

Supplementary Material for FAST-LIVO2: Fast, Direct LiDAR-Inertial-Visual Odometry

Chunran Zheng, Wei Xu, Zuhao Zou, Tong Hua, Chongjian Yuan, Dongjiao He, Bingyang Zhou, Zheng Liu, Jiarong Lin, Fangcheng Zhu, Yunfan Ren, Rong Wang, Fanle Meng, Fu Zhang

I. SYSTEM MODULE VALIDATION

In this section, we validate the key modules of our system, including the affine warping, normal refinement, reference patch update, on-demanding raycasting query, exposure time estimation, and ESIKF sequential update utilizing the FAST-LIVO2 private dataset and MARS-LVIG dataset.

A. Evaluation of Affine Warping

In this experiment, we aim to comprehensively evaluate the various affine warping effects based on the constant depth assumption (a common technique utilized in semi-dense methods), the plane prior from point clouds, and the refined plane normal of our proposed system (denoted as “Constant depth”, “Plane prior”, and “Plane normal refined”, respectively). To achieve this, we compare the mapping results and drift metrics of the three methods on “CBD Building 02” and “Office Building Wall”. As depicted in Fig. S2, “Plane normal refined” delivers the most clear and accurate mapping results and the next best is “Plane prior”. Notably, “Plane normal refined” renders text and patterns on the ground and walls, as well as lane markings, with remarkable clarity. Moreover, the drift associated with “Plane prior” and “Plane normal refined” remains below 0.01 m, whereas “Constant depth” does not return to the starting position, experiencing a drift of 0.22 m. Such results confirm the enhanced performance of affine warping based on the plane prior and enhancements by the plane normal refinement.

Besides, we compare warped projection effects based on “Constant depth” and “Plane prior” on the sequence “CBD Building 02” and “Office Building Wall”. We randomly select several image frames from these two sequences for the qualitative analysis. For each frame, we project the reference patches attached to the visual map points visible in the frame onto a blank image of the current frame. This process yields a novel RGB image. If the affine warping and pose estimation are both performed well, areas with patch projections will produce a seamless and minimally distorted appearance, closely resembling the raw RGB image. The comparison of warped patches is presented in Fig. S1. The results indicate that the pose accuracy and warped performance under “Plane prior” significantly outperform those under “Constant depth”.

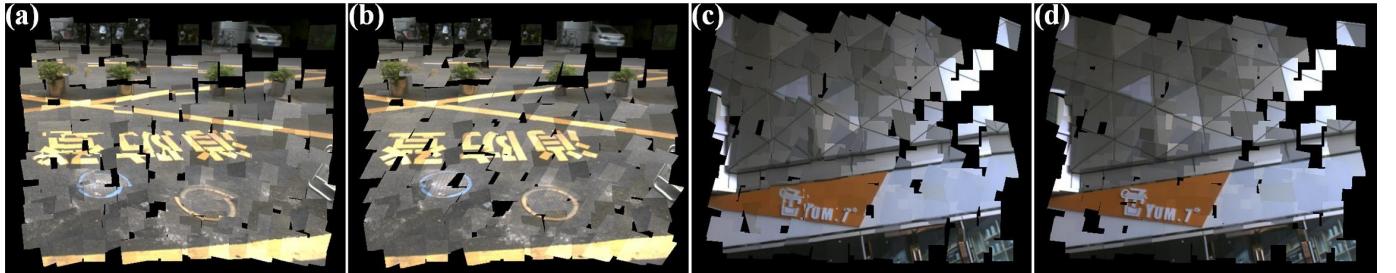


Fig. S1: (a) and (c) depict the warped patches derived from the constant depth assumption, while (b) and (d) represent those based on the plane prior.

B. Evaluation of Reference Patch Update and Normal Convergence

In this experiment, we validate the effects of the reference patch update strategy and normal convergence on “HIT Graffiti Wall” and “HKU Centennial Garden”. As shown in Fig. S3, (a) and (b) are the reconstructed point clouds of these two sequences. On the right, for each region from A to H, we respectively present five patch observations captured at different poses, each with a patch size of 40×40 pixels for visualization. These patches are observed in camera frames located at corresponding numbers on the left. As can be seen, our reference path update strategy tends to choose a high-resolution reference patch that faces the plane along its normal. It is also noticed that, despite of visual map points and patches generated at non-planar locations (e.g., tree leaves, trunk, and a lamp stand), the overall mapping quality is still high.

We also evaluate the convergence of our proposed normal estimation across patches in regions A to H. Each patch is in the size of 11×11 pixels. The initial normal vector is estimated from the LiDAR points. The convergence curves, which represent the angle change between the initial and optimized normal vectors at each iteration number, are shown in Fig. S3. Regions A, C, D, and E are structured areas, whereas regions B, F, G, and H are unstructured ones. It can be observed that normal

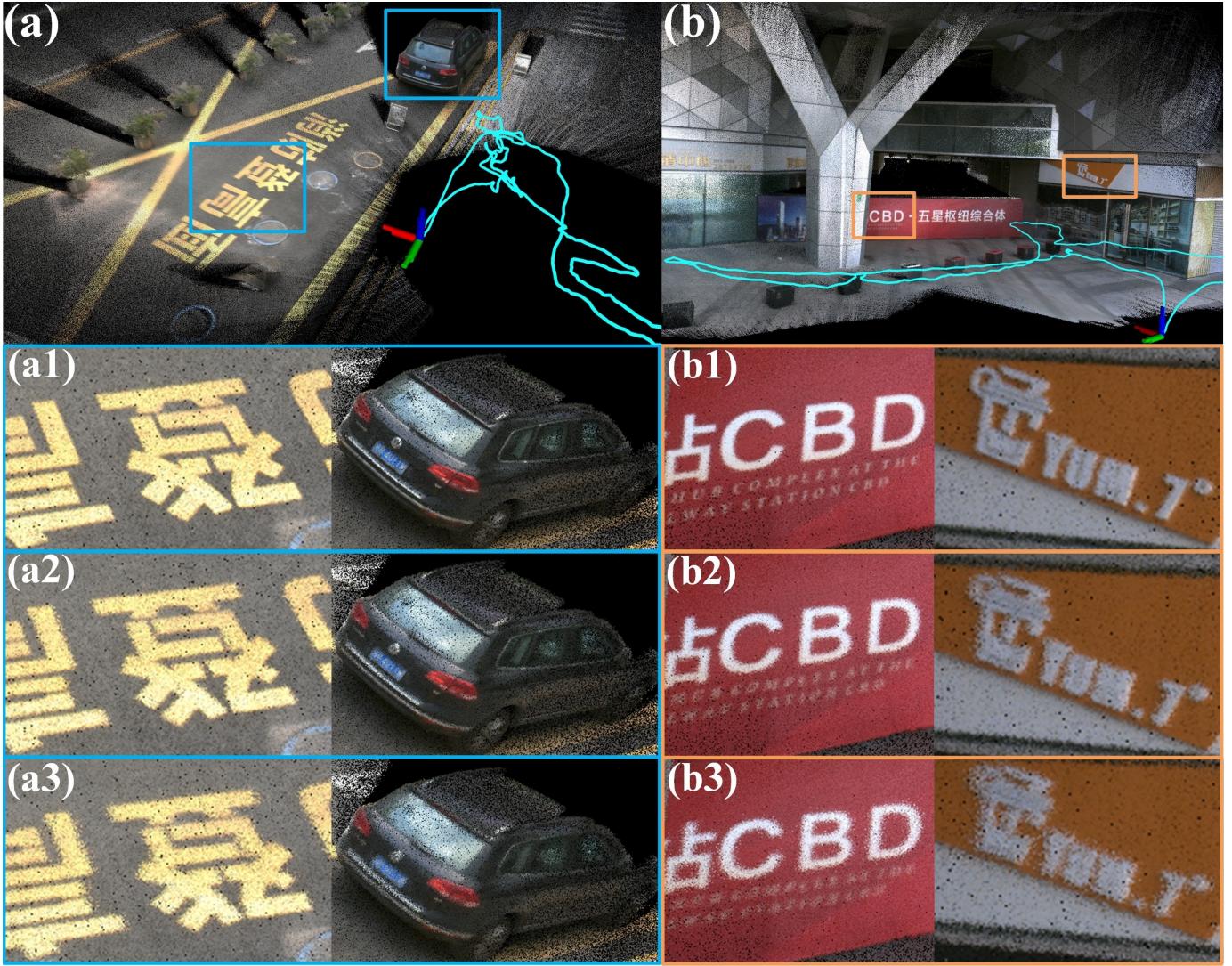


Fig. S2: (a) and (b) are FAST-LIVO2 default mapping results in “CBD Building 02” and “Office Building Wall”, respectively. (a1, b1), (a2, b2), (a3, b3) are enlarged views of the point clouds using “Plane normal refined”, “Plane prior”, and “Constant depth”, respectively.

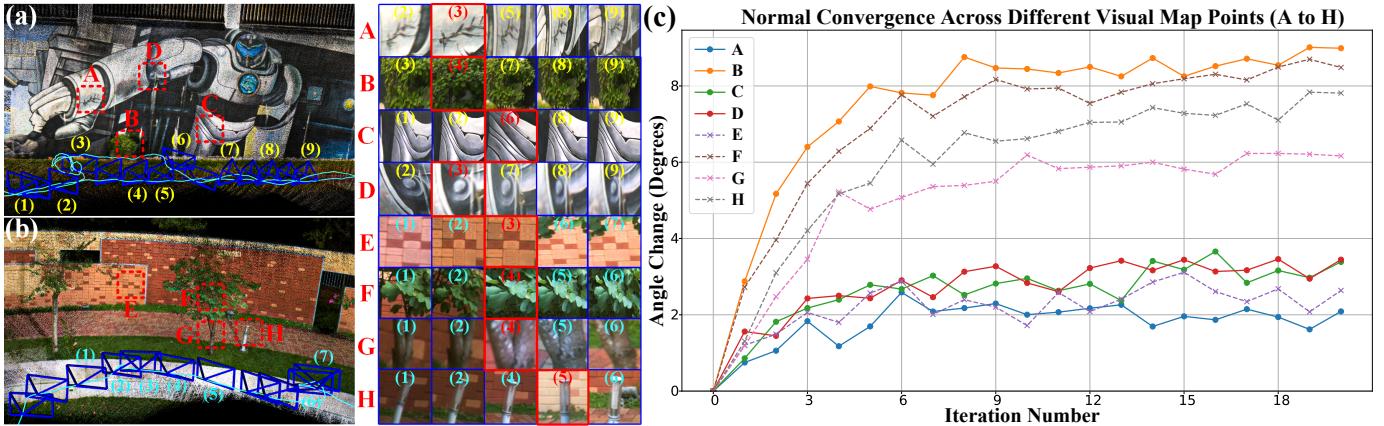


Fig. S3: The illustration of reference patch update. (a) and (b) are the reconstructed point clouds of sequences “HIT Graffiti Wall” and “HKU Centennial Garden”, respectively. Regions A to H encompass both structured and unstructured areas in the scene. On the right, the 40×40 image patches illustrate various observations for the same region, with numbers indicating the corresponding camera frame on the left. The reference patch for each region is highlighted by a red box. (c) shows the convergence of patch normal across regions A to H.

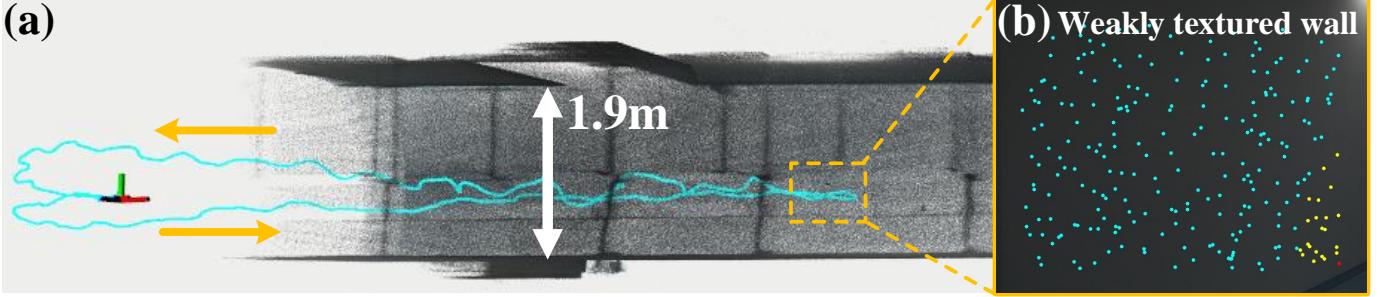


Fig. S4: The illustration of on-demand raycasting and voxel query. (a) displays a top-down view of the reconstructed point cloud for “Narrow Corridor”, with a corridor width of 1.9 m. In (a), the blue line indicates the trajectory and the yellow arrows represent the direction of movement. The LiDAR camera sensor suite is turned to face a wall on the left side at the location contained in the dashed box. (b) shows the camera image during the turn, having dim lighting conditions and very few LiDAR points. In (b), blue dots represent points acquired from raycasting, yellow dots indicate points from voxel query, and red dots are points in the LiDAR scan.

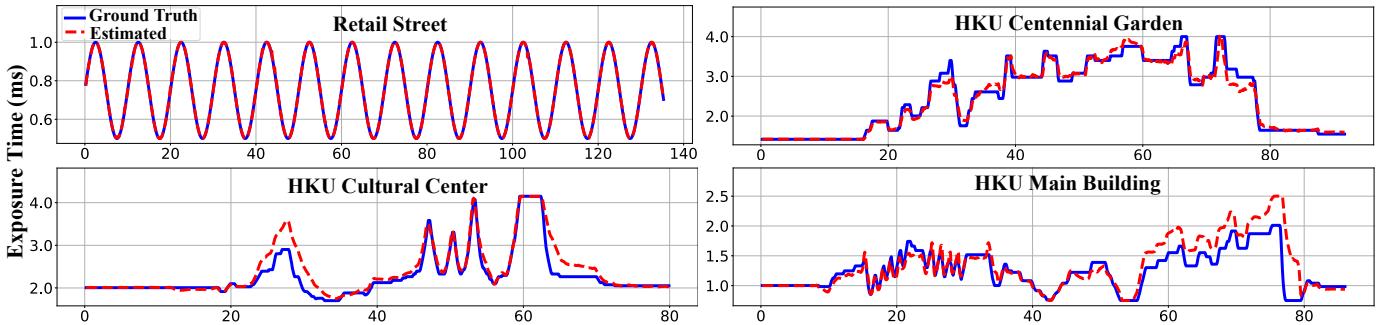


Fig. S5: Comparison of estimated exposure time versus ground truth exposure time in “Retail Street”, “HKU Centennial Garden”, “HKU Cultural Center”, and “HKU Main Building”.

vectors for structured areas converge faster (within 6 iterations) with small normal refinement (2 to 4 degrees) because the initial normal provided by point clouds is relatively accurate. In unstructured areas, such as shrubs and tree leaves (i.e., B and F), the normal refinement is significant (up to 9 degrees) and requires 9 iterations to converge. Overall, normal refinement across these 8 regions demonstrates good convergence properties.

C. Evaluation of On-demand Raycasting

In this experiment, we assess the performance of the on-demand raycasting module under extreme conditions where the current and recent LiDAR scans have few or even no points due to LiDARs’ close proximity blind zone [1]. We use the sequence “Narrow Corridor” for an in-depth analysis illustrated in Fig. S4. In this sequence, we traverse an extremely narrow tunnel, approximately 1.9 m in width, and turn to face the weakly textured wall on one side. Due to the limited points in the LiDAR scans when facing the wall, we can only acquire few visual map points through voxel query (yellow dots in Fig. S4 (b)). In this case, raycasting offers sufficient visual constraints to mitigate degeneration (blue dots in Fig. S4 (b)). The visualization results demonstrate that the on-demand raycasting module works well under challenging conditions with few points in LiDAR scans.

D. Evaluation of Exposure Time Estimation

In this experiment, we validate the exposure time estimation module in two parts: 1) For the sequence with fixed exposure and gain, we multiply each pixel of the received raw image by an exposure factor that changes sinusoidally over time. We verify the effectiveness of our estimation by comparing the estimated exposure times against the sinusoidal function applied. 2) For sequences with auto-exposure and either fixed or auto-gain settings, we evaluate the accuracy of our estimated exposure times by comparing them with the ground truth values retrieved from the camera’s API.

In part one, we test on the sequence “Retail Street”, applying an exposure factor to images with fixed exposure and gain. As shown in Fig. S5, the estimated relative inverse exposure time matches the true values very well, evidencing the convergence of our exposure estimation in synthetic conditions. In part two, we use the sequences “HKU Centennial Garden”, “HKU Cultural Garden” and “HKU Main Building”, which have significant exposure time changes, for testing. We scale the estimated relative exposure times by the first frame to recover the actual exposure time (ms) of each frame. As shown in Fig. S5, the estimated exposure time follows closely the ground-truth values, which validates the effectiveness of our exposure time estimation module. The occasional mismatches are possibly due to the unmodeled response function and vignetting factor [2].

E. Evaluation of ESIKF Sequential Update

In this experiment, we evaluate different ESIKF update strategies for LiDAR and camera states. We compare asynchronous versus synchronous updates, as well as standard versus sequential updates. Specifically, we assess three strategies: “asynchronous (standard update)”, where the camera and LiDAR states are updated at their respective sampling times without scan recombination; “synchronous (standard update)”, where LiDAR scans are recombined to sync with camera images and the state is updated with both LiDAR and camera measurements within a standard ESIKF; and “synchronous (sequential update)”, where the LiDAR and camera are synced but the state is first updated by LiDAR measurements and then updated by camera measurements. These strategies are evaluated in terms of accuracy, robustness, and efficiency using the “AMvalley03” sequence of the MARS-LVIG dataset. We select this sequence for several key reasons:

- (1) This sequence includes slopes that lead to both LiDAR and visual degenerations, making it a challenging test case.
- (2) This sequence represents an extremely large-scale scenario (approximately 901 m × 500 m × 130 m) with long-term and high-speed data collection (covering 600 s at a speed of 12 m/s), where pose deviations are prone to occur (due to the long-term and high-speed conditions), and even slight drifts can cause significant blurring in the colored point clouds (due to the large scale), leading to more pronounced comparative results.
- (3) This sequence provides the RTK ground truth data, allowing for more accurate quantitative comparisons.

We compare the qualitative mapping results, quantitative APE, and the average processing time of the three update strategies. The experimental configuration is as follows: LiDAR updates involve up to 5 iterations, visual updates use a three-level pyramid with up to 5 iterations per level and no more than 3 iterations per level when the camera and LiDAR are updated simultaneously in a standard ESIKF, and the scale normalization factor (from visual photometric error to LiDAR point-to-plane distance) for the “synchronous (standard update)” is set to 0.0032, which has been meticulously tuned for optimal performance.

Fig. S6, (a-c) present the reconstructed colored point clouds for this sequence. It is evident that the “synchronous (sequential update)” strategy produces accurate mapping results, particularly in the areas highlighted by the blue and orange boxes, where the mountain roads are reconstructed without any layering. In contrast, the other two strategies exhibit misalignments in these areas, although the “synchronous (standard update)” performs slightly better than the “asynchronous (standard update)”. The superior performance of the “synchronous (sequential update)” strategy is mainly attributed to its robustness in handling significant LiDAR and visual degenerations, as seen in the white box (c3). This area features a large, textureless slope, and the UAV passes over it at a high speed, heavily relying on a strong prior. The other two methods, which rely solely on the IMU prior, struggle to compute a relatively accurate image gradient descent direction, leading to significant linearization errors.

The APE (RMSE) metrics for the “AMvalley03” sequence are 3.12 m, 2.45 m, and 0.68 m for the “asynchronous (standard update)”, “synchronous (standard update)”, and “synchronous (sequential update)”, respectively. The average processing times on a desktop PC (Section IX-A) are approximately 27.6 ms, 49.9 ms, and 23.1 ms. Our proposed “synchronous (sequential update)” achieves the highest efficiency and accuracy, while the “asynchronous (standard update)” has the lowest accuracy. The “synchronous (standard update)” is the most time-consuming, primarily because it requires fusing all LiDAR measurements at each level of the image pyramid.

Overall, our proposed “synchronous (sequential update)” offers superior accuracy and efficiency, while the “asynchronous (standard update)” has the lowest accuracy, and the “synchronous (standard update)” is the most time-consuming.

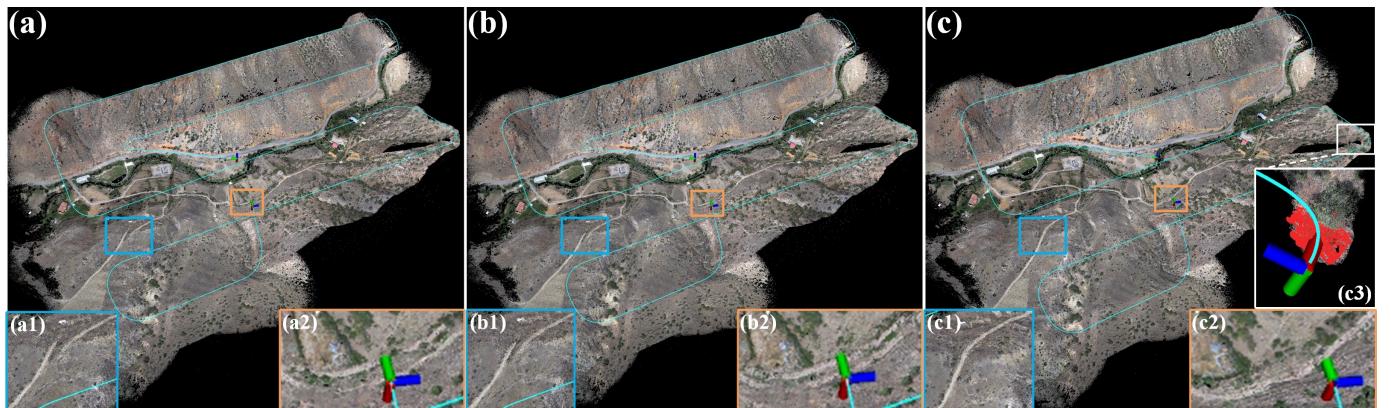


Fig. S6: (a), (b), and (c) are the mapping results in “AMvalley03” for the “asynchronous (standard update)”, “synchronous (standard update)”, and “synchronous (sequential update)” strategies, respectively. (a1, b1, c1) and (a2, b2, c2) are enlarged views of the point clouds for different update strategies. The blue line represents the UAV’s flight path, and the red points in (c3) indicate the LiDAR scan at that moment.

II. ADDITIONAL INFORMATION

TABLE S1: Overview of FAST-LIVO2 Private Dataset

Sequence	Duration (minute:second)	Sensor Degeneration	Return Origin ¹	Exposure / Gain Mode	Exposure Time ²	Close Surface Distance ³	Scene Characteristics
Retail Street	2 min : 15 sec	—	✓	Fixed / Fixed			Outdoor
CBD Building 01	1 min : 58 sec	—	✓	Fixed / Auto			Outdoor
CBD Building 02	3 min : 54 sec	LiDAR	✓	Auto / Fixed	✓		Outdoor
CBD Building 03	5 min : 01 sec	Camera, LiDAR	✓	Auto / Fixed	✓		Outdoor, Textureless
HKU Landmark	1 min : 31 sec	—	✓	Auto / Fixed	✓		Outdoor
HKU Lecture Center	1 min : 17 sec	LiDAR	✓	Auto / Fixed	✓		Indoor
HKU Centennial Garden	1 min : 32 sec	LiDAR	✓	Auto / Fixed	✓		Outdoor
HKU Cultural Center	4 min : 01 sec	LiDAR	✓	Auto / Fixed	✓		Outdoor
HKU Main Building	1 min : 36 sec	Camera, LiDAR	✓	Auto / Auto	✓		Out(Indoor) ⁴ , Textureless
HKUST Red Sculpture	2 min : 10 sec	—	✓	Fixed / Auto			Outdoor, Cluttered
HIT Graffiti Wall	8 min : 57 sec	LiDAR	✓	Fixed / Auto			Outdoor
Banner Wall	1 min : 40 sec	LiDAR	✓	Fixed / Auto			Indoor, Dim
Bright Screen Wall	1 min : 18 sec	LiDAR	✓	Fixed / Auto			Indoor
Black Screen Wall	0 min : 56 sec	Camera, LiDAR	✓	Fixed / Auto			Indoor, Textureless
Office Building Wall	1 min : 9 sec	LiDAR	✓	Fixed / Auto		✓	Outdoor
Narrow Corridor	1 min : 1 sec	Camera, LiDAR	✓	Fixed / Auto		✓	Indoor, Dim, Textureless
Long Corridor	1 min : 35 sec	LiDAR	✓	Fixed / Auto		✓	Indoor, Dim
Mining Tunnel	15 min : 49 sec	Camera, LiDAR	✓	Fixed / Auto			Indoor, Dim, Textureless
SYSU 01	4 min : 40 sec	—	✓	Auto / Auto			Outdoor, Light ⁵
SYSU 02	4 min : 30 sec	Camera, LiDAR	✓	Auto / Auto			Out(Indoor), Light
Total	66 min : 50 sec	Camera, LiDAR	✓	Auto (Fixed) / Auto (Fixed)	✓	✓	Out(Indoor), Light, Dim, Textureless

¹ Sequences are collected by traveling in a loop, starting and ending at the same position.

² Sequences with ground truth camera exposure time read from camera's API.

³ Sequences exist where LiDAR captures no/limited point clouds due to close proximity blind zones.

⁴ Sequences include the process of moving from indoor to outdoor environments.

⁵ Sequences characterized by significant lighting variations.

TABLE S2: Overview of UAV Autonomous Navigation Experiments

Experiment	Flight Duration (minute:second)	Flight Mode ⁶	Sensor Degeneration	Return origin	Exposure / Gain Mode	Exposure Time	Close Proximity to Obstacles	Scene Characteristics
Basement	5 min	Autonomous	Camera, LiDAR	✓	Auto / Auto	✓		Out(Indoor), Light
Narrow Opening	2 min:32 sec	Manual	Camera, LiDAR	✓	Auto / Auto	✓	✓	Out(Indoor), Light
SYSU Campus	3 min:6 sec	Manual	—	✓	Auto / Auto			Outdoor, Light
Woods	2 min:49 sec	Autonomous	Camera	✓	Auto / Fixed	✓		Outdoor, Cluttered

⁶ In autonomous mode, FAST-LIVO2, planning, and MPC modules are all running. In manual mode, FAST-LIVO2 and MPC are running while planning is disabled.

TABLE S3: Average processing time per frame for FAST-LIO2 across different datasets

Hilti'22 & Hilti'23	Processing Time (ms)	NTU VIRAL	Processing Time (ms)	Private Dataset	Processing Time (ms)
Construction Ground	24.32	eee_01	17.63	Retail Street	10.23
Construction Multilevel	23.21	eee_02	17.33	CBD Building 01	10.40
Construction Stairs	×	eee_03	15.32	CBD Building 02	×
Long Corridor	28.43	nya_01	19.99	CBD Building 03	×
Cupola	×	nya_02	18.62	HKU Landmark	11.62
Lower Gallery	24.49	nya_03	19.32	HKU Lecture Center	×
Attic to Upper Gallery	×	nya_04	15.71	HKU Centennial Garden	×
Outside Building	18.67	sbs_02	16.10	HKU Cultural Center	×
Floor 0	27.34	sbs_03	17.71	HKU Main Building	×
Floor 1	27.12			HKUST Red Sculpture	10.98
Floor 2	25.22			HIT Graffiti Wall	×
Basement	22.31			Banner Wall	×
Stairs	20.91			Bright Screen Wall	×
Parking 3x floors down	×			Black Screen Wall	×
Large room	29.26			Office Building Wall	×
Large room (dark)	28.20			Narrow Corridor	×
				Long Corridor	×
				Mining Tunnel	×
				SYSU 01	11.21
				SYSU 02	×
Overall Average	19.68				

⁷ × denotes the system totally failed.



Fig. S7: Challenging environments captured in the FAST-LIVO2 private dataset.

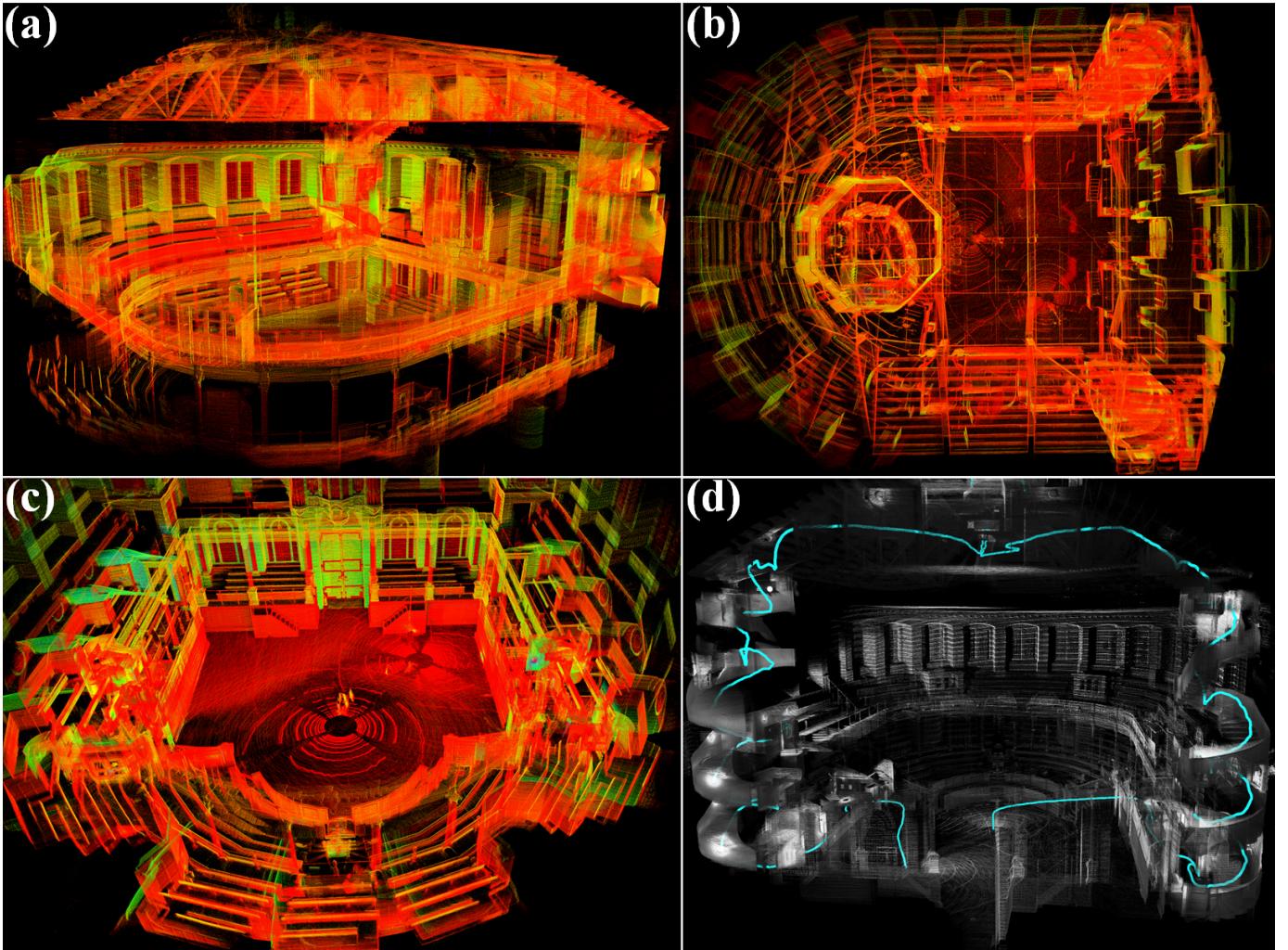


Fig. S8: The real-time mapping results of FAST-LIVO2 in the Hilti'22 dataset. (a) “Attic to Upper Gallery”, (b) “Cupola”, (c) “Lower Gallery”, and (d) “Construction Stairs”. The point clouds in (a-c) are colored by intensity, while (d) is colored using grayscale images.



Fig. S9: The real-time mapping results generated online in rich texture and structured scenes. The point clouds from left to right correspond to “HKU Landmark”, “SYSU 01” and “CBD Building 01”, respectively, showing the comparison of colored point cloud accuracy among FAST-LIVO2, FAST-LIVO, R3LIVE, and FAST-LIO2.

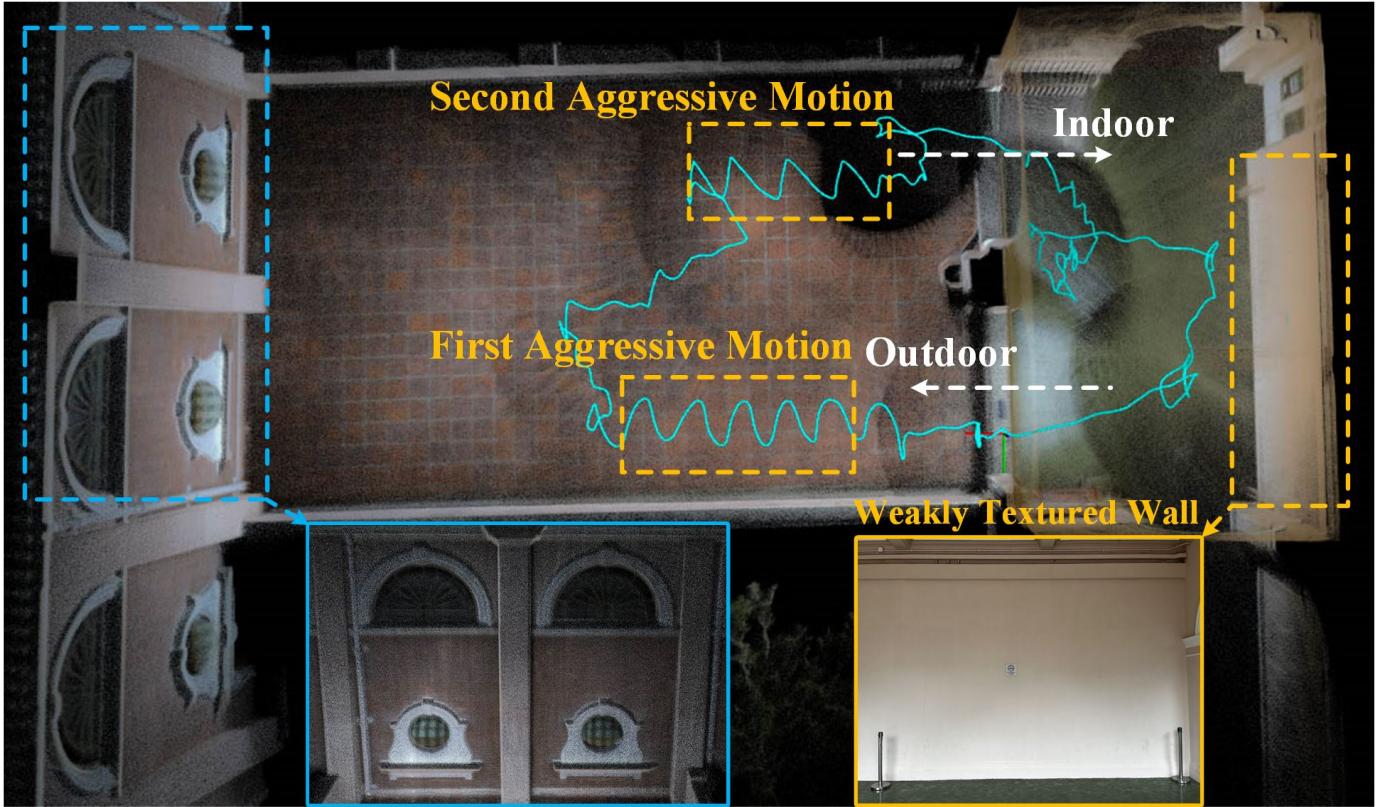


Fig. S10: The real-time mapping results of FAST-LIVO2 in “HKU Main Building”, containing aggressive motions, indoor-to-outdoor, outdoor-to-indoor, and a weakly textured wall.

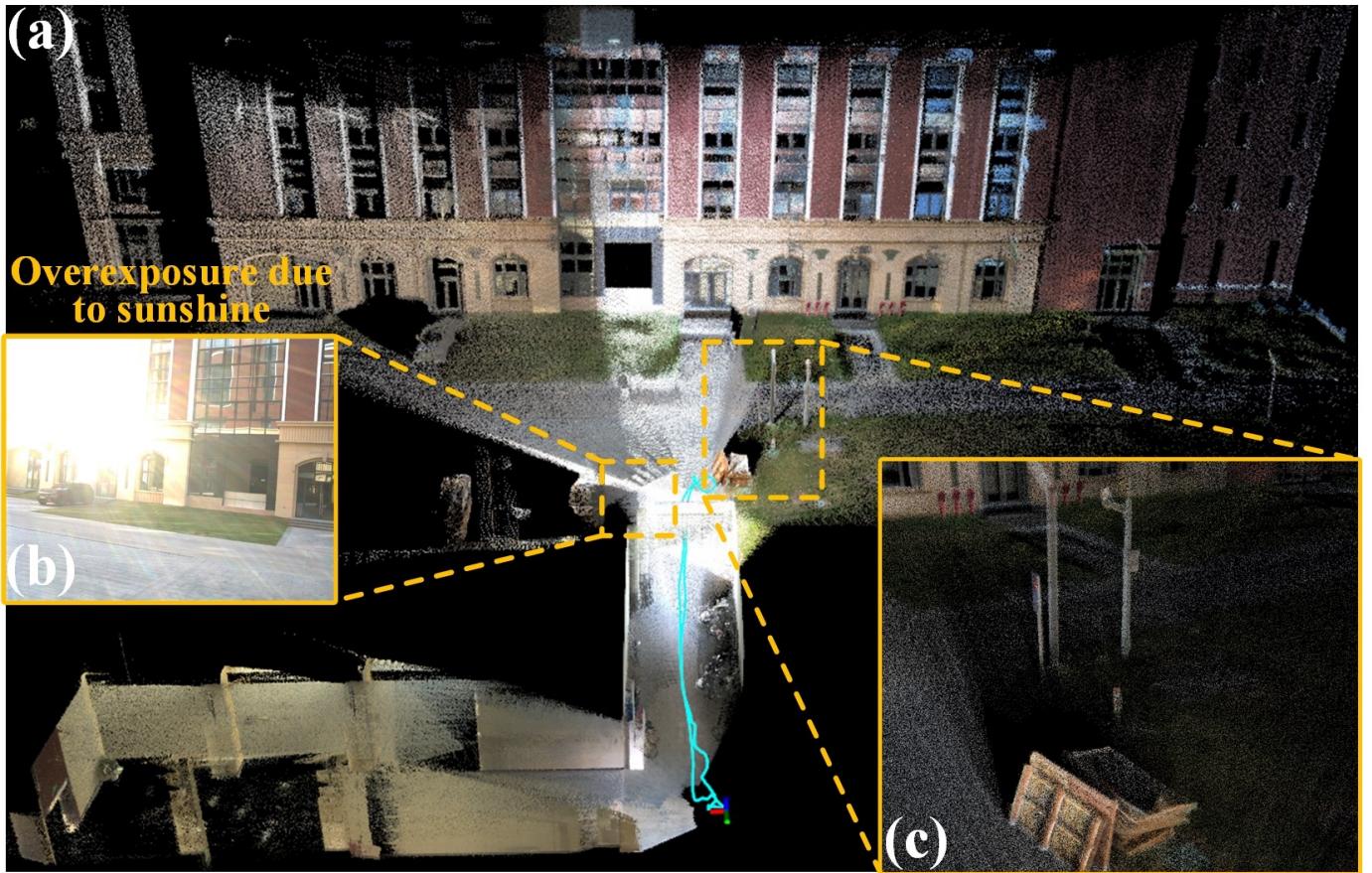


Fig. S11: The real-time mapping results of FAST-LIVO2 in “SYSU 02”. (a) birdview of the colored point map, (b) drastic lighting changes from indoor to outdoor facing the sun, (c) closeup view of details.

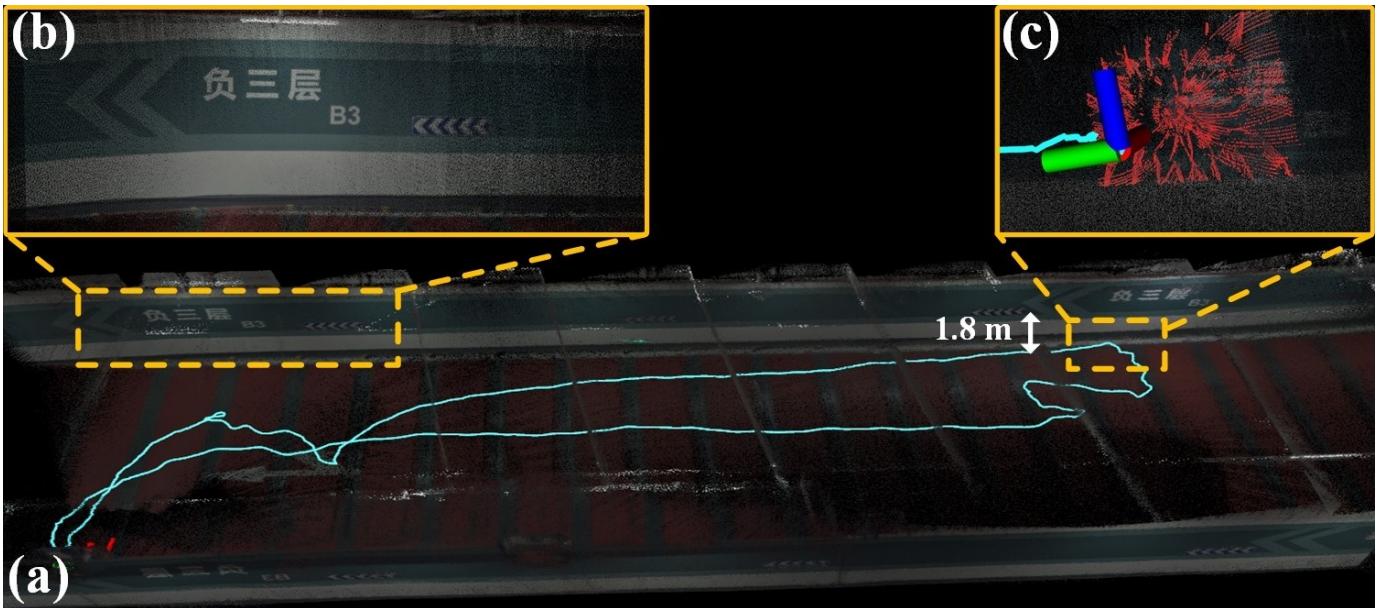


Fig. S12: The real-time mapping results of FAST-LIVO2 in “Long Corridor”. (a) birdview of the colored point map, (b) closeup view of details, (c) LiDAR facing a single wall causing LiDAR degeneration.



Fig. S13: The real-time mapping results of FAST-LIVO2 in rich texture and structured scenes. (a) “HKUST Red Sculpture”, (b) “CBD Building 01”.

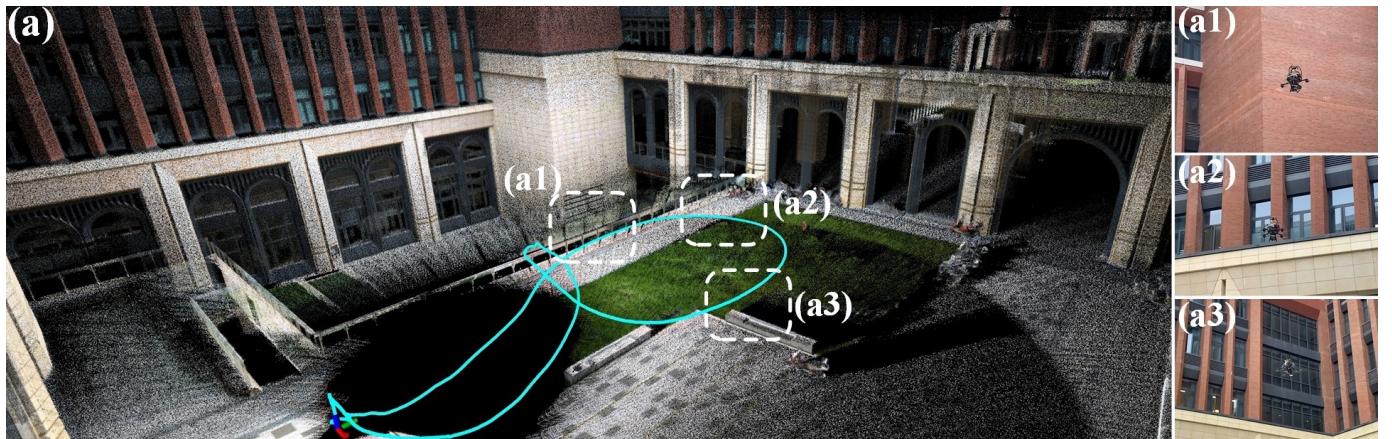


Fig. S14: (a) is the enlarged point cloud image of the “SYSU Campus” experiment. (a1), (a2), and (a3) represent the third-person view at the corresponding locations.

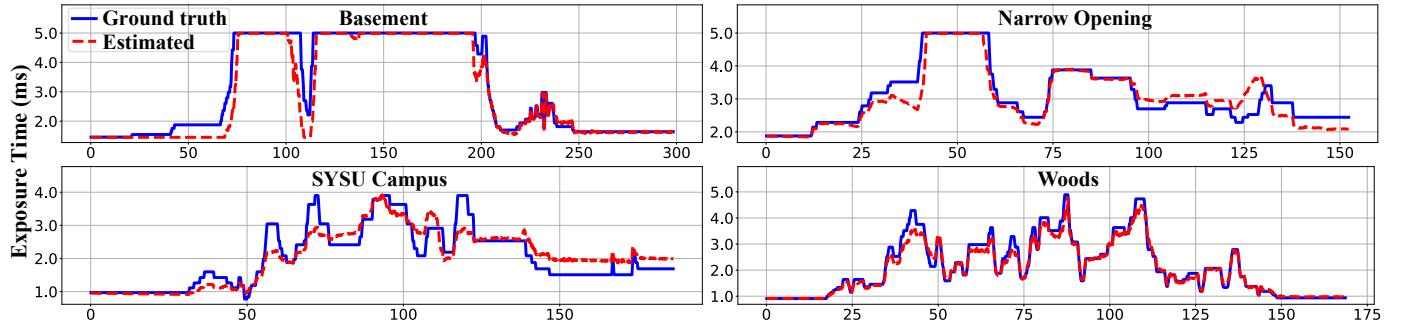


Fig. S15: Comparison of estimated exposure time versus the ground truth exposure time in “Basement”, “Narrow Opening”, “SYSU Campus”, and “Woods”.

REFERENCES

- [1] “Specs: Livox avia lidar sensor,” available online: <https://www.livoxtech.com/avia/specs>.
- [2] J. Engel, V. Usenko, and D. Cremers, “A photometrically calibrated benchmark for monocular visual odometry,” *arXiv preprint arXiv:1607.02555*, 2016.