Hilti-Trimble-Oxford Dataset: 360 Visual-Inertial Benchmark with Floor Plan Priors for SLAM and Localization

Samuele Centanni[1*], Yuhao Zhang[2*], Yifu Tao[2], Julien Kindle[1,3], Frank Neuhaus[4], Tilman Koß[4], Aryaman Patel[5], Michael Helmberger[1], Emilia Szymańska[1], Torben Gräber[1], and Maurice Fallon[2]

***Abstract*—Automated progress monitoring on construction sites is an active area of research and development. Robot and human-carried mapping systems have been developed to build 3D maps of building and infrastructure projects. While LiDAR-based mapping systems achieve high accuracy, the cost of LiDAR can be prohibitive. Consumer-grade cameras with wide field of view (*"360 cameras"*) combined with embedded inertial measurement units (IMUs) provide a cost-effective alternative. To support change detection and progress monitoring, highly accurate visual simultaneous localization and mapping (SLAM) and floor plan localization systems are required. In this paper, we present a high-quality dataset collected at an active construction site, which captures realistic challenges such as variable lighting conditions, moving workers, fast motions, and repetitive structures. The dataset offers thirty visual-inertial sequences recorded across ten floors over an eight-month period of the construction project. Ground truth trajectories are collected using a high quality LiDAR-inertial SLAM system rigidly attached to the 360 camera. Additionally, we report the results of an open research challenge evaluating the best visual SLAM and localization systems from around the world. The Challenge attracted substantially higher participation in SLAM, with 62 teams compared to 22 in floor plan localization, which suggests that this task receives less attention from researchers and motivates further exploration through proposed dataset and challenge.**

***Index Terms*—Data Sets for SLAM, Robotics and Automation in Construction, Omnidirectional Vision.**

I. Introduction

THE buildings and construction industry accounts for approximately 11–13 % of global GDP [1]. It is, however, a sector that has seen minimal improvements to productivity, in part due to the limited digitization of work processes, rework due to defects and quality deviations, occupational safety hazards, and demographic changes in the construction workforce [2]. These challenges motivate automated methods for tracking the progress of construction to provide site managers with timely insights into the state of the project.

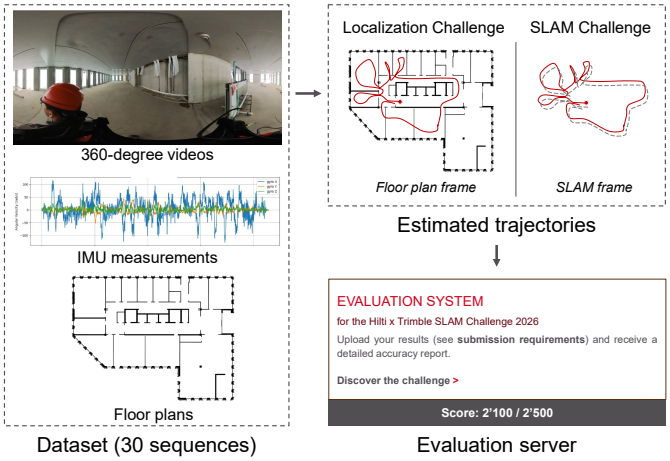


Fig. 1: The Hilti-Trimble-Oxford Dataset, consisting of 360-degree videos, IMU measurements and floor plans, allows for benchmarking algorithms for SLAM and floor-plan-referenced localization.

A common approach to construction site monitoring is the regular mapping of the environment using a variety of sensors such as LiDAR, RGB cameras, or terrestrial laser scanners [3], [4]. These sensors may be operated manually or mounted on robotic platforms, including legged robots, drones, and wheeled platforms [5], [6]. The usefulness of the resulting maps depends on the extent to which they can support change detection across consecutive mapping sessions or comparison against architectural models [7]. Detecting errors or deviations early can lead to substantial savings in both cost and time. The maps can also serve as documentary evidence of the building project [8].

The core technology allowing for the generation of such maps is simultaneous localization and mapping (SLAM), which is an area of active research in robotics and computer vision [9]. While LiDAR sensors remain widely used for SLAM, color cameras provide a richer understanding of the environment. Due to their low cost, wide availability, and easy

Manuscript received: July 7, 2026; Revised: September 22, 2026; Accepted: October 11, 2026. This paper was recommended for publication by Editor Hyun Myung upon evaluation of the Associate Editor and Reviewers' comments. *(*Samuele Centanni and Yuhao Zhang contributed equally to this work.) (Corresponding author: Emilia Szymańska.)*

[1]Samuele Centanni, Julien Kindle, Michael Helmberger, Emilia Szymańska, and Torben Gräber are with Hilti AG, Corporate Research & Technology, Schaan, Liechtenstein (e-mail: emilia.szymanska@hilti.com).

[2]Yuhao Zhang, Yifu Tao, and Maurice Fallon are with Dynamic Robot Systems Group, University of Oxford, Oxford, U.K. (e-mail: yuhao@robots.ox.ac.uk).

[3]Julien Kindle is also with Robotics Systems Lab, ETH Zürich, Zürich, Switzerland (e-mail: jkindle@ethz.ch).

[4]Frank Neuhaus and Tilman Koß are with Vision & Robotics GmbH, Koblenz, Germany (e-mail: frank.neuhaus@vision-robotics.de).

[5]Aryaman Patel is with Trimble Inc., Denver, USA (e-mail: aryaman_patel@trimble.com).

Dataset and benchmark: https://hilti-trimble-challenge.com

Digital Object Identifier (DOI): see top of this page.

© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

deployment, passive cameras are increasingly used as the sole sensor for data collection. In particular, 360 cameras provide a full field-of-view coverage, which enables efficient environment mapping. Additionally, many 360 cameras incorporate inertial measurements, which can support gravity alignment and scale estimation.

However, maps and trajectories obtained from different data-collection sessions are not sufficient on their own. For longitudinal monitoring, they need to be registered to a common reference frame, such as the construction site plan. Prior work focuses localization within complete 3D building representations [10], though building information modeling (BIM) models are not universally available, and tend to be large and complex to process. In contrast, 2D floor plans are widely adopted, compact, and still encode important geometric and semantic information about the environment. This makes floor plan-based localization a practical and relevant problem for construction-site monitoring.

This setting highlights the need for benchmark datasets that evaluate the spatial localization capabilities underpinning construction site monitoring, including visual-inertial SLAM and localization with respect to 2D floor plans in realistic, evolving construction environments. To address this gap, we introduce the Hilti-Trimble-Oxford Dataset and the corresponding Hilti × Trimble SLAM Challenge 2026, which provide data, ground truth, and an evaluation protocol for benchmarking trajectory estimation methods, as shown in Fig. 1.

The main contributions of this work are:

- The Hilti-Trimble-Oxford Dataset of 30 challenging sequences of 360° field of view image streams with synchronized IMU measurements, captured over a span of 8 months at the same construction site, together with the corresponding 2D floor plans.
- Ground truth trajectories collected with a high-accuracy LiDAR-inertial system which are released alongside the dataset.
- An evaluation system for benchmarking algorithms from two categories, where participants estimate the trajectory of the 360 camera: (i) in an arbitrary reference frame (**SLAM** category) (ii) in the floor plan's reference frame (**Localization** category).
- The discussion of the results of the Hilti × Trimble SLAM Challenge 2026, which received submissions from 22 teams in the Localization category and 62 teams in the SLAM category, demonstrating substantial interest in the proposed benchmark.

II. Related Work

Research into visual SLAM has progressed hand in hand with the availability of high quality datasets. Accordingly, we review datasets for visual SLAM and localization before turning to datasets for construction monitoring.

A. Datasets for Visual SLAM and Localization

Common visual SLAM datasets exist for autonomous vehicles [11], aerial robots [12], and legged robots [13]. The manner in which the sensors are mounted on the platforms and how they move has a significant impact on the usefulness of any dataset. Early autonomous vehicle and field robotics datasets, such as KITTI [11], mostly contain forward translation with low rotation rates and larger distances to nearby objects. With the advent of body-mounted or egocentric sensing systems, there are new challenges for visual SLAM systems to solve, notably the high rotation rates which occur when a person moves their head abruptly. Subsequent datasets have included inertial data and were used to develop visual-inertial odometry systems [12], [14]. More recent datasets such as LaMAR [15] target egocentric scenarios, while LaMAria [16] extends this approach to city-scale large environments.

For evaluation of visual SLAM systems, ground truth poses can be obtained from GNSS (especially in autonomous driving [11]) or from motion capture systems in indoor environments [17]. Control points have also been used to provide precise but sparse ground truth locations [18], [16]. In addition to ground truth trajectories, there are datasets that also provide a ground truth 3D map to evaluate the SLAM reconstruction. Datasets including Matterport3D [19], ScanNet [20] and LaMAR [15] use the output from an accurate SLAM system as the ground truth map. Alternatively, a survey-grade laser scanner can produce more precise 3D maps, and is used to provide ground truth maps in datasets including ETH3D [21], ScanNet++ [22], FusionPortableV2 [13] and Oxford Spires [23].

Visual localization or place recognition (in a prior map) is a common and related problem. Early place recognition datasets such as 7-Scenes [24] were limited to small static scenes. Later efforts pushed the scale of viewpoint, illumination and structural change. InLoc [25] features textureless indoor areas while Aachen Day-Night [26] has sequences captured with extreme variations in image appearance. However, these approaches localize in prior maps of visual features as opposed to localizing directly within an architectural floor plan, which is the focus of the Hilti-Trimble-Oxford Dataset.

B. Datasets Specific to Construction Monitoring

Despite the emergence of a wide range of technologies for digitizing construction project monitoring, there are few high quality public SLAM datasets available. This is in part because site owners and contractors prefer to retain confidentiality on their projects. The following datasets therefore provide important benchmarks, but differ in terms of sensing modalities, prior information, and the type of localization or mapping task they support.

ConSLAM is a dataset of five sequences collected with a portable LiDAR scanner on the same part of a construction site over an extended period [27], which supports multi-session LiDAR SLAM. However, its limited spatial coverage and focus on LiDAR sensing make it less suitable for evaluating low-cost visual localization approaches.

Hilti Research has organized a series of SLAM Challenges focused on benchmarking performance in realistic construction environments. The initial Hilti SLAM Challenge (2021) [28] introduced a multi-sensor dataset combining visual, LiDAR, and inertial measurements with millimeter-accurate ground

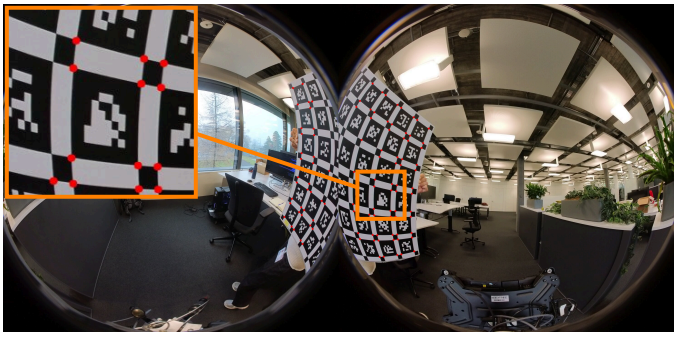


Fig. 2: Visualization of the multi-camera calibration. Extrinsic constraints are obtained when the AprilTag board is observed by two cameras simultaneously. The red markers indicate the forward-projected image-plane locations of the extracted calibration target corners under the calibrated cameras.

truth, highlighting the importance of sensor fusion for robustness in real-world scenarios. Follow-on Challenges in 2022 and 2023 extended the benchmark to more diverse environments, sparse survey-grade ground truth, multi-device setups, and multi-session SLAM [18], [29]. Across these Challenges, reported results showed strong performance from LiDAR-based methods, including sub-centimeter drift in some 2022 sequences, while also revealing that multi-session capabilities remain comparatively less mature. The associated datasets, however, lack explicit map priors, which have the potential to improve the reliability of SLAM solutions.

The SLABIM [10] dataset is notably relevant. The authors focus on direct LiDAR localization inside a detailed 3D BIM—in this case the HKUST campus. While industry trends are promoting the use of BIMs, we feel that a useful commercial solution should not rely on such a detailed prior map for localization. The dataset proposed in the Hilti × Trimble SLAM Challenge 2026 instead targets visual-only localization with respect to a minimal 2D floor plan. This combination is more practical due to the low power and cost requirements of the 360 camera we selected.

III. Hardware

The Insta360 ONE RS 1-Inch Edition is selected as the primary sensor for its high-quality omnidirectional imaging and portability. Its compact, self-contained design enables rapid and reliable deployment in confined, cluttered, and unstructured construction and industrial environments.

This device captures front and rear hemispherical images, each providing a field of view (FoV) exceeding $180°$. The camera also integrates a 6-axis inertial measurement unit (IMU) that provides high-frequency motion measurements. These can be used to estimate metric motion and to carry out motion compensation of the rolling shutter images.

A. Calibration

1) Multi-Camera Intrinsic and Extrinsic Calibration: We employ the open-source toolbox Kalibr [30] for intrinsic and extrinsic calibration of two fisheye cameras. Due to the limited visual overlap between the two lenses, standard checkerboard patterns are not suitable, as they require both cameras to simultaneously observe a large, contiguous set of shared inner corners. Instead, we use a 6×6 AprilTag calibration board. We keep the device stationary while moving the board around the field of view of each camera to ensure spatial coverage and feature diversity, as shown in Fig. 2. We move the calibration board slowly to ensure that the rolling shutter effect is negligible.

To achieve stable and accurate calibration under wide-angle imaging conditions, we adopt the enhanced unified camera model (EUCM) [31]. Since many SLAM systems support the Kannala-Brandt four-parameter model (KB4) [32], we also provide KB4 parameters fitted to the calibrated EUCM intrinsics. Fig. 2 shows an example of the reprojection of the extracted calibration target corners using the calibrated intrinsics and extrinsics.

2) Rolling Shutter Parameter: We obtain the rolling-shutter frame readout time directly from the camera metadata and validate it using a high-frequency blinking LED, where the spacing of the induced horizontal brightness bands provides an independent estimate of the sensor readout time.

3) IMU Calibration: We estimate the IMU noise characteristics through Allan variance analysis. Specifically, we collected approximately three hours of stationary IMU data to compute the noise density and random walk parameters for both the accelerometer and the gyroscope.

4) Camera-IMU Extrinsic Calibration: Given the calibrated camera and IMU noise parameters, we perform camera–IMU extrinsic calibration using Kalibr with a recording that includes rich translational motion and rotational excitation about the yaw, pitch, and roll axes to ensure full observability of all extrinsic parameters.

IV. Dataset

The dataset spans an eight-month period at a construction site in Buchs, Switzerland. This long collection period captures the site's structural evolution, offering a novel benchmark for long-term mapping in dynamic environments. The building footprint is approximately $40\,\mathrm{m} \times 100\,\mathrm{m}$ and spans eight floors as well as two underground parking levels. In its early stages, the environment is characterized by limited texture and repetitive concrete structures; as construction progresses, the scene complexity increases as ducting, fixtures and dividing walls are installed.

We provide the dataset as ROS 2 bag files (`.db3`), a standard logging format for time-stamped sensor streams commonly used in robotics. Raw images and high-frequency IMU measurements are extracted directly from the `.insv` file (Insta360 native format for storing recorded video and associated metadata). Each recording is self-contained and captured on a single floor, enabling modular benchmarking and evaluation. Specifically, each ROS 2 bag records a set of time synchronized data streams ("topic") as summarized in Table I. These topics provide the complete sensor data required for visual-inertial SLAM, including IMU measurements and two fisheye camera streams.

TABLE I: Sensor ROS Topics and Frequencies.

Sensor	Topic name	Rate [Hz]
IMU	`/imu/data_raw/compressed`	1000
Front camera	`/cam0/image_raw/compressed`	30
Rear camera	`/cam1/image_raw/compressed`	30

TABLE II: Characterisation of challenge sequences.

Clutter										✓		✓
Low light					✓	✓		✓	✓	✓	✓	✓
Dynamic init.				✓		✓	✓				✓	✓
Aggr. motion			✓					✓		✓	✓	✓
Poor texture		✓					✓		✓			
# Sequences	**6**	**4**	**2**	**4**	**1**	**2**	**3**	**1**	**3**	**2**	**1**	**1**

Note: Each column represents a unique combination of conditions, with the number of sequences shown in the bottom row. Sequence properties and videos are available on the dataset webpage.

A. Challenge Sequences

The dataset consists of 30 sequences with an average duration of 180 s. We categorize the runs according to five factors that influence sequence difficulty:

1) **Clutter:** Selected runs offer intentionally cluttered environments (e.g. close proximity to walls or moving people), resulting in severe occlusions.
2) **Low-Light:** Sequences include near-dark conditions, testing the limits of feature tracking.
3) **Initialization:** *Static* runs start with a stationary period and gentle excitation for IMU bias estimation. *Dynamic* runs begin in motion, requiring the estimator to initialize under active dynamics.
4) **Aggressive Motion:** Certain runs involve rapid 6-DoF movements, introducing motion blur and stress-testing IMU-based de-skewing. These sequences may require rolling shutter compensation for optimal results.
5) **Poor texture:** Captures contain regions with weak visual structure, reducing the quality of feature tracking.

By combining these factors, the dataset spans diverse operating conditions, as summarized in Table II, ranging from simple cases with static initialization, sufficient illumination, rich visual features, and smooth motion, to challenging cases with dynamic initialization, low-light or near-dark scenes, aggressive 6-DoF motion, and severe visual occlusions.

The sequences intentionally avoid large repeated loops, making loop-closure detection and global map optimization more difficult and emphasizing wide-field-of-view loop closure using both cameras.

B. Floor Plans

We provide anonymized building floor plans as PNG images, resized to a metric resolution of $1\,\mathrm{px} = 1\,\mathrm{cm}$ and simplified to contain only structural elements such as walls and concrete columns. The resulting maps are represented as binary images, where structural elements are encoded as 1 (black) and free space as 0 (white), allowing direct conversion into ROS occupancy maps. Two variants of each floor plan are available: with and without windows included as structural elements. These elements were manually labelled on professional architectural floor plans at a higher resolution,

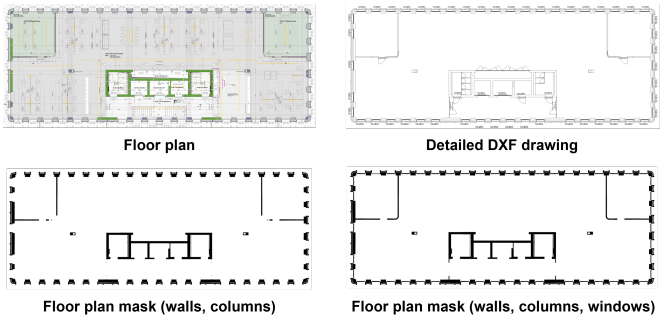


Fig. 3: Types of provided floor plans.

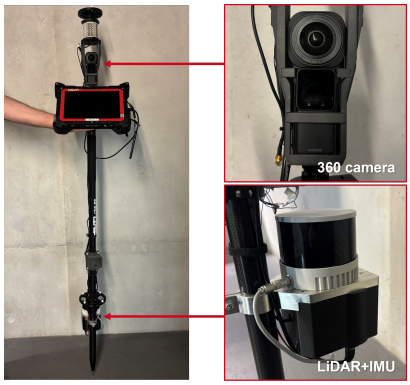


Fig. 4: Data collection platform: rigidly connected V&R LiDAR-inertial SLAM system and Insta360 ONE RS 1-Inch 360 Edition.

with an uncertainty of 1–3 pixels after scaling. Additionally, we provide floor plans in DXF format, which contain more detailed structural information. Although these are not registered with respect to the PNG maps, they offer additional geometric or semantic data. All floor plans' formats available in the dataset are presented in Fig. 3.

Localization in this setting remains challenging, as it requires associating visual observations from the sequences with the simplified 2D structural representation. To reduce ambiguity and shift the focus from global place recognition to local map matching, the initial pose of each sequence within the floor plan is provided.

V. Ground Truth

We obtain ground truth for both camera trajectories and the reconstructed scene geometry using a LiDAR-inertial SLAM system from our industrial partner, rigidly attached to the 360 camera as depicted in Fig. 4. The LiDAR-inertial subpart consists of a Hesai XT32M2X LiDAR sensor and a high-quality, factory-calibrated, tactical-grade IMU, with hardware synchronization between the two components.

A. Temporal Synchronization

While the Insta360 and the LiDAR-inertial system are each internally hardware-synchronized with their own IMU, they are not synchronized with each other. We temporally align the two IMUs by correlating gyroscope magnitudes to initialize the time offset, then refine a linear timestamp mapping $t_L = \alpha t_C + \Delta t$, where α is the clock-scale factor, Δt is the time

offset and t_L, t_C correspond to LiDAR and camera timestamps respectively. The resulting time mapping is used to convert timestamps between the two clock domains.

B. LiDAR–Inertial Reference Trajectory

Ideally, the reference trajectories would be independently measured using survey-grade terrestrial laser scanners or surveyed control points. However, such measurements were not possible due to the pace of the ongoing construction project. We therefore use trajectories estimated by a high-quality LiDAR–inertial SLAM system as the reference. The underlying technology has demonstrated centimeter-level accuracy ($\approx$3.94 cm), and sub-centimeter accuracy (0.7 cm) under favorable conditions, in previous construction-site SLAM Challenges with independently surveyed ground truth [18]. Compared to the consumer-grade IMU in the Insta360 camera, the reference system combines a Hesai LiDAR with a factory-calibrated tactical-grade IMU and performs full offline optimization over all recorded measurements using a continuous-time trajectory representation. Given that the errors of the best visual–inertial submissions in the present benchmark are substantially larger than the expected centimeter-scale error of this reference system, we consider the resulting trajectories sufficiently accurate for the intended evaluation.

C. Camera Ground Truth

To transfer the LiDAR trajectory to the camera frame, we perform the LiDAR–camera extrinsic calibration. Given the calibrated camera intrinsics, an AprilTag grid is moved to multiple poses observed by both sensors. From the camera detections and the known physical dimensions of the printed target, we compute the metric pose of the calibration plane for each placement. Starting from a manually defined initial LiDAR–camera transform, the optimization incorporates two complementary error terms for joint optimization: (i) the reprojection error between detected AprilTag corners and their predicted image locations, and (ii) point-to-plane distances between LiDAR measurements and the observed calibration board, which is assumed to be planar. While we use an internal calibration tool from our industrial partner, we note that a similar open-source tool is available [33].

D. Floor Plan Alignment

We achieve the spatial alignment of the trajectories to the architectural floor plans through a multi-stage registration process. Initially, we perform a coarse alignment by manually rotating and translating the LiDAR point cloud to match the primary structural features of the floor plan. This result serves as the initial seed for a fine-grained refinement using the point-to-point iterative closest point (ICP) algorithm [34], ensuring a high-precision fit between the global LiDAR map and the reference geometry (Fig. 5).

Beyond alignment, this process quantifies the discrepancies between the as-built state and the as-planned architectural models. We observe an average deviation of approximately 2 cm between the LiDAR data and the floor plans. These values reflect the standard structural tolerances and inevitable variances inherent in large-scale, real-world construction environments. Thus, differences of a few centimeters between methods should not be overinterpreted.

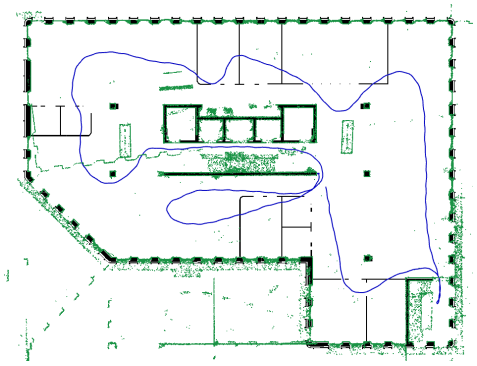


Fig. 5: A ground truth trajectory (**blue**) with LiDAR SLAM map (**green**) overlaid on a floor plan (**black**). The trajectory passes through some as-yet unbuilt walls due to the early construction phase, and some temporary wall-like structures are captured in the LiDAR point cloud. Glass reflections can also be seen.

VI. Challenge Results and Findings

A. Benchmark Overview

The proposed dataset is accompanied by a benchmarking system that evaluates submitted trajectories according to the selected evaluation track:

SLAM: This track follows standard SLAM evaluation. The submitted and ground truth trajectories are first rigidly aligned in a least-squares sense using the Kabsch algorithm [35]. The score is then computed from the resulting 3D absolute trajectory error (ATE).

Localization: This track requires localization with respect to the floor plan. The submitted trajectory has to be expressed in the map frame (without any alignment). In this track, only the 2D ATE in the floor plan plane is evaluated, with the z-coordinate ignored.

To encourage submissions to the benchmark, we organized the Hilti × Trimble SLAM Challenge 2026, which accepted submissions over a period of 3.5 months. After the completion of the Challenge, all ground truth trajectories were made available to support independent evaluation, further algorithm development, and model training.

B. Evaluation Metrics

We required each submission to provide a trajectory with one pose for each camera-frame timestamp. A run is evaluated only if the submitted trajectory achieves at least 99 % coverage of the evaluated ground truth poses, with coverage defined as the ratio of matched predicted poses to ground truth poses; otherwise, the score for that run is set to zero. The initial 5 s of each recording are excluded to ignore any effects during initialization.

TABLE III: Comparison of the top five SLAM teams.

	Team	Base method	View type	Map prior	Real time	Global BA	Loop closure	Same params	Average RMSE [m]	Score
1	KAIST	OKVIS2-X	Fisheye	✗	-	✓	✓	-	0.089	2410.0
2	Inha Univ.	$\sqrt{\text{VINS}}$	Fisheye	✗	✗	✓	✓	-	0.109	2393.9
3	-	OKVIS2-X	Fisheye	✗	✗	✓	✓	✗	0.189	2320.7
4	ICL	OpenVINS + DBoW	Fisheye	✗	✗	✗	✓	✓	0.202	2308.8
5	Univ. of Luxembourg	ORB-SLAM3	Fisheye	✓	-	✓	✓	-	0.224	2299.3

Note: Teams are ranked by their challenge results. The comparison includes the base SLAM method, image view type, use of map priors, real-time capability, global bundle adjustment, loop closure, parameter consistency across sequences, average RMSE, and overall score.

TABLE IV: Comparison of the top five localization teams.

	Team	Trajectory estimation approach	Map alignment approach	Semantic segment.	Average RMSE [m]	Score
1	ETH-THU-Stuttgart	OKVIS2-X	Z-FLoc anchors	✓ [36]	0.238	2196.3
2	Univ. of Luxembourg	ORB-SLAM3 with map priors	BEV-wall-points-to-map ICP in 2D	✓ [37]	0.286	2161.4
3	ETH OmniRecon	OpenVINS + COLMAP	BEV-wall-points-to-map ICP in 2D	✓ [36]	0.399	2041.3
4	Cairo University	ORB-SLAM3	Ray-based wall scoring	✗	0.497	1919.0
5	ETH Map-It Ralph	OpenVINS + COLMAP	Starting-pose-based alignment	✗	1.287	1569.9

Note: Teams are ranked by their challenge results. The comparison includes the trajectory generation method, trajectory-to-map alignment approach, use of video-based semantic segmentation for alignment, average RMSE, and overall score.

For each evaluated trajectory i, we compute an exponentially decaying position score by averaging the per-pose scores over the trajectory:

$$\text{score}_i = \frac{1}{N_i} \sum_{t=1}^{N_i} a e^{-c\, e_t^{(i)}}, \tag{1}$$

$$e_t^{(i)} = \left\| \mathbf{p}_{\mathcal{M},\mathcal{C}}^{(i)}(t) - \hat{\mathbf{p}}_{\mathcal{M},\mathcal{C}}^{(i)}(t) \right\|_2 . \tag{2}$$

Here, N_i denotes the number of evaluated poses in trajectory i; $\mathbf{p}_{\mathcal{M},\mathcal{C}}^{(i)}(t)$ and $\hat{\mathbf{p}}_{\mathcal{M},\mathcal{C}}^{(i)}(t)$ denote the ground truth and predicted camera position in the map frame $\mathcal{M}$; and $e_t^{(i)}$ is the Euclidean position error. For the SLAM track, $e_t^{(i)}$ is the 3D Euclidean distance after rigid alignment. For the Localization track, $e_t^{(i)}$ is the 2D Euclidean distance in the floorplan plane. The constants a and c parameterize the exponential decay and are chosen such that an error of $0\,\text{m}$ corresponds to a score of 100 and an error of $10\,\text{m}$ corresponds to a score of 1, yielding $a = 100$ and $c = \ln(10)/5 \approx 0.4605$. The final score is the sum of the trajectory scores over all runs:

$$\text{finalScore} = \sum_i \text{score}_i. \tag{3}$$

For the official benchmark, both the SLAM and Localization tracks exclude 5 sequences that were made available with ground truth. These sequences are intended solely for algorithm fine-tuning and not used for evaluation. Therefore, the SLAM track is evaluated on 25 trajectories out of 30 in total. The Localization track is evaluated on 24 trajectories, excluding the same 5 runs as SLAM, with one additional sequence omitted due to a missing floor plan. Consequently, the maximum achievable total score is 2500 points for SLAM and 2400 points for Localization. All the aforementioned metrics are computed using the `evo` toolbox [38].

C. Results and Leading Approaches

At the time of the Challenge deadline, the evaluation system has received over 1800 submissions, with 22 unique teams competing in the Localization track and 62 teams in the SLAM track. The results of the Hilti × Trimble SLAM Challenge 2026 were announced publicly and the top performing solutions are shown in Table III and Table IV, with a dash signifying the lack of information provided by the participants. The full leaderboard and detailed technical reports from each team are available on the Challenge webpage.

The winning SLAM submission, ACDC-VSLAM from KAIST, achieved a total score of 2410.0 points. Their system is based on OKVIS2-X [39] and introduces several adaptations tailored to the benchmark. In the frontend, spherical equal-area feature selection improves angular coverage of the fisheye images, while line features provide additional constraints in texture-poor regions. Its local bundle adjustment further re-weights visual constraints based on track reliability and depth. In the backend, a two-stage loop-closure module detects both intra-camera and cross-camera revisits, and adaptively selects between $\text{SE}(3)$, $\text{Sim}(3)$, and 4-DoF corrections according to the quality of each loop candidate. These design choices directly target the wide-FoV distortion, weak texture, and front–rear camera configuration present in the dataset. Other high-performing SLAM submissions use complementary strategies to address similar challenges, including low-light enhancement and dynamic-object masking, point-and-line feature fusion, cross-camera constraints, robust initialization, and global bundle adjustment or factor-graph optimization.

A joint team from ETH, THU, and University of Stuttgart won the Localization track with the score of 2196.3 points. Their approach first estimates camera poses using OKVIS2-X with online extrinsic calibration. They then build upon Z-FLoc [40], which estimates a $\text{Sim}(2)$ transformation from recent image frames. It first uses semantic segmentation to extract structurally stable wall and floor regions, reconstructs

a wall-only 3D point cloud, and projects it into a bird's-eye-view (BEV) map. It then matches extracted line/circular primitives to floor plan primitives with RANSAC and non-linear refinement, and refines the overall trajectory using pose graph optimization. Their solution achieves an average RMSE of 24 cm. Other strong Localization submissions similarly exploit semantically stable wall structures, typically using semantic segmentation followed by BEV-based geometric alignment to the floor plan.

D. Discussion

The top SLAM solutions all rely on visual-inertial odometry, usually using both cameras and the IMU, and followed by a subsequent optimization step (BA). Most are based on established systems such as OKVIS2-X [39], OpenVINS [41], ORB-SLAM3 [42], or $\sqrt{\text{VINS}}$ [43]. Notably, OKVIS2-X was used by several high-performing teams, including the winners of both the SLAM and Localization tracks.

Some teams explicitly reported challenges with initialization, especially for sequences that start in motion, and addressed this through dynamic or delayed initialization strategies. Image preprocessing was also common, including contrast enhancement in dark scenes, static masks, and segmentation to remove the operator or dynamic objects. Beyond classical point features, multiple systems incorporate line features, which are particularly useful in construction environments with texture-poor walls, corridors, and strong structural edges. In the backend, loop closure, pose-graph optimization, and global bundle adjustment are frequently used to reduce drift beyond what local visual-inertial odometry can recover. Particularly, cross-camera loop closure is commonly adopted to improve robustness by exploiting the complementary front and rear cameras, which together provide 360° visual coverage. These results suggest that strong performance is generally associated with combining robust visual-inertial estimation with dataset-specific frontend processing and global trajectory refinement.

To localize within the floorplan, top-performing teams typically detect walls in the images, construct wall-only point clouds, and align their BEV map with floor plans using methods such as Z-FLoc or ICP. This suggests that performance gains can come not only from geometric alignment, but also from selecting stable, semantically meaningful structural elements. However, no team reported improving alignment by detecting windows or other structural components other than walls that are present in detailed drawings.

Sequences recorded in the underground floor consistently challenged the top teams in both the localization and SLAM tracks, as shown in Fig. 6. We attribute this to the combination of low-light and texture-poor conditions, larger open spaces, and longer trajectory lengths.

E. Known Issues

There are a couple of practical considerations when using this dataset. The most important issue is the accuracy or concurrency of the provided floor plans. The bitmap floor plans were derived from a single version of the architectural



Fig. 6: Average RMSE with standard deviations of the top three teams for each sequence in the SLAM and Localization categories. Bar color indicates error magnitude, transitioning from green to red as error increases. The coding for run names is floor, date, and numbered run on that date.

floor plans. The building project was in constant progress with certain elements of the floor plan (cubicles, partitions or ducting) not yet present during the early data recordings. This means that the Challenge required localization to a similar but different scene—with only some of the features usable for matching. We feel this is a realistic challenge because up-to-the-minute floor plans would not be feasible in active construction environments.

Separately from this, a few deviations or errors in the floor plan were identified when comparing them to our LiDAR-based ground truth point clouds (e.g., partitions installed with a slightly different length). This may lead to small residual errors in our ground truth pipeline. These effects are minor with respect to the intended evaluation scale and do not affect the overall validity of the benchmark.

VII. Conclusion

The Hilti-Trimble-Oxford Dataset provides a construction-focused visual-inertial state estimation benchmark with prior 2D maps. Using the Hilti × Trimble SLAM Challenge 2026, we evaluated SLAM and localization submissions from teams worldwide. The top team in the Localization category achieved an average error below 24 cm, with increased performance observed among solutions that incorporate semantic scene information from video frames. The SLAM category attracted nearly three times as many teams, demonstrating high interest, with the lowest reported average error reaching 9 cm.

Beyond the original Challenge sequences, we have also released the ground truth trajectories and evaluation code. We hope that the Hilti-Trimble-Oxford Dataset and benchmark will support further progress in SLAM and localization research and contribute to more effective construction-site monitoring.

References

[1] United Nations Environment Programme and Global Alliance for Buildings and Construction, "Global status report for buildings and

construction 2025–2026," United Nations Environment Programme, Tech. Rep., 2026, accessed 7 July 2026. [Online]. Available: www.unep.org/resources/report/global-status-report-buildings-and-construction-2025-2026

[2] I. Armeni, J. Chen, C. Feng, I. Jeong, C.-J. Liang, K. Morin, N. Tsagarakis, X. Wang, and L. Zhang, "Construction robotics and automation [TC spot light]," *IEEE Robot. Autom. Mag.*, vol. 31, no. 4, pp. 186–192, 2024.

[3] F. Bosche, C. T. Haas, and B. Akinci, "Automated recognition of 3D CAD objects in site laser scans for project 3D status visualization and performance control," *J. Comput. Civ. Eng.*, vol. 23, no. 6, pp. 311–318, 2009.

[4] Y. Turkan, F. Bosche, C. T. Haas, and R. Haas, "Automated progress tracking using 4D schedule and 3D sensing technologies," *Autom. Constr.*, vol. 22, pp. 414–421, 2012.

[5] S. Halder, K. Afsari, J. Serdakowski, and S. DeVito, "A methodology for BIM-enabled automated reality capture in construction inspection with quadruped robots," in *Proc. Int. Symp. Autom. Robot. Constr.*, 2021.

[6] M. Arjmand, J. Jung, M. J. Olsen, H. A. Lassiter, and M. Jafari, "Conceptual design of advanced construction progress monitoring with terrestrial and robotic laser scanning systems," in *Proc. FIG Working Week*, 2023.

[7] F. Bosche, M. Ahmed, Y. Turkan, C. T. Haas, and R. Haas, "The value of integrating scan-to-BIM and scan-vs-BIM techniques for construction monitoring using laser scanning and BIM: The case of cylindrical MEP components," *Autom. Constr.*, vol. 49, pp. 201–213, 2015.

[8] Building Safety Regulator. (2024) Keeping information about a higher-risk building: The golden thread. Accessed 7 July 2026. [Online]. Available: www.gov.uk/guidance/keeping-information-about-a-higher-risk-building-the-golden-thread

[9] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, 2016.

[10] H. Huang, Z. Qiao, Z. Yu, C. Liu, S. Shen, F. Zhang, and H. Yin, "SLABIM: A SLAM-BIM coupled dataset in HKUST main building," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2025.

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[12] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, 2016.

[13] H. Wei, J. Jiao, X. Hu, J. Yu, X. Xie, J. Wu, Y. Zhu, Y. Liu, L. Wang, and M. Liu, "FusionPortableV2: A unified multi-sensor dataset for generalized SLAM across diverse platforms and scalable environments," *Int. J. Robot. Res.*, 2025.

[14] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stueckler, and D. Cremers, "The TUM VI benchmark for evaluating visual-inertial odometry," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 1680–1687.

[15] P.-E. Sarlin, M. Dusmanu, J. L. Schönberger, P. Speciale, L. Gruber, V. Larsson, O. Miksik, and M. Pollefeys, "LaMAR: Benchmarking localization and mapping for augmented reality," in *Proc. Eur. Conf. Comput. Vis.*, 2022.

[16] A. Krishnan, S. Liu, P.-E. Sarlin, O. Gentilhomme, D. Caruso, M. Monge, R. Newcombe, J. Engel, and M. Pollefeys, "Benchmarking egocentric visual-inertial SLAM at city scale," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2025.

[17] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 573–580.

[18] L. Zhang, M. Helmberger, L. F. T. Fu, D. Wisth, M. Camurri, D. Scaramuzza, and M. Fallon, "Hilti-Oxford dataset: A millimeter-accurate benchmark for simultaneous localization and mapping," *IEEE Robot. Autom. Lett.*, vol. 8, no. 1, pp. 408–415, 2023.

[19] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D data in indoor environments," in *Proc. Int. Conf. 3D Vis.*, 2017, pp. 667–676.

[20] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5828–5839.

[21] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3260–3269.

[22] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, "ScanNet++: A high-fidelity dataset of 3D indoor scenes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 12–22.

[23] Y. Tao, M. Á. Muñoz-Bañón, L. Zhang, J. Wang, L. F. T. Fu, and M. Fallon, "The Oxford Spires dataset: Benchmarking large-scale LiDAR-visual localisation, reconstruction and radiance field methods," *Int. J. Robot. Res.*, 2026.

[24] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in RGB-D images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2930–2937.

[25] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "InLoc: Indoor visual localization with dense matching and view synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7199–7209.

[26] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla, "Benchmarking 6DOF outdoor visual localization in changing conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8601–8610.

[27] M. Trzeciak, K. Pluta, Y. Fathy, L. Alcalde, S. Chee, A. Bromley, I. Brilakis, and P. Alliez, "ConSLAM: Construction data set for SLAM," *J. Comput. Civ. Eng.*, 2023.

[28] M. Helmberger, K. Morin, B. Berner, N. Kumar, G. Cioffi, and D. Scaramuzza, "The Hilti SLAM challenge dataset," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 7518–7525, 2022.

[29] A. D. Nair, J. Kindle, P. Levchev, and D. Scaramuzza, "Hilti SLAM challenge 2023: Benchmarking single + multi-session SLAM across sensor constellations in construction," *IEEE Robot. Autom. Lett.*, vol. 9, no. 8, 2024.

[30] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 4304–4311.

[31] B. Khomutenko, G. Garcia, and P. Martinet, "An enhanced unified camera model," *IEEE Robot. Autom. Lett.*, vol. 1, no. 1, pp. 137–144, 2016.

[32] J. Kannala and S. Brandt, "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1335–1340, 2006.

[33] X. Zhi, J. Hou, Y. Lu, L. Kneip, and S. Schwertfeger, "Multical: Spatiotemporal calibration for multiple IMUs, cameras and LiDARs," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 2446–2453.

[34] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Proc. SPIE Sensor Fusion IV*, vol. 1611, 1992, pp. 586–606.

[35] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallogr. A*, vol. 32, no. 5, 1976.

[36] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022.

[37] T. Kerssies, N. Cavagnero, A. Hermans, N. Norouzi, G. Averta, B. Leibe, G. Dubbelman, and D. de Geus, "Your ViT is secretly an image segmentation model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025.

[38] M. Grupp, "evo: Python package for the evaluation of odometry and SLAM," https://github.com/MichaelGrupp/evo, 2017.

[39] S. Boche, J. Jung, S. B. Laina, and S. Leutenegger, "OKVIS2-X: Open keyframe-based visual-inertial SLAM configurable with dense depth or LiDAR, and GNSS," *IEEE Trans. Robot.*, vol. 41, pp. 6064–6083, 2025.

[40] A. Umemura, T. Kuwahara, M. Pollefeys, and D. Barath, "Z-FLoc: Zero-shot floorplan localization via geometric primitives," *arXiv preprint arXiv:2606.04788*, 2026.

[41] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A research platform for visual-inertial estimation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020.

[42] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. Montiel, and J. D. Tardos, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, 2021.

[43] Y. Peng, C. Chen, K. Wu, and G. Huang, "sqrt-VINS: Robust and ultrafast square-root filter-based 3D motion tracking," *IEEE Trans. Robot.*, vol. 41, pp. 6570–6589, 2025.