

TXR-SLAM System: Hilti SLAM Challenge 2023 Entry (Team “XR Penguin” – Vision/IMU-only track)

Yifu Wang, Yonhon Ng, Inkyu Sa, Álvaro Parra, Cristian Rodriguez-Opazo, Taojun Lin*, Hongdong Li

XR Vision Labs, Tencent Games; Australian National University*

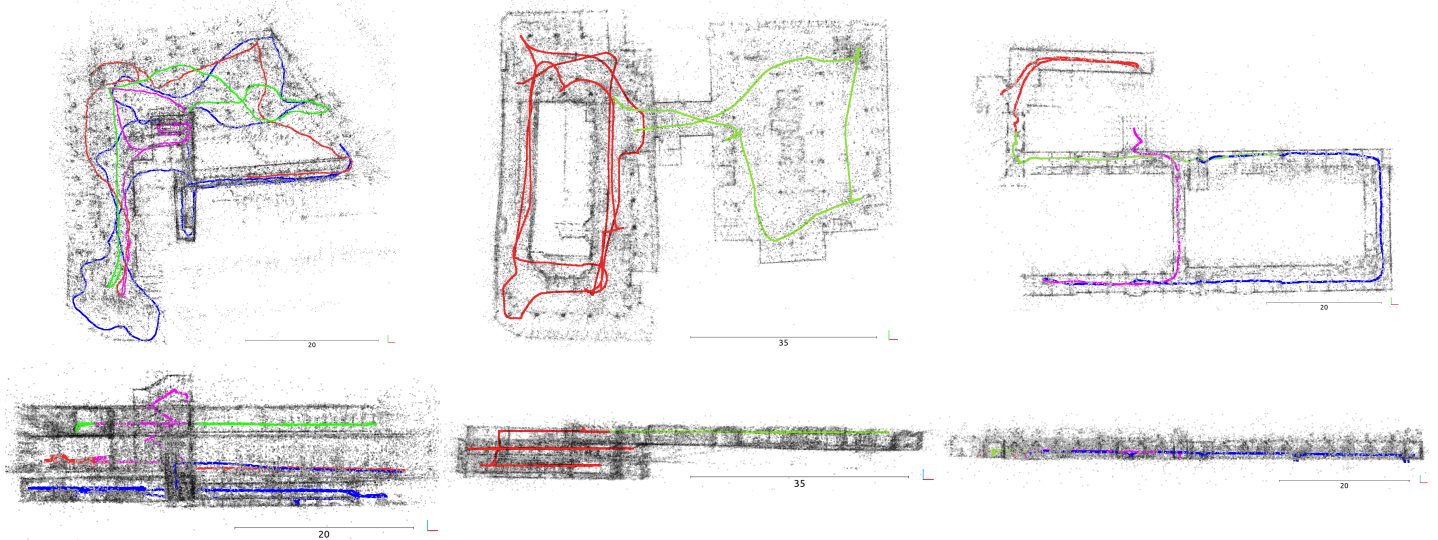


Figure 1: Merged maps and trajectories. Top row shows the top-down view, bottom row shows the side view. From left to right: site 1, site 2, site 3. Trajectory is coloured in sequence from red, green, blue, magenta. Note that automatic map merging fails for *site1_handheld_3* and *site2_robot_3* due to insufficient overlap.

Abstract

This report describes our entry for the Hilti SLAM Challenge 2023, the vision and IMU only track. Our SLAM system is named as TXR-SLAM, which is an optimisation-based Visual-Inertial SLAM system designed for multi-camera systems. Our system uses both sliding-window Bundle Adjustment to optimise the trajectory and support for multi-camera loop-closure, as well as full visual-inertial Bundle Adjustment for improved performance. Our system can be applied to XR (VR/AR) virtual and augmented reality, as well as robot navigation and autonomous driving.

1. Our responses to the competition’s requirements (Q&A)

This section presents our responses to the specific requirements as follows:

- **Q: Filter or optimization-based (or else)**
A: Optimisation based SLAM method.
- **Q: Is the method causal? (i.e. does not use information from the future to predict the pose at a given time).**
A: Non-Causal. We have employed a full-scale bundle adjustment as the final refinement, as well as loop closures.
- **Q: Is Bundle Adjustment (BA) used? What type of BA, e.g. full BA or sliding window BA?**
A: Yes, we use both window-BA and a full BA.
- **Q: Is loop closing used?**
A: Yes.
- **Q: Exact sensor modalities used (IMU, stereo or mono, LIDAR data?)**

A: Only cameras and IMU are used throughout the experiments. No Lidar data is ever used. We use a combination of stereo and monocular cameras set up.

- **Q: Total processing time for each sequence and the used hardware**

A: See the report.

- **Q: Whether the same set of parameters is used throughout all the sequences**

A: For all hand-held sequences we used the same set of parameters. For robot sequences, we use a different set of parameters, as the sensors are different.

- **Q: Whether manual alignment was performed for maps/trajectories in the multi-session submission.**

A: No. Only automatic map merging is used.

2. Single Session SLAM

Existing open-source VI-SLAM / VIO methods[1, 8, 9, 10] are typically designed to support only the monocular/dual-camera hardware configuration along with an IMU. However, some market products and self-developed engineering prototypes, such as the Alphasense Core Development Kit utilized in the competition, are equipped with multiple partially overlapped cameras, including forward-facing dual-camera, left/right/upward cameras, etc. Our TXR-SLAM system operates with the provided IMU and multiple cameras without incorporating any Lidar measurements. The main system components of our solution can be summarized as follows.

2.1. Sensor Calibration

The accuracy and reliability of the SLAM algorithm are directly impacted by the calibration of both the intrinsics and extrinsics parameters of the sensors. To ensure precise calibration, we use the widely-adopted Kalibr[5] tool for calibrating the handheld sensor suite and the Multical tool[11] for the robot sensor suite. We also calibrate for IMU intrinsics such as scale, misalignment and skew for handheld sequences and compensate them in our TXR-SLAM.

2.2. Data preprocessing

We apply histogram equalisation to compensate for dark frame (especially useful for *site1_handheld_1* and *site2_robot_3*). In the newly released *site3* dataset, there are numerous image frame and IMU drops. In particular, the second half of *site3_handheld_2* and *site3_handheld_3*, do not have all the five cameras image. We still make use of the remaining images in feature tracking to avoid integrating IMU (dead reckoning) for a long duration. For dropped

IMU, we use cubic spline to interpolate gyroscope and accelerometer data. This way, we obtained visible improvements to the SLAM performance.

2.3. Camera-IMU initialization

We follow the standard procedure to perform IMU initialization in a VIO system [3], which estimates the biases of accelerometer and gyroscope measurements, as well as the initial velocity and gravity direction. More advanced algorithms such as [7] can also be employed though we have not tried yet.

2.4. Front-end tracking

Both the handheld and robot-based sensor suites come with multiple cameras, with the forward-facing stereo cameras usually offering a more extensive co-visibility area compared to the left, right, and upwards cameras, which have only limited overlap with the forward-facing stereo pair. We employ a localization strategy based on a local map, which matches 2D extracted features and local 3D map points for pose estimation. This is done by first projecting all local map points onto the multi-camera image at the current time, and feature matching are done for both intra-cameras and inter-cameras to enhance the co-visibility relationships. These relationships are then used in the back-end optimization of TXR-SLAM to augment co-visibility edges and improve the positioning accuracy.

2.5. IMU Pre-integration

We derived a novel, exact IMU pre-integration formulation based on the $SE_2(3)$ exponential. This resulted in more accurate integration of IMU for fast rotational motion and for long integration time compared to existing methods [4, 2]. The corresponding covariance propagation is also derived for improved SLAM consistency. This improvement directly contributes the improved tracking performance of our SLAM system.

2.6. Back-end optimization / Loop-Closure

Similar to many other visual-inertial SLAM/VIO systems based on optimization scheme, our TXR-SLAM updates the body poses, velocity, IMU biases, and 3D landmarks' position by minimizing both visual reprojection errors based on observations from all cameras and error terms from pre-integrated IMU measurement using a sliding-window bundle adjustment scheme. Once a closed loop is detected, a global bundle adjustment will be executed. It optimizes the entire trajectory by utilizing all previous information, which greatly reduces drift across multiple sequences, especially for *site1_handheld_2*, *site1_handheld_4*, *site1_handheld_5*, *site1_handheld_2*, *site2_robot_2*, and *site3_handheld_1*.

Sequence name	Difficulties	Sequence length	Processing time	Single session score	Multi-session score
site1_handheld_1	dark around stairs going to Floor 1	204.71s	224.59s	32.5	4.17
site1_handheld_2	dark around stairs going to Floor 2	167.11s	211.98s	23.75	
site1_handheld_3	insufficient overlap for multi-session	170.63s	204.09s	22.5	
site1_handheld_4	-	295.42s	364.41s	30	
site1_handheld_5	-	159.29s	196.86s	26.67	
site2_robot_1	unsynchronised cameras, long, no loop closure	699.31s	531.89s	15.71	3.33
site2_robot_2	unsynchronised cameras	305.79s	194.40s	53.33	
site2_robot_3	dark, insufficient overlap for multi-session	359.00s	187.30s	19.0	
site3_handheld_1	-	97.18s	124.07s	105.0	19.55
site3_handheld_2	dropped data	148.13s	182.31s	35.0	
site3_handheld_3	dropped data	189.60s	243.46s	23.75	
site3_handheld_4	-	106.88s	130.34s	65.0	
Total score				452.21	27.05

Table 1: Difficulties, timing and score information for test sequences

3. Multi-Session SLAM

In order to participate in multi-session challenges, we utilize a similar methodology to merge multiple maps using our loop-closure correction module. The bag-of-words (BoW) model [6] is employed to retrieve potential overlapping keyframes, and local maps are used to aid in aligning the geometric maps. By fusing multiple sub-maps and conducting successful verification checks, we finally generate a globally-consistent map. In the context of the multi-session challenge, we select one of the single-session sequences as the foundation for the global map. Once SLAM processing is completed for a given sequence, the map is saved locally, and prior to processing subsequent data sequences, the saved map is pre-loaded in advance. By systematically processing all sequences and performing map fusion, we can integrate all the submaps into the global map. However, due to limited overlap between certain submaps and the global map, as well as challenging visual conditions such as inadequate lighting or texture-less areas, certain sequences like *site1_handheld_3* and *site2_handheld_3* cannot be merged into the global map.

4. Results

We use window-based BA and global BA together to obtain the fully optimised IMU poses, which means our result is non-causal. We also use loop closure whenever possible to reduce drift. We use IMU along with one stereo cameras and multiple monocular cameras setting in our TXR-

SLAM.

We used the same parameters for all handheld sequences in our experiments. However, in the case of robot sequences, we encountered inter-stereo-pair time-synchronization issues in *site2_robot_1*. Therefore, we only use a pair of stereo cameras with IMU in our SLAM system for this sequence. For the remaining two robot sequences, we selected the best synchronized four cameras in each sequence. Figure 2 depicts the estimated trajectories and sparse reconstruction results of all the sequences.

The timings and score obtained in the HILTI challenge for all sequences are presented in Table 1, which were evaluated on a desktop equipped with an AMD Ryzen 9 5950X 16-Core Processor and running Ubuntu 20.04.

5. Further Remarks

We would like to thank the organiser for the HILTI Challenge, and would like to add a few remarks for future references:

- The calibration dataset provided for the HILTI 2023 Challenge may have insufficient coverage of the calibration pattern across the whole image plane, which resulted in some issues in calibration based on our experiments. Thus, we choose to use the calibration dataset from the HILTI 2022 Challenge for the handheld sequence. In addition, the calibration dataset for the robot may have limited excitations for full 6DoF motion (mainly due difficulty of moving the heavy

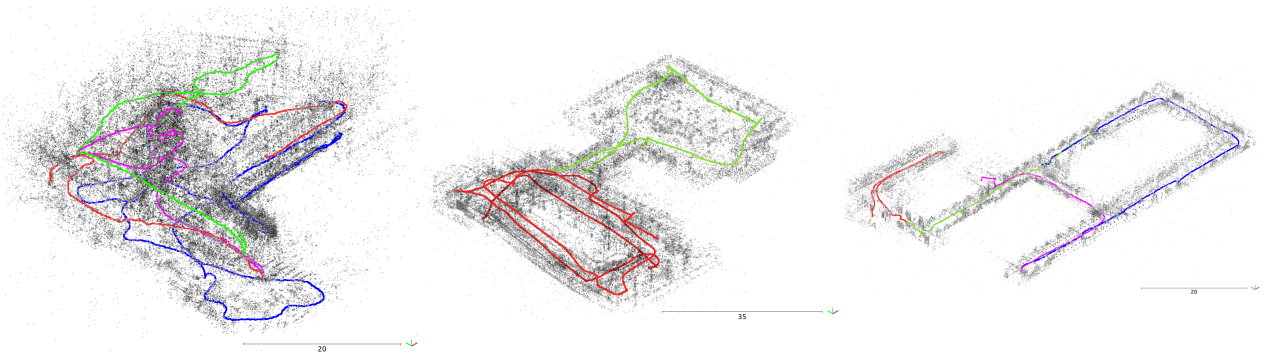


Figure 2: Perspective view of reconstructed scenes. From left to right: site 1, site 2, site 3. Trajectory is coloured in sequence from red, green, blue, magenta.

robot), which makes the extrinsics calibration more difficult than the handheld set up.

- Despite millimeter accuracy ground truth, we think there may be insufficient control points during the evaluation. This causes any slight deviation of the estimated trajectory to result in a significantly different Umeyama alignment. The evaluation is also performed with a heuristically chosen threshold. These causes a significant difference in the score with any slight variation in the estimated trajectory. In particular, our TXR-SLAM has an inherent variability between runs due to the use of RANSAC (for outlier rejection) and multi-threading (for shorter computational time). We observed a multi-run variation of absolute translation difference of less than 3cm RMSE even when using the same parameters. However, this may lead to a maximum difference of 30 points between runs.
- The front cameras for robot sequences (site 2) has visible circular Lidar lines that interfere with the vision-based feature tracking. Thus, we choose not to use the front cameras for our experiment.

Acknowledgement Taojun Lin is a current PhD student with ANU-Australian National University. This work was completed when he was a research intern with Tencent XR Vision Labs. The team had benefited from numerous discussions with Dr Yijia He *et al.* former members of the SLAM team of Tencent XR Labs.

References

- [1] 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 multi-map slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.
- [2] Jia-shan Cui, Fang-rui Zhang, Dong-zhu Feng, Cong Li, Fei Li, and Qi-chen Tian. An improved slam based on rk-vif: Vision and inertial information fusion via runge-kutta method. *Defence Technology*, 2021.
- [3] Tue-Cuong Dong-Si and Anastasios I Mourikis. Estimator initialization in vision-aided inertial navigation with unknown camera-imu calibration. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1064–1071. IEEE, 2012.
- [4] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. On-manifold preintegration theory for fast and accurate visual-inertial navigation. *IEEE Transactions on Robotics*, pages 1–18, 2015.
- [5] 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*, pages 1280–1286. IEEE, 2013.
- [6] Dorian Gálvez-López and Juan D Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012.
- [7] Yijia He, Bo Xu, Zhanpeng Ouyang, and Hongdong Li. A rotation-translation-decoupled solution for robust and efficient visual-inertial initialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 739–748, June 2023.
- [8] 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, 2018.
- [9] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1689–1696. IEEE, 2020.
- [10] Vladyslav Usenko, Nikolaus Demmel, David Schubert, Jörg Stückler, and Daniel Cremers. Visual-inertial mapping with non-linear factor recovery. *IEEE Robotics and Automation Letters*, 5(2):422–429, 2019.
- [11] Xiangyang Zhi, Jiawei Hou, Yiren Lu, Laurent Kneip, and Sören Schwertfeger. Multical: Spatiotemporal calibration for multiple imus, cameras and lidars. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2446–2453. IEEE, 2022.