modular design, making it a popular choice for both research and practical applications. In our experiment, it is used for mapping and visualization to support different visual SLAM algorithms.

The final results of mapping and localization are shown in Figure 9. Figure 9a illustrates the outcome of a single run, where the thin blue line represents the estimated trajectory. Figure 9b shows the results from different runs in the same environment, with the point cloud map merging each sub-map across sessions. The differently colored lines correspond to individual robot paths, and the node points indicate keyframe poses from the visual odometry. This experiment aims to provide a simple benchmark and a visualization evaluation based on the

(a) Localization results (b) The corresponding area to the red circle

Figure 8: A comparison between several visual frontends.

collected dataset.

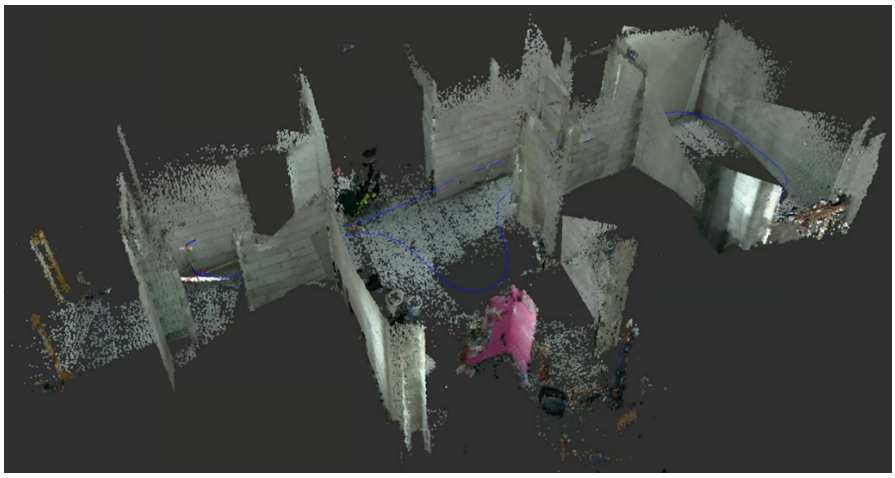
The mapping results in Figure 9a and Figure 9b are not perfect, with some mismatches observed, particularly in the multiple mapping results shown in Figure 9b. This highlights the challenges faced by current visual localization and mapping methods in real construction environments. Future work could focus on developing more robust SLAM algorithms that better leverage prior knowledge, such as 2D drawing, BIM models and structure information to enhance accuracy and reliability in these challenging construction scenarios.

4 Conclusions

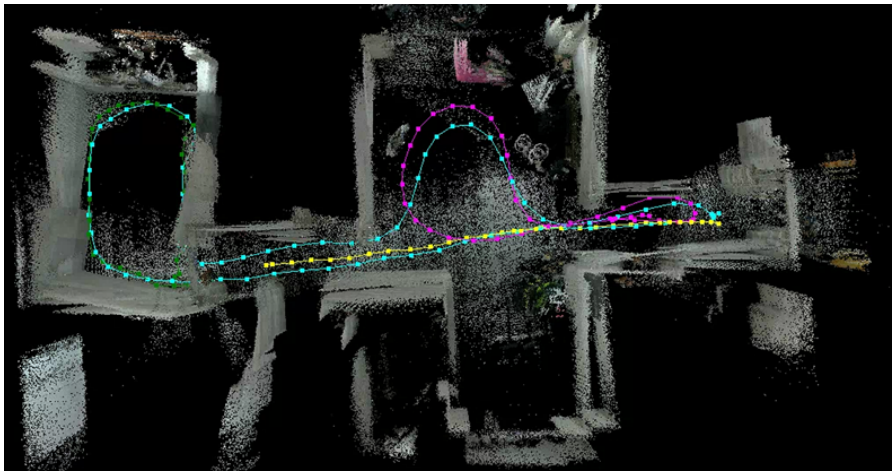
This study introduces an RGB-D visual SLAM dataset specifically collected in real construction sites, which is often overlooked in existing benchmarks. It uniquely includes the dynamic and complex nature of real construction environments, which pose significant challenges for traditional SLAM algorithms. By offering detailed environmental data and realistic settings, we hope this dataset will be a useful resource for construction automation and robotics community, and can facilitate advancements in SLAM localization systems tailored for construction sites.

5 Acknowledgments

This work was supported by the HKUST-BDR Joint Research Institute (Grant No. OKT23EG01), the HKUST Bridge Gap Fund (Grant No. BGF.012.2022), the Collaborative Research Fund (C6044-23GF) supported by the University Grants Committee of the Government of the Hong Kong Special Administrative Region, and the 30 for 30 Scheme (3030007) funded by The Hong Kong University of Science and Technology.



(a) Result on one sequence in our dataset



(b) Results from multiple experiments within the same room

Figure 9: Mapping and localization results by combining VINS RGBD and RTABMAP.

References

[1] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 573–580. IEEE, 2012.

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

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

[4] Jie Yin, Ang Li, Tao Li, Wenxian Yu, and Danping Zou. M2dgr: A multi-sensor and multi-scenario slam dataset for ground robots. *IEEE Robotics and Automation Letters*, 7(2):2266–2273, 2021.

[5] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José MM Montiel, and Juan D. Tardós. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE Transactions on Robotics*, 2021.

[6] Michael Helmberger, Kristian Morin, Beda Berner, Nitish Kumar, Giovanni Cioffi, and Davide Scaramuzza. The Hilti SLAM Challenge Dataset. *IEEE Robotics and Automation Letters*, 7(3):7518–7525, July 2022.

[7] Maciej Trzeciak, Kacper Pluta, Yasmin Fathy, Lucio Alcalde, Stanley Chee, Antony Bromley, Ioannis Brilakis, and Pierre Alliez. ConSLAM: Construction data set for SLAM. *Journal of Computing in Civil Engineering*, 37(3):04023009, 2023.

[8] Li Qingqing, Yu Xianjia, Jorge Pena Queralta, and Tomi Westerlund. Multi-modal lidar dataset for benchmarking general-purpose localization and mapping algorithms. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3837–3844. IEEE, 2022.

[9] Dongjae LEE, Minwoo Jung, and Ayoung Kim. ConPR: Ongoing construction site dataset for place recognition. In *IROS 2023 Workshop on Closing the Loop on Localization: What Are We Localizing for, and How Does That Shape Everything We Should Do?*, 2023.

[10] Wei Xu, Yixi Cai, Dongjiao He, Jiarong Lin, and Fu Zhang. Fast-lio2: Fast direct lidar-inertial odometry. *arXiv preprint arXiv:2107.06829*, 2021.

[11] Peize Li, Kaiwen Cai, Muhamad Risqi U. Saputra, Zhuangzhuang Dai, and Chris Xiaoxuan Lu. OdomBeyondVision: An Indoor Multi-modal Multi-platform Odometry Dataset Beyond the Visible Spectrum. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3845–3850, October 2022.

[12] Jie Xu, Ruifeng Li, Song Huang, Xiongwei Zhao, Shuxin Qiu, Zhijun Chen, and Lijun Zhao. R2DIO: A Robust and Real-Time Depth-Inertial Odometry Leveraging Multi-Modal Constraints for Challenging Environments. *IEEE Transactions on Instrumentation and Measurement*, pages 1–1, 2023.

[13] Han Wang, Chen Wang, and Lihua Xie. Lightweight 3-D Localization and Mapping for Solid-State LiDAR. *IEEE Robotics and Automation Letters*, 6(2):1801–1807, April 2021.

[14] Pengfei Gu and Ziyang Meng. S-VIO: Exploiting Structural Constraints for RGB-D Visual Inertial Odometry. *IEEE Robotics and Automation Letters*, 8(6):3542–3549, June 2023.

[15] Henri Rebecq, Timo Horstschaefer, Guillermo Gallego, and Davide Scaramuzza. EVO: A Geometric Approach to Event-Based 6-DOF Parallel Tracking and Mapping in Real Time. *IEEE Robotics and Automation Letters*, 2(2):593–600, April 2017.

[16] Xinyu Chen and Yantao Yu. HLE-SLAM: SLAM for overexposed construction environment. In *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, volume 40, pages 585–588. IAARC Publications, 2023.

[17] Xinyu Chen and Yantao Yu. An unsupervised low-light image enhancement method for improving V-SLAM localization in uneven low-light construction sites. *Automation in Construction*, 162:105404, June 2024.

[18] Andrew Yarovoi and Yong Kwon Cho. Review of simultaneous localization and mapping (SLAM) for construction robotics applications. *Automation in Construction*, 162:105344, June 2024.

[19] Liu Yang and Hubo Cai. Enhanced visual SLAM for construction robots by efficient integration of dynamic object segmentation and scene semantics. *Advanced Engineering Informatics*, 59:102313, 2024.

[20] Jianheng Liu, Xuanfu Li, Yueqian Liu, and Haoyao Chen. RGB-D Inertial Odometry for a Resource-Restricted Robot in Dynamic Environments. *IEEE Robotics and Automation Letters*, 7(4):9573–9580, October 2022.

[21] Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, August 2018.

[22] Zeyong Shan, Ruijian Li, and Sören Schwertfeger. RGBD-Inertial Trajectory Estimation and Mapping for Ground Robots. *Sensors*, 19(10):2251, January 2019.

[23] Mathieu Labbé and François Michaud. RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. *Journal of Field Robotics*, 36(2):416–446, 2019.