Floorplan-Anchored Localization with a 360° Camera for the Hilti-Trimble Challenge 2026

Ayumi Umemura[1,3], Wei Zhang[2], Zihan Zhu[1], Toshinori Kuwahara[3], Marc Pollefeys[1], Norbert Haala[2], Daniel Barath[1]

***Abstract*— We present our approach to the Localization track of the Hilti-Trimble SLAM Challenge 2026. The challenge uses the Insta360 One-RS camera, which is equipped with two back-to-back fisheye lenses providing full 360° coverage together with an integrated IMU. As a front-end, we run OKVIS2-X [1] configured with both fisheye cameras under the equidistant distortion model and with online extrinsic calibration enabled; the resulting keyframe trajectory is exported as a pose graph in g2o format. We then incorporate building floorplan priors derived from Z-FLoc to anchor the trajectory in the floorplan coordinate frame. Our system is evaluated on all sequences of the dataset spanning multiple floors of a building under active construction.**

I. Introduction

Robust visual-inertial SLAM in construction environments is a long-standing open challenge. Construction sites combine poor and repetitive texture, rapidly changing layouts, mixed indoor/outdoor lighting, and large open spaces that starve feature-based systems of stable correspondences. The Hilti-Trimble SLAM Challenge 2026 [2] provides a realistic benchmark for these conditions: 30 sequences captured across multiple floors of an active construction site using an Insta360 One-RS 1-Inch Edition camera rigidly attached to a LiDAR-inertial reference sensor.

A distinctive property of the Insta360 sensor is its dual-fisheye design: two Leica Summicron-A 6.52 mm f/2.2 lenses, one pointing forward and one pointing backward, each covering approximately 200° of field of view for full 360° coverage. While standard visual-inertial odometry (VIO) pipelines typically assume a narrow-baseline perspective camera, exploiting both wide-angle views simultaneously substantially improves tracking robustness in texturally sparse construction interiors.

The challenge offers two tracks. The *SLAM* task requires a trajectory estimate in an arbitrary reference frame; the *Localization* task additionally requires the estimate to be expressed in the building floorplan coordinate frame. We participate in the Localization track. Section II describes our visual-inertial odometry front-end based on OKVIS2-X, and Section III describes our localization pipeline.

[1]Computer Vision and Geometry Group (CVG), ETH Zürich, Switzerland.
[2]Institute for Photogrammetry (IFP), University of Stuttgart, Germany.
[3]Space Infrastructure Engineering Lab, Tohoku University, Japan.

II. Visual-Inertial Odometry Front-End

A. Base System: OKVIS2-X

We use OKVIS2-X [1], an open-source visual-inertial SLAM system that supports multiple camera configurations and additional sensors. OKVIS2-X jointly optimizes camera poses and IMU states within a sliding-window bundle adjustment, and performs a final bundle adjustment over all keyframes at the end of each sequence to produce a globally refined trajectory.

B. Dual Fisheye Camera Configuration

We configure OKVIS2-X to use both the front and back cameras simultaneously. Each lens covers approximately 200° field of view, which exceeds what the pinhole model can represent. We therefore use the Kannala-Brandt equidistant model [3] for both cameras. Intrinsic parameters were taken from the challenge-provided Kalibr [4] calibration.

C. Online Extrinsic Calibration

We found the challenge-provided extrinsics (camera-to-IMU and front-to-back camera) to vary across sequences, likely due to mechanical play in the mount and thermal expansion as the device warms up. We therefore enable the online extrinsic calibration in OKVIS2-X, which refines the camera-to-IMU transformations jointly within the sliding-window optimization.

D. Pose Graph Export

After each sequence, the final bundle adjustment estimates of all keyframe camera poses are exported as a pose graph in g2o format [5]. Nodes represent keyframe poses; edges are relative-pose constraints derived from the covisibility graph. This pose graph is the primary input to the localization pipeline described in Section III.

III. Localization

Our localization pipeline is invoked offline once per sequence after the SLAM session completes.

A. Base System: Z-FLoc

To incorporate a floorplan-based global prior into the SLAM trajectory, we build upon Z-FLoc, a global localization system that takes as input the last N RGB frames together with their SE(3) poses from VI-SLAM, and outputs a Sim(2) transformation into the floorplan coordinate frame. Z-FLoc reconstructs a wall-only 3D point cloud and

projects it along gravity to obtain a bird's-eye-view (BEV) map. Line segments and circular primitives extracted from the BEV are matched to floorplan primitives via dedicated minimal solvers within a hybrid RANSAC loop. Top-ranked hypothesis is refined via nonlinear optimization to yield the final Sim(2) transform. The resulting transform is used as a global prior via pose graph optimization (PGO). Details of Z-FLoc will be described in a forthcoming paper.

B. Automatic Annotation on Floorplan

Given a rasterized floorplan image, wall pixels are isolated via morphological erosion. Line segments are then detected from the resulting wall edge map using the probabilistic Hough transform. Circular primitives (e.g., columns) are detected by clustering the remaining wall pixels with HDBSCAN [6]; the centroid of each cluster is taken as a keypoint. The extracted geometric primitives are saved per floor.

C. Gravity-Aligned Perspective Crops

We exploit 360° coverage from the Insta360 to extract perspective crops that maximize visible wall structure for the Z-FLoc matching stage. From each panorama, we crop six perspective views in the horizontal plane. Each crop is gravity-aligned using a coarse pose estimate derived from the VI-SLAM relative pose and the challenge-provided initial pose. A rotation aligning the image plane with gravity is then applied to the ray directions before sampling the equirectangular image via remapping.

D. Floorplan Prior

Z-FLoc produces top-10 pose hypotheses per keyframe. To suppress false localizations, we filter Z-FLoc's top-10 hypotheses per frame using a sliding-window consistency check: a candidate is accepted only if it is consistent with the SLAM-derived relative poses across the window, within translation and yaw thresholds. The accepted priors are then incorporated into a PGO over the VI-SLAM keyframe graph using GTSAM. Each accepted prior contributes a position-only GPS factor with a Tukey robust noise model, which suppresses residual outliers that pass the sliding-window filter.

IV. Results

We evaluated our system on all 24 sequences submitted to the challenge. Table I reports the per-sequence scores using the challenge exponential scoring metric:

$$\text{score} = \frac{1}{N}\sum_{t=1}^{N} 100 \cdot e^{-c\cdot \text{error}_t}, \quad c = 0.461. \tag{1}$$

The global floorplan prior provided by Z-FLoc substantially reduced trajectory drift: without floorplan correction the mean RMSE is 0.19 m, whereas anchoring the keyframe graph to the floorplan coordinate frame via PGO reduces it to 0.24 m.

TABLE I
PER-SEQUENCE LOCALIZATION RESULTS.

Dataset	RMSE (m)	Exp Score
floor_1_2025-07-07_run_1	0.2887	89.3602
floor_1_2025-12-02_run_1	0.1778	93.3493
floor_2_2025-12-02_run_1	0.3255	89.5551
floor_2_2025-12-03_run_1	0.2998	90.0148
floor_3_2025-05-19_run_1	0.1754	93.0813
floor_3_2025-12-02_run_1	0.3645	87.1897
floor_4_2025-05-19_run_1	0.2422	90.7738
floor_4_2025-12-02_run_1	0.1260	95.1339
floor_5_2025-12-02_run_1	0.1431	94.7929
floor_6_2025-06-18_run_1	0.2769	89.8547
floor_6_2025-07-07_run_1	0.3357	87.4158
floor_6_2025-12-02_run_1	0.1610	93.5871
floor_6_2025-12-02_run_2	0.1908	93.5472
floor_7_2025-12-02_run_1	0.1330	95.0458
floor_7_2025-12-02_run_2	0.2104	92.4993
floor_7_2025-12-03_run_1	0.2261	93.0427
floor_EG_2025-10-16_run_1	0.1507	94.3218
floor_EG_2025-12-02_run_1	0.1248	95.2424
floor_EG_2025-12-02_run_2	0.1837	93.6770
floor_UG1_2025-05-19_run_1	0.3217	88.6782
floor_UG1_2025-06-18_run_1	0.3356	87.1462
floor_UG1_2025-12-02_run_1	0.2425	91.6358
floor_UG1_2025-12-02_run_2	0.4262	86.5281
floor_UG1_2025-12-03_run_1	0.2377	90.8342
Mean Results	0.2375	–
Final Score	–	2196.3075

V. Conclusion

We presented our approach to the Localization track of the Hilti-Trimble SLAM Challenge 2026. We use OKVIS2-X with both fisheye cameras, the equidistant distortion model, and online extrinsic calibration as a visual-inertial odometry front-end, and export the final bundle adjustment result as a pose graph. Building floorplan priors from Z-FLoc are then incorporated via pose graph optimization to anchor the trajectory in the floorplan coordinate frame, substantially reducing drift in the challenging construction-site environment.

References

[1] 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 Transactions on Robotics*, 2025.

[2] S. Centanni, Y. Zhang, Y. Tao, J. Kindle, F. Neuhaus, T. Koß, A. Patel, M. Helmberger, E. Szymańska, T. Gräber, and M. Fallon, "Hilti-Trimble-Oxford dataset: 360 visual-inertial benchmark with floor plan priors for SLAM and localization," 2026.

[3] J. Kannala and S. S. Brandt, "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1335–1340, 2006.

[4] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1280–1286, IEEE, 2013.

[5] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g 2 o: A general framework for graph optimization," in *2011 IEEE International Conference on Robotics and Automation*, pp. 3607–3613, IEEE, 2011.

[6] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.