Supplementary Materials —— EVI-SAM: Robust, Real-time, Tightly-coupled Event-Visual-Inertial State Estimation and 3D Dense Mapping

Abstract

In this report, we assess and compare the dense mapping performance of our EVI-SAM [1], which employs a monocular event camera, with the baseline approaches that utilize RGB-D, stereo cameras, and monocular cameras.

I. Global Mapping Performance of our EVI-SAM

In the EVI-SAM, we have presented the dense mapping performance of our EVI-SAM, illustrated in Fig. 1, showing the local and global dense map generated using monocular event camera. In section II, we used the data sequence "LG_office" from the EVI-SAM dataset[1]during the evaluation. For detailed information regarding this evaluation and the EVI-SAM dataset, we refer the reader to the paper [1].

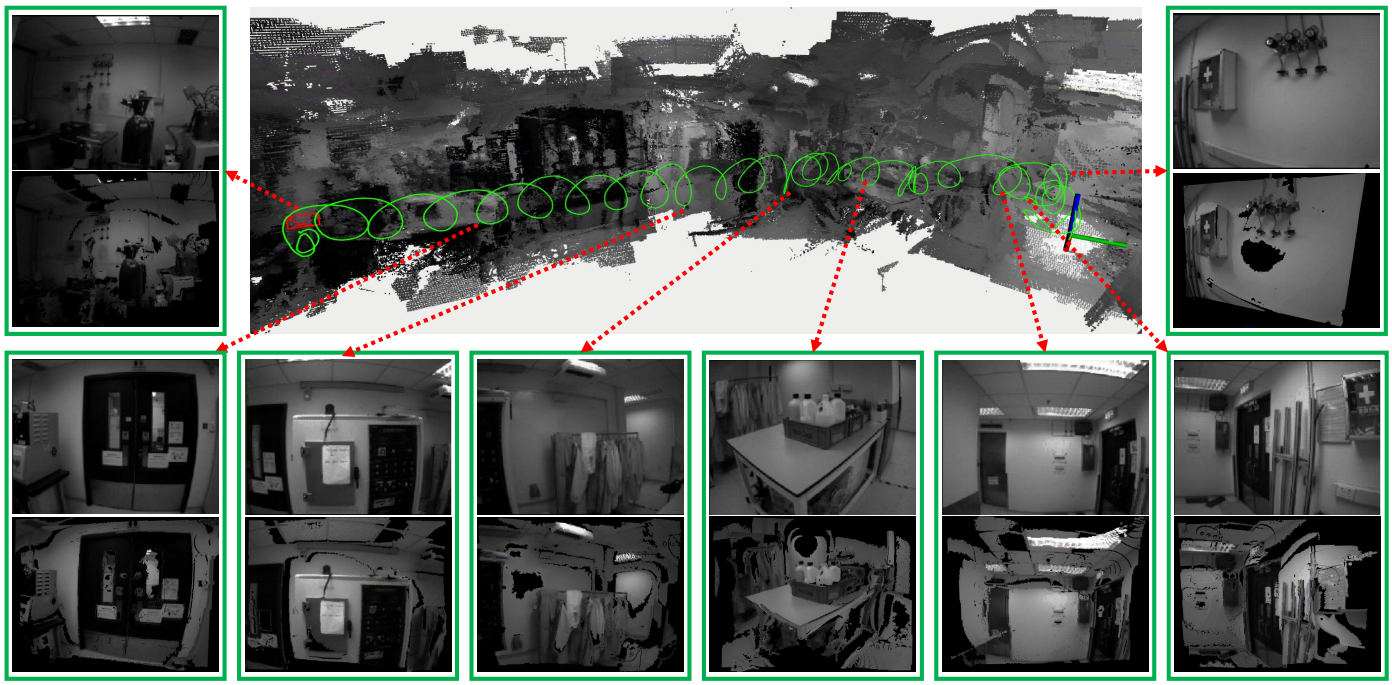

Fig. 1. Visualization of the estimated camera trajectory and global 3D reconstruction (surface mesh) of our EVI-SAM. Sequentially display from right to left includes the event-based dense point clouds with texture information and intensity images, at selected viewpoints.

[1]https://github.com/arclab-hku/Event_based_VO-VIO-SLAM/blob/main/EVI-SAM/data_srcipt.md

Fig. 2. Local mapping performance comparison. The first column shows the intensity image. Column 2 to 5 show the color point cloud results from our EVI-SAM, stereo RGB camera [2], RGB-D camera [3], and monocular camera [4], respectively.

II. Local Mapping Performance Comparison

In this section, we conducted qualitative comparisons of the depth estimation results presented as texture point clouds obtained from our EVI-SAM, with the depth estimation results from stereo RGB camera, monocular RGB camera, and RGB-D camera serving as the baseline. To ensure a fair comparison, firstly, we only employ the lightweight and real-time methods to estimate depth from standard camera, as opposed to more complex approaches that rely on GPU acceleration. Our EVI-SAM operates in real-time solely on CPU. Secondly, we resize the resolution depth generated from the stereo RGB camera, monocular RGB camera, and RGB-D camera to match the resolution of our event camera (346x260). Thirdly, we render the texture information from the intensity image of the DAVIS346 onto the estimated depth to generate the textured point cloud.

Our EVI-SAM is run in a full system model which includes the tracking and mapping pipeline. While textured depth results of the baseline methods are obtained as follows:

- The textured point cloud form RGB-D camera is directly obtained from the depth image of Realsense D455 [3], with its infrared ranging sensor obscured. While the texture information is transformed from the intensity image of the DAVIS346.
- The textured point cloud form stereo RGB cameras is generated by the image_undistort tool [2].
- The textured point cloud form monocular RGB camera is generated using the monocular dense mapping algorithm of Ref. [4], while it required the pose as input which is calculated from the pose estimation results from VINS-MONO [5].

As can be seen from Fig. 2, our EVI-SAM performs better than these baselines in terms of both the accuracy of depth recovery and the integrity of textures. The textured point cloud obtained from the RGB-D camera is directly captured from a commercial Realsense camera. However, its infrared ranging sensor is obstructed, as infrared light can interfere with the perception of event cameras. When the infrared ranging sensor is blocked, the depth map produced by the Realsense camera tends to contain a considerable amount of noise, as illustrated in column 4 of Fig. 2. This observation aligns with our expectations, as demonstrated in section VI.E.2 of EVI-SAM, where it is evident that the local depth map generated by the RGB-D camera exhibits more noise compared to that of our EVI-SAM. Our proposed event-based dense mapping approach demonstrates performance comparable to commercial depth cameras.

Regarding the results from monocular and stereo RGB cameras, as shown in columns 3 and 5 of Fig. 2, although their estimated depths exhibit less noise compared to those estimated by RGB-D cameras, the completeness of depth recovery falls far behind that of RGB-D and our EVI-SAM. This might be caused by the challenging testing conditions characterized by varying illumination and intense motion. Additionally, the regions with limited texture and the drawbacks such as narrow dynamic range of image sensor perception contribute to the poor depth recovery performance of monocular and stereo RGB cameras on this data sequence.

Certainly, image-based dense reconstruction has been extensively researched and has shown promising results. However, the exceptional advantages of event cameras highlight the importance of exploring event-based dense reconstruction, especially considering that non-learning methods for event-based dense reconstruction remain unexplored territory. Our method generates dense depths from event streams, with images only serving as guidance. As a result, the reconstruction results may not match the impressive performance of some state-of-the-art RGB-only or RGB-D methods, such as those utilizing NeRF [6] or 3D Gaussian Splatting [7]. Nevertheless, the dense mapping performance of our EVI-SAM is comparable to some approaches that use RGB, stereo RGB, and RGB-D cameras. Besides, our focus lies in addressing dense reconstruction using event cameras, which offer a potential solution for challenging scenarios not adequately addressed by standard images. This study not only aims to achieve optimal mapping performance in challenging situations but also serves as an inspiration for the research community encouraging future research in event-based dense reconstruction.

References

[1] W. Guan, P. Chen, H. Zhao, Y. Wang, and P. Lu, "Evi-sam: Robust, real-time, tightly-coupled event-visual-inertial state estimation and 3d dense mapping," *arXiv preprint arXiv:2312.11911*, 2023.

[2] Z. Taylor, "A compact package for undistorting images directly from kalibr calibration files," https://github.com/ethz-asl/image_undistort.

[3] "intelrealsense," https://www.intelrealsense.com/depth-camera-d455/.

[4] K. Wang, W. Ding, and S. Shen, "Quadtree-accelerated real-time monocular dense mapping," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–9.

[5] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[6] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[7] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.