

FAST-LIVO: Fast and Tightly-coupled Sparse-Direct LiDAR-Inertial-Visual Odometry

Chunran Zheng^{1*}, Qingyan Zhu^{1*}, Wei Xu¹, Xiyuan Liu¹, Qizhi Guo¹ and Fu Zhang¹,

Abstract—To achieve accurate and robust pose estimation in Simultaneous Localization and Mapping (SLAM) task, multi-sensor fusion is proven to be an effective solution and thus provides great potential in robotic applications. This paper proposes FAST-LIVO, a fast LiDAR-Inertial-Visual Odometry system, which builds on two tightly-coupled and direct odometry subsystems: a VIO subsystem and a LIO subsystem. The LIO subsystem registers raw points (instead of feature points on e.g., edges or planes) of a new scan to an incrementally-built point cloud map. The map points are additionally attached with image patches, which are then used in the VIO subsystem to align a new image by minimizing the direct photometric errors without extracting any visual features (e.g., ORB or FAST corner features). To further improve the VIO robustness and accuracy, a novel outlier rejection method is proposed to reject unstable map points that lie on edges or are occluded in the image view. Experiments on both open data sequences and our customized device data are conducted. The results show our proposed system outperforms other counterparts and can handle challenging environments at reduced computation cost. The system supports both multi-line spinning LiDARs and emerging solid-state LiDARs with completely different scanning patterns, and can run in real-time on both Intel and ARM processors. We open source our code and dataset of this work on Github² to benefit the robotics community.

I. INTRODUCTION

In recent years, simultaneous localization and mapping (SLAM) has made great progress in real-time 3D reconstruction and localization in unknown environments. Currently, there are several successful implemented framework using a single measurement sensor such as camera [1]–[3] or LiDAR [4]–[6]. However, with the increasing demand of operating intelligent robots in real world that usually contains challenging structure-less or texture-less environments, existing systems using a single sensor cannot achieve an accurate and robust pose estimation as required. To address this issue, multi-sensor fusion [7]–[10] has drawn increasing recent attention to combine the advantages of different sensors and provide an effective pose estimation in sensor degraded environments, showing great potentials in robotic applications.

Among those sensors used in robotics, camera, LiDAR and Inertial Measurement Units (IMUs) are possibly the mostly widely used sensors in SLAM tasks. Several recent LiDAR-inertial-visual odometry systems (LIVO) have been proposed to achieve robust state estimation such as R2LIVE

[11] and LVI-SAM [12]. They usually include a LiDAR-inertial odometry (LIO) subsystem and a visual-inertial odometry (VIO) subsystem that jointly fuses the state vector but separately process each data without considering their measurement-level coupling. The resultant system usually takes up significant computation resources. To address this issue, we propose a fast and tightly-coupled sparse-direct LiDAR-inertial-visual odometry system (FAST-LIVO), combining the advantage of sparse direct image align with direct raw points registration to achieve accurate and reliable pose estimation at a reduced computation cost. The contributions of this paper are listed as below:

- 1) A compact LiDAR-inertial-visual odometry framework, which builds on two direct and tightly-coupled odometry systems: a LIO subsystem and a VIO subsystem. These two subsystems estimates the system state jointly by fusing their respective LiDAR or visual data with IMUs.
- 2) A direct and efficient VIO subsystem that maximally re-use the point cloud map built in LIO subsystem. Specifically, points in the map are attached with image patches previously observed and then projected to a new image to align its pose (hence full system state), by minimizing the direct photometric errors. The LiDAR points re-use in VIO subsystem avoids the extraction, triangulation or optimization of visual features and couples the two sensors at the measurement level.
- 3) Implementation of the proposed system into a practical open software that can run in real-time on both Intel and ARM processors, and supports both multi-line spinning LiDARs and emerging solid-state LiDARs with completely different scanning patterns.
- 4) Verification of the developed system on both open data sequences (i.e., NTU VIRAL Dataset [13]) and our customized device data. Results show that our system outperforms other counterparts and can handle challenging sensor-degenerated environments at a reduced computation cost.

II. RELATED WORKS

Previous works related to our system and the involved techniques can be divided into the following two parts.

A. Direct Methods

Direct method is one of the most popular ways to achieve fast pose estimation in both visual and LiDAR SLAM. Different from feature-based methods ([4, 5, 7, 14]) that require to extract salient feature points (e.g., corners and edge

*These two authors contribute equally to this work.

¹C. Zheng, Q. Zhu, W. Xu, X. Liu, Q. Guo and F. Zhang are with the Department of Mechanical Engineering, The University of Hong Kong, Hong Kong Special Administrative Region, People's Republic of China. {zhengcr, xuwei, xliu, qguo, fzhang}@connect.hku.hk

²<https://github.com/hku-mars/FAST-LIVO>

pixels in images; plane and edge points in LiDAR scans), and generate robust feature descriptors for correspondence matching. Direct methods directly use raw measurements to optimize sensor pose [15] by minimizing an error function using photometric error or point-to-plane residuals, e.g., [16], [17]. This can achieve fast pose estimation by eliminating the feature extraction and matching that are usually time-consuming. Due to the lack of robust features matching, these direct methods are highly dependent on good state initialization. Dense direct method is mostly used in tracking for RGB-D cameras like [18, 19] and [20], which includes a dense depth-map in each image frame, and apply image-to-model alignment for estimation. Sparse direct method [16] shows a robust and accurate state estimation using only a few well-selected raw patches, thus further reduce the computation load compared by dense direct method.

Our work maintains the idea of direct method in both LIO and VIO. The LIO of our system is directly adapted from FAST-LIO2 [17]. The VIO subsystem is based on sparse direct image align similar to [16] but re-uses the map points built in the LIO subsystem to save the time-consuming backend (i.e., feature alignment, sliding window optimization and/or depth filtering).

B. LiDAR-Visual-Inertial SLAM

The multiple sensors used in LiDAR-visual-inertial SLAM make it possible to handle various of challenging environments where one sensor fails or is partially degenerated. Motivated by this, there have been several LiDAR-visual inertial SLAM systems developed in the community. To name a few, LIMO [21] extracts the depth information from LiDAR measurements for image features and estimates motion among keyframes based on bundle adjustment. Similarly, LIC-Fusion [22] fuses IMU measurements, sparse visual features, and LiDAR plane and edge features based on the MSCKF [23] framework. The subsequent LIC-Fusion2.0 [24] performs plane-feature tracking within a sliding window for better LiDAR pose estimation. VIL-SLAM [25] combines visual measurements obtained from frame-to-frame KLT tracking and ORB descriptor matching, and IMU measurements over a pose graph. The LIO module performs scan-to-map registration using the motion prior from the VIO module. Besides, it adds the loop closure module and the global pose graph optimization. VILENS [26] jointly optimizes visual, LiDAR, and inertial information within a single integrated factor graph, which uses fixed lag smoothing. R2LIVE [11] is a tightly-coupled system to fuse the LiDAR-visual-inertial sensors, which achieves an accurate state estimation by extracting LiDAR and image features and then conducts re-projection error within the framework of Iterated Error State Kalman Filter [27]. For the part of VIO subsystem, a sliding window optimization is used to further improve the accuracy of visual features in the map. LVI-SAM [12] is a feature-based framework similar to [11] but also incorporates loop closure. These feature-based methods are usually computationally-costly due to the extraction of image and/or LiDAR feature points and

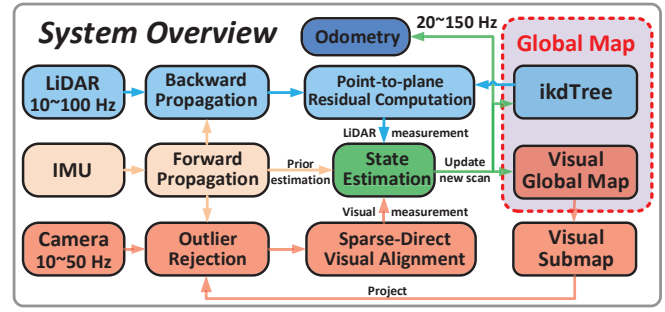


Fig. 1. System overview of FAST-LIVO

TABLE I

SOME IMPORTANT NOTATIONS

Notations	Meaning
$G(\cdot)$	The vector (\cdot) in global frame.
$C(\cdot)$	The vector (\cdot) in camera frame.
${}^I\mathbf{T}_L$	The extrinsic of LiDAR frame w.r.t. IMU frame.
${}^I\mathbf{T}_C$	The extrinsic of camera frame w.r.t. IMU frame.
$\mathbf{x}, \hat{\mathbf{x}}, \bar{\mathbf{x}}$	The ground-truth, predicted and updated estimation of \mathbf{x} .
$\delta\mathbf{x}$	The error state between ground-truth \mathbf{x} and its estimation.

the subsequent sliding window optimization. Compared with these methods, our proposed FAST-LIVO uses raw LiDAR points and image pixels to respectively track LiDAR scans and images without extracting any features, resulting in higher computation efficiency.

The VIO subsystem of our FAST-LIVO is most similar to DVL-SLAM [28, 29], which projects LiDAR points into a new image and tracks the image by minimizing the direct photometric error. However, DVL-SLAM conducts only frame-to-frame image alignment and does not incorporate any IMU measurements nor LiDAR scan registration. In contrast, our system FAST-LIVO tightly couples the frame-to-map image alignment, LiDAR scan registration, and IMU measurements in an integrated Kalman filter. We also propose a reliable and robust outlier rejection solution when projecting LiDAR measurements to image planes, which further improves the system accuracy.

III. SYSTEM OVERVIEW

This paper uses notations in Table I. The overview of our system is shown in Fig. 1, which contains two subsystems: the LIO subsystem (the blue part) and the VIO subsystem (the red part). The LIO subsystem first compensates the motion distortion in a LiDAR scan by backward propagation [30] and then computes the frame-to-map point-to-plane residual. Similarly, the VIO subsystem extracts the visual submap in the current FoV from the visual global map and rejects the outlier (points that are occluded or have depth discontinuity) in the submap. Then, a sparse-direct visual alignment is conducted to compute the frame-to-map image photometric errors. The LiDAR point-to-plane residual and the image photometric errors are tightly fused with the IMU propagation in an error-state iterated Kalman Filter. The fused pose is used to append new points to global map.

IV. STATE ESTIMATION

The state estimation of FAST-LIVO is a tightly-coupled error-state iterated Kalman filter (ESIKF) fusing measure-

ments from LiDAR, camera and IMU. Here we mainly explain the system model (state transition model and measurement model). Readers can refer to [31] for the detailed structure and implementation of an on-manifold ISIKF.

A. The boxplus “ \boxplus ” and boxminus “ \boxminus ” operator

In this section, we use the “ \boxplus ” and “ \boxminus ” operations to express error of state on a manifold \mathcal{M} . Specifically, for $\mathcal{M} = SO(3) \times \mathbb{R}^n$ considered in this paper, we have:

$$\begin{bmatrix} \mathbf{R} \\ \mathbf{a} \end{bmatrix} \boxplus \begin{bmatrix} \mathbf{r} \\ \mathbf{b} \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{R} \cdot \text{Exp}(\mathbf{r}) \\ \mathbf{a} + \mathbf{b} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{a} \end{bmatrix} \boxminus \begin{bmatrix} \mathbf{R}_2 \\ \mathbf{b} \end{bmatrix} \triangleq \begin{bmatrix} \text{Log}(\mathbf{R}_2^T \mathbf{R}_1) \\ \mathbf{a} - \mathbf{b} \end{bmatrix}$$

where $\mathbf{r} \in \mathbb{R}^3$, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, $\text{Exp}(\cdot)$ and $\text{Log}(\cdot)$ represent the bidirectional mapping between the rotation matrix and rotation vector derived from the Rodrigues' formula³.

B. State Transition Model

In our system, we assume the time offsets among the three sensors (LiDAR, IMU and camera) are known, which can be calibrated or synchronized in advance. We take IMU frame (denoted as I) as the body frame and the first body frame as the global frame (denoted as G). Besides, we assume that the three sensors are rigidly attached together and the extrinsics, defined in Table I, are pre-calibrated. Then, the discrete state transition model at the i -th IMU measurement:

$$\mathbf{x}_{i+1} = \mathbf{x}_i \boxplus (\Delta t \mathbf{f}(\mathbf{x}_i, \mathbf{u}_i, \mathbf{w}_i)) \quad (1)$$

where Δt is the IMU sample period, the state \mathbf{x} , input \mathbf{u} , process noise \mathbf{w} , and function \mathbf{f} are defined as follows:

$$\mathcal{M} \triangleq SO(3) \times \mathbb{R}^{15}, \quad \dim(\mathcal{M}) = 18 \quad (2)$$

$$\mathbf{x} \triangleq [{}^G\mathbf{R}_I^T \quad {}^G\mathbf{p}_I^T \quad {}^G\mathbf{v}^T \quad \mathbf{b}_g^T \quad \mathbf{b}_a^T \quad {}^G\mathbf{g}^T]^T \in \mathcal{M} \quad (3)$$

$$\mathbf{u} \triangleq [\omega_m^T \quad \mathbf{a}_m^T]^T, \quad \mathbf{w} \triangleq [\mathbf{n}_g^T \quad \mathbf{n}_a^T \quad \mathbf{n}_{bg}^T \quad \mathbf{n}_{ba}^T]^T \quad (4)$$

$$\mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{w}) = \begin{bmatrix} \omega_m - \mathbf{b}_g - \mathbf{n}_g \\ {}^G\mathbf{v} + \frac{1}{2}({}^G\mathbf{R}_I(\mathbf{a}_m - \mathbf{b}_a - \mathbf{n}_a) + {}^G\mathbf{g})\Delta t \\ {}^G\mathbf{R}_I(\mathbf{a}_m - \mathbf{b}_a - \mathbf{n}_a) + {}^G\mathbf{g} \\ \mathbf{n}_{bg} \\ \mathbf{n}_{ba} \\ \mathbf{0}_{3 \times 1} \end{bmatrix} \in \mathbb{R}^{18} \quad (5)$$

where ${}^G\mathbf{R}_I$ and ${}^G\mathbf{p}_I$ denote the IMU attitude and position in the global frame, ${}^G\mathbf{g}$ is the gravity vector in the global frame, ω_m and \mathbf{a}_m are the raw IMU measurements, \mathbf{n}_g and \mathbf{n}_a are measurement noises in ω_m and \mathbf{a}_m , \mathbf{b}_a and \mathbf{b}_g are IMU bias, which are modeled as random walk driven by Gaussian noise \mathbf{n}_{bg} and \mathbf{n}_{ba} , respectively.

C. Forward Propagation

We use the forward propagation to predict the state $\hat{\mathbf{x}}_{i+1}$ and its covariance $\hat{\mathbf{P}}_{i+1}$, at each IMU input \mathbf{u}_i . More specifically, the state is propagated by setting the process noise \mathbf{w}_i in (1) to zero:

$$\hat{\mathbf{x}}_{i+1} = \hat{\mathbf{x}}_i \boxplus (\Delta t \mathbf{f}(\hat{\mathbf{x}}_i, \mathbf{u}_i, \mathbf{0})). \quad (6)$$

³[https://en.wikipedia.org/wiki/Rodrigues' rotation formula](https://en.wikipedia.org/wiki/Rodrigues%27_rotation_formula)

with covariance propagated as below:

$$\begin{aligned} \hat{\mathbf{P}}_{i+1} &= \mathbf{F}_{\delta \hat{\mathbf{x}}} \hat{\mathbf{P}}_i \mathbf{F}_{\delta \hat{\mathbf{x}}}^T + \mathbf{F}_w \mathbf{Q} \mathbf{F}_w^T \\ \mathbf{F}_{\delta \hat{\mathbf{x}}} &= \left. \frac{\partial \delta \hat{\mathbf{x}}_{i+1}}{\partial \delta \hat{\mathbf{x}}_i} \right|_{\delta \hat{\mathbf{x}}_i=0, \mathbf{w}_i=0}, \quad \mathbf{F}_w = \left. \frac{\partial \delta \hat{\mathbf{x}}_{i+1}}{\partial \mathbf{w}_i} \right|_{\delta \hat{\mathbf{x}}_i=0, \mathbf{w}_i=0} \end{aligned} \quad (7)$$

where \mathbf{Q} is covariance of \mathbf{w} , $\delta \hat{\mathbf{x}}_i \triangleq \mathbf{x}_i \boxminus \hat{\mathbf{x}}_i$, and the concrete forms of $\mathbf{F}_{\delta \hat{\mathbf{x}}}$ and \mathbf{F}_w can be found in [11, 30].

The state prediction (6) and covariance (7) propagate from time t_{k-1} , where the last LiDAR or image measurements are received, until time t_k , where the current LiDAR or image measurements are received, with the reception of each IMU measurements \mathbf{u}_i in between t_{k-1} and t_k . The initial state and covariance in (6) and (7) are $\bar{\mathbf{x}}_{k-1}$ and $\bar{\mathbf{P}}_{k-1}$, which were obtained by fusing the last LiDAR or image measurement (see Section IV-E). We denote the state and covariance propagated until t_k as $\hat{\mathbf{x}}_k$ and $\hat{\mathbf{P}}_k$, respectively. Notice that we do not assume LiDAR scans and images are received at the same time. The arrival of either a LiDAR scan or image will cause an update of the state as detailed in Section IV-E.

D. Frame-to-map Measurement Model

1) *LiDAR Measurement Model*: If a LiDAR scan is received at time t_k , we first conduct backward propagation proposed in [30] to compensate the motion distortion. The resultant points $\{{}^L\mathbf{p}_j\}$ in the scan can be viewed as being sampled simultaneously at t_k and expressed in the same LiDAR local frame L . When registering the scan points $\{{}^L\mathbf{p}_j\}$ to the map, we assume each point lies on a neighboring plane in the map with normal \mathbf{u}_j and center point \mathbf{q}_j . That is, if transforming the measured ${}^L\mathbf{p}_j$ expressed in the LiDAR local frame to the global frame using the ground-truth state (i.e., pose) \mathbf{x}_k , the residual should be zero:

$$\mathbf{0} = \mathbf{r}_l(\mathbf{x}_k, {}^L\mathbf{p}_j) = \mathbf{u}_j^T ({}^G\mathbf{T}_{I_k}^T \mathbf{T}_L {}^L\mathbf{p}_j - \mathbf{q}_j) \quad (8)$$

In practice, in order to find the neighboring plane, we transform ${}^L\mathbf{p}_j$ to the global frame using pose in the predicted state $\hat{\mathbf{x}}_k$ by ${}^G\hat{\mathbf{p}}_j = {}^G\hat{\mathbf{T}}_{I_k}^T \mathbf{T}_L {}^L\mathbf{p}_j$ and search for the nearest 5 points in the LiDAR global map organized by an incremental kd-tree structure, *ikd-tree* [17], to fit a plane. Then, the equation in (8) defines an implicit measurement model for state \mathbf{x}_k . In order to account for the measurements noise in ${}^L\mathbf{p}_j$, the equation is weighted by a factor Σ_l .

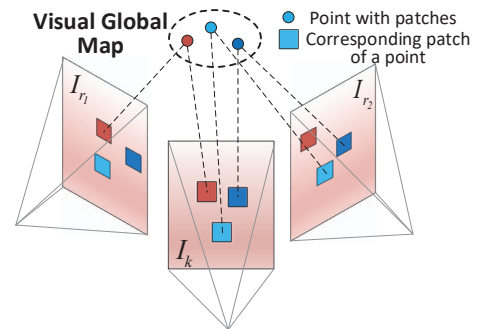


Fig. 2. Changing the pose of current frame I_k in the global frame to minimize the photometric errors between the reference image patches in the visual global map and the corresponding patch in the current frame.

2) *Sparse-Direct Visual Alignment Measurement Model*: Unlike the frame-to-frame image alignment in [16], we conduct sparse-direct frame-to-map image alignment by minimizing the photometric errors in a coarse-to-fine manner, see Fig. 2. Specifically, if an image is received at time t_k , we extract map points $\{^G\mathbf{p}_i\}$ from the global visual map, that fall within the image FoV (see Section V-B.2). For each map point $^G\mathbf{p}_i$, it has been attached with patches observed in different previous images (see Section V-B), we choose the patch contained in the image that observes the point with the closest observation angle with the current image as the reference patch (denoted as \mathbf{Q}_i). Also denote \mathbf{A}_i the affine wrapping matrix of \mathbf{Q}_i when projected to the current image pyramid. Then, transforming the map point $^G\mathbf{p}_i$ to the current image $\mathbf{I}_k(\cdot)$ with the ground-truth state (i.e., pose) \mathbf{x}_k , the photometric error between \mathbf{Q}_i and respective patch in the current image should be zero:

$$\mathbf{0} = \mathbf{r}_c(\mathbf{x}_k, ^G\mathbf{p}_i) = \mathbf{I}_k(\pi(^T\mathbf{T}_C^{-1} ^G\mathbf{T}_{I_k}^{-1} ^G\mathbf{p}_i)) - \mathbf{A}_i\mathbf{Q}_i \quad (9)$$

where $\pi(\cdot)$ is the pinhole projection model. The equation in (9) defines another implicit measurement model for state \mathbf{x}_k and is optimized (see Section IV-E) at three levels, where on each level the current image and reference patch is half sampled from the previous one. The optimization starts from the coarsest level, after the convergence of a level, the optimization proceeds to the next finer level. In order to account for the measurements noise in the image \mathbf{I}_k , the equation is weighted by a factor Σ_c .

E. Error-state Iterated Kalman Filter Update

The propagated state $\hat{\mathbf{x}}_k$ and covariance $\hat{\mathbf{P}}_k$ derived from Section IV-C impose the prior distribution for \mathbf{x}_k as follows:

$$\mathbf{x}_k \boxminus \hat{\mathbf{x}}_k \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{P}}_k). \quad (10)$$

Combining the prior distribution in (10), the measurement distribution for LiDAR measurement in (8) and visual measurement in (9), we obtain the maximum a posteriori (MAP) estimation for \mathbf{x}_k :

$$\min_{\mathbf{x}_k \in \mathcal{M}} \left(\|\mathbf{x}_k \boxminus \hat{\mathbf{x}}_k\|_{\hat{\mathbf{P}}_k}^2 + \sum_{j=1}^{m_l} \|\mathbf{r}_l(\mathbf{x}_k, ^L\mathbf{p}_j)\|_{\Sigma_l}^2 + \sum_{i=1}^{m_c} \|\mathbf{r}_c(\mathbf{x}_k, ^G\mathbf{p}_i)\|_{\Sigma_c}^2 \right) \quad (11)$$

where $\|\mathbf{x}\|_{\Sigma}^2 = \mathbf{x}^T \Sigma^{-1} \mathbf{x}$. Notice that if a LiDAR scan is received at t_k , (11) fuses only LiDAR residual \mathbf{r}_l with IMU propagation (i.e., $m_c = 0$). Similarly, if an image is received at t_k , (11) fuses only visual photometric error \mathbf{r}_c with IMU propagation (i.e., $m_l = 0$).

The optimization in (11) is non-convex and can be iteratively solved by a Gauss-Newton method. Such an iterative optimization has been proven to be equivalent to an iterated Kalman filter [27]. To deal with the manifold constraint \mathcal{M} , in each optimization iteration, we parameterize the state in the tangent space (i.e., the error state) of the current state estimate via the \boxminus operation in Section IV-A. The solved error state then updates the current state estimate

and proceeds to the next iteration until convergence. The converged state estimate, denoted as $\bar{\mathbf{x}}_k$, and the Hessian matrix of (11) at convergence, denoted as $\bar{\mathbf{P}}_k$, are used to propagate the incoming IMU measurements as described in Section IV-C. The converged state is also used to update the new LiDAR scan to global map in Section V-A and in Section V-B for visual global map.

V. MAP MANAGEMENT

Our map consists of a point cloud map (the LiDAR global map) for the LIO subsystem and a point cloud map attached with patches (the visual global map) for the VIO subsystem.

A. LiDAR Global Map

Our LiDAR global map is adopted from FAST-LIO2 [17], which consists of all past 3D points organized into an incremental k-d tree structure *ikd-Tree* [32]. The *ikd-Tree* provides interfaces of points inquiry, insertion, and delete. It also internally down samples the point cloud map at a given resolution, repeatedly monitors its tree structure, and dynamically balances the tree structure by rebuilding the respective sub-trees. When receiving a new LiDAR scan, we poll each point transformed with the predicted pose in the *ikd-Tree* for nearest points (Section IV-D.1). After the scan is fused with IMU to obtain $\bar{\mathbf{x}}_k$ (Section IV-E), we use it to transform the scan points to the global frame and insert them to the *ikd-Tree* at the LIO rate.

B. Visual Global Map

The visual global map is a collection of LiDAR points previously observed. Each point is attached with multiple patches from the images observing it. The data structure and update of the visual global map are explained below:

1) *Data Structure*: To quickly find the visual map points falling in the current FoV, we use axis-aligned voxels to contain the points in visual global map. Voxels are of the same size ($0.5 \times 0.5 \times 0.5$ meters) and organized by a Hash table for fast indexing. A point contained in a voxel is saved with its position, three-level patch pyramids ($8 \times 8 \times 3$ pixels) extracted from different reference images, and the camera pose of each patch pyramids.

2) *Visual Submap and Outlier Rejection*: Even the number of voxels is much less than that of visual map points, determining which of them are within current frame FoV could still be very time consuming, especially when the map points (hence voxels) are large in number. To address this issue, we poll those voxels which contain LiDAR points in the last scan. This can be done very efficiently by inquiring the voxel Hash table using the measured points. If the camera FoV is largely overlapped with the LiDAR FoV, map points falling in the camera FoV are mostly likely contained in these voxels. Hence, the visual submap can be obtained by the points contained in these voxels followed by a FoV check.

The visual submap could contain map points that are occluded in the current image frame or have discontinuous depth, which severely degrade the VIO accuracy. To address this issue, we project all the points in the visual submap

TABLE II
ABSOLUTE TRANSLATIONAL ERRORS (RMSE, METERS) IN NTU-VIRAL SEQUENCES WITH GOOD QUALITY GROUND TRUTH

	eee_01	eee_02	eee_03	nya_01	nya_02	nya_03	sbs_01	sbs_02	sbs_03
FAST-LIVO	0.19	0.13	0.19	0.12	0.18	0.11	0.25	0.13	0.13
FAST-LIO2	0.21	0.17	0.21	0.14	0.21	0.13	0.18	0.16	0.14
SVO2.0 (edgelets+prior)	Fail	Fail	4.12	2.29	2.91	3.32	7.84	Fail	Fail
R2LIVE	0.45	0.21	0.97	0.19	0.63	0.31	0.56	0.24	0.44
DVL-SLAM (no loop closure)	2.88	1.65	3.08	2.09	1.45	1.82	1.08	2.31	2.23

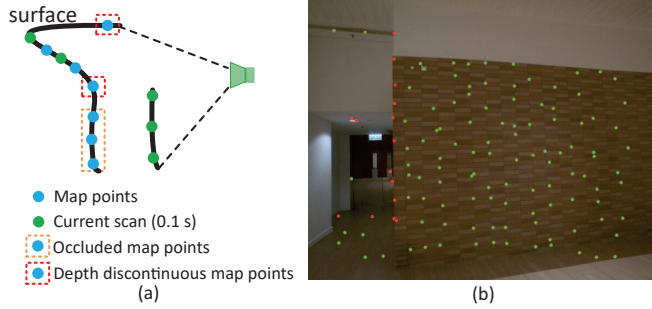


Fig. 3. Outlier rejection. (a) shows the diagrammatic drawing of occluded and depth-discontinuous map points. (b) shows the effect visualization of outlier rejection module in real scenes. The red dots are the rejected map points, and the green dots are the accepted map points.

onto the current frame using the predicted pose in $\hat{\mathbf{x}}_k$ and keep the lowest-depth points in each grid of 40×40 pixels. Furthermore, we project the points in the most recent LiDAR scan to the current frame and check if they occlude any map points projected within 9×9 neighbor by examining their depth. Occluded map points are rejected (see Fig. 3) and the rest will be used to align the current image (Section IV-D.2).

3) *Update Visual Global Map*: After a new image frame is aligned (Section IV-D.2), we attach patches from the current image to map points within the FoV, so that the map points will likely have effective patches with uniformly distributed viewing angles. Specifically, we select map points with high photometric errors after the frame alignment, if it has been more than 20 frames from the last time the map point is added with a patch or the map point in the current frame is more than 40 pixels away from its pixel position in the last reference frame where a patch was added, we add a new patch to it. The new patch is extracted from the current image with a size of 8×8 pixels. Along with the patch pyramid, we also attach the frame pose to the map point.

Besides adding patches to the map points, we also need to add new map points to the visual global map. To do so, we divide the current image into grids of 40×40 pixels and project on it the points in the most recent LiDAR scan. Projected LiDAR points with the highest gradient in each grid will be added to the visual global map, along with a patch extracted there and the image pose. To avoid LiDAR points on edges to be added to the visual map, we skip edge points with high local curvature [4, 5].

VI. EXPERIMENT AND RESULTS

In this section, we verify our proposed method in both open and private datasets.

A. Benchmark Dataset

In this section, quantitative experiments are conducted on all the nine sequences of NTU-VIRAL open datasets [13].

The sensors used in the dataset are the left camera, the horizontal 16-channel OS1 gen1⁴ LiDAR and its internal IMU. We compare our method with various open-source odometry system, including R2LIVE [11], a feature-based, full LiDAR-inertial-visual odometry, FAST-LIO2 [17], a direct, LiDAR-inertial odometry, SVO2.0 [33], a semi-direct, visual-inertial odometry, and DVL-SLAM [29], a direct, LiDAR-visual SLAM system. All these systems are downloaded from their respective Github websites where the FAST-LIO2, R2LIVE, and DVL-SLAM all use their recommended parameters for outdoor scenarios. The parameters of SVO2.0, including grid size, feature-selection threshold, and sliding window size, are tuned to to attain the best results to the authors' efforts. Moreover, since all systems in comparison are odometry with no loop closure, except DVL-SLAM, we remove the loop-closure module of DVL-SLAM to make a fair comparison. The modified DVL-SLAM, which we also opened on Github⁵, keeps the sliding window optimization to retain the accuracy as much as possible.

The results of all methods are shown in Table II, where each method uses the same parameters across all sequences. As can be seen, our method achieves the best accuracy among all sequences except the sequence "sbs.01", where our system has a slightly higher error than the LiDAR-inertial only odometry, FAST-LIO2. This is because the image is extremely blurred due to the fast UAV motion, fusing these low-quality images did not help on the odometry accuracy. Other than "sbs.01", our method benefits from the tight couple of all LiDAR, inertial, and visual information and its performance surpasses the LiDAR-inertial only odometry, FAST-LIO2.0 (by an extent depending on the image quality) and also other subsystems (e.g., visual-inertial only odometry, SVO2.0, and LiDAR-visual only odometry, DVL-SLAM) on all the rest sequences. Moreover, our system achieves higher accuracy than R2LIVE, a state-of-the-art feature-based tightly-coupled full LiDAR-inertial-visual odometry system. It is also interesting to see that R2LIVE occasionally achieves lower accuracy than the LiDAR-inertial only odometry, FAST-LIO2. The reason is due to the feature-based LIO subsystem of R2LIVE, which leads to fewer LiDAR points to be registered, while FAST-LIO2 has a direct LIO subsystem that registers more raw LiDAR points. The fusion of images in R2LIVE can compensate this loss of accuracy and therefore its performance surpasses FAST-LIO2 in some sequences. Finally, it is noted that SVO2.0 fails in eee_01, eee_02, sbs_02 and sbs_03 due to the insufficient prior

⁴<https://ouster.com/products/scanning-lidar/os1-sensor/>

⁵<https://github.com/xuankuzcr/DVL-SLAM.ROS>

information caused by the fast UAV rotation and the strong image blur, while DVL-SLAM survived in these sequences due to the use of LiDAR measurements in its VO module.

B. Private Dataset

1) *Equipment Setup*: Our sensor suite for data collection is shown in Fig. 4, which consists of an onboard computer DJI manifold-2c (Intel i7-8550u CPU and 8 GB RAM), two industrial cameras (MV-CA013-21UC), and a Livox Avia LiDAR. All sensors are hard synchronized with a trigger signal at 10 Hz generated by STM32 synchronized timers. Notice that Livox converter only supports 1 Hz pulse per second (PPS) trigger signal, and Livox SDK will automatically divide it into 10 Hz.

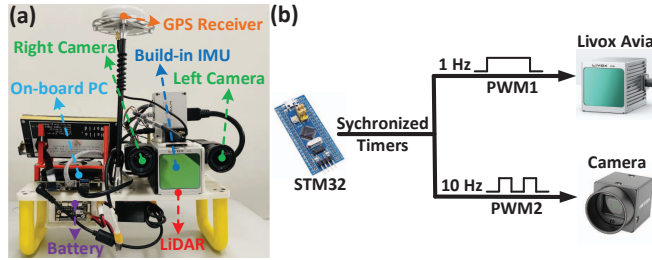


Fig. 4. Our platform with hardware synchronization for data acquisition. (a) our hardware system, (b) the hardware synchronization diagram.

2) *LiDAR Degenerated Experiment*: In this experiment, we evaluate our system in a LiDAR degenerated environment, facing a wall with a length of about 30 meters. The results are shown in Fig. 5. Due to the lack of constraints in the direction along the wall, the LiDAR-inertial odometry, FAST-LIO2, cannot have a good state estimation in this direction, especially only a side wall can be seen in the FoV. Similarly, the highly repeated visual features on the wall and insensitive visual constraints in the vertical direction of the wall bring drifts to visual-inertial odometry, SVO2.0. Compared with FAST-LIO2 and SVO2.0, our system is robust and accurate to handle the scene and achieves the best performance, with the lowest end-to-end drift of 0.05m.

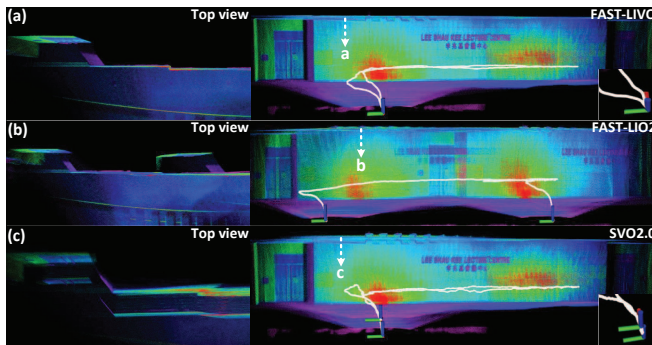


Fig. 5. The performance of FAST-LIVO, FAST-LIO2, and SVO2.0 in a LiDAR degeneration scene. The point cloud is colored by LiDAR intensity. (a), (b) and (c) represent the top view of the corresponding position a, b, c respectively.

3) *Visual Challenge Experiment*: In this experiment, we challenge an extremely difficult scene in visual SLAM. As shown in Fig. 6, the motion of the sensor includes an indoor-to-outdoor process, an outdoor-to-indoor process, two aggressive motions and a visual texture-less white wall. Our

proposed algorithm can reliably survives in this challenging scene and returns to the starting point with an end-to-end error of 0.04 m over the total path length 79.52 m.

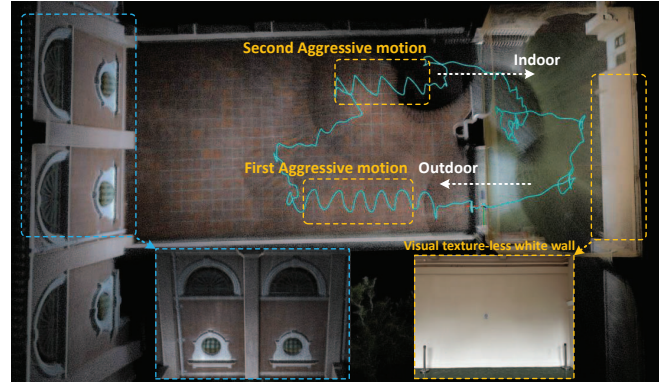


Fig. 6. The RGB-colored point cloud of the visual challenging scene containing aggressive motions, indoor-to-outdoor and outdoor-to-indoor movements, and a visual texture-less wall. The green path is the computed trajectory.

4) High Precision Mapping with Colored Point Cloud:

In this experiment, we take advantage of our algorithm to reconstruct, in real-time, a precise, dense, 3D, and RGB-colored map of a HKU campus environment, which is shown in Fig. 7. What worth to be mentioned is, the enlarged view of the colored point cloud has fine visual details similar to the actual RGB image.

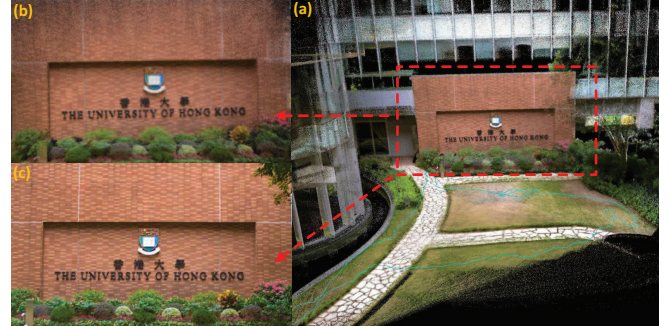


Fig. 7. (a) RGB-colored point cloud, (b) partially enlarged view of the colored point cloud and (c) the corresponding RGB image.

TABLE III

MEAN TIME CONSUMPTION IN MILLISECONDS

	Intel i7	ARM
VIO subsystem	Visual Submap	3.81
	Outlier Rejection	2.58
	Sparse-Direct Visual Alignment	3.88
	ESIKF Update	0.52
	Update Visual Global Map	3.03
	Total Time	13.82
LIO subsystem	26.52	51.51

C. Time Analysis

In this section, we evaluate the computation efficiency of our full system FAST-LIVO and compare it with a feature-based LiDAR-inertial-visual odometry system, R2LIVE. The comparison is performed among all private datasets on a desktop PC with a 8-core Intel Core i7-10700U processor, the mean time consumption per LiDAR and image frame of R2LIVE is 45.16 ms for the LIO and VIO frontend plus 59.27 ms for the VIO backend sliding window optimization,

while FAST-LIVO is only 36.75 ms for the complete processing. Table III further displays a mean time breakdown of FAST-LIVO on both the desktop PC and an embedded computation platform RB5 with a Qualcomm Kryo585 CPU. As can be seen, the FAST-LIVO can run in real-time on both the Intel and ARM processors with a significant computation margin.

VII. CONCLUSION

This paper proposes a fast, robust, sparse-direct, frame-to-map LiDAR-inertial-visual fusion framework, which is faster than the current state-of-the-art LIVO algorithms. We fuse the measurements of LiDAR, inertial, and camera sensors within an error-state iterated Kalman filter, utilizing the patch-based photometric error. Our system was tested in aggressive motions, indoor-outdoor, texture-less white wall, and LiDAR degenerated environments. The experiments in open datasets show our system can achieve the best overall performance among the state-of-the-art LIO, VIO, and LIVO algorithms.

ACKNOWLEDGMENT

The authors gratefully acknowledge Livox Technology for the equipment support during the whole work. This work is supported in part by the University Grants Committee of Hong Kong General Research Fund under Project 17206421. The authors would like to thank Jiarong Lin for the helps in the experiments.

REFERENCES

- [1] R. Mur-Artal and J. D. Tardós, “Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras,” *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [2] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.
- [3] J. Engel, T. Schöps, and D. Cremers, “Lsd-slam: Large-scale direct monocular slam,” in *European conference on computer vision*. Springer, 2014, pp. 834–849.
- [4] J. Zhang and S. Singh, “Loam: Lidar odometry and mapping in real-time,” in *Robotics: Science and Systems*, vol. 2, no. 9, 2014.
- [5] J. Lin and F. Zhang, “Loam livox: A fast, robust, high-precision lidar odometry and mapping package for lidars of small fov,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3126–3131.
- [6] T. Shan and B. Englot, “Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4758–4765.
- [7] 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.
- [8] C. Qin, H. Ye, C. E. Pranata, J. Han, S. Zhang, and M. Liu, “Lins: A lidar-inertial state estimator for robust and efficient navigation,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8899–8906.
- [9] J. HYUN and H. Myung, “Uwb-inertial slam based on tightly-coupled ekf framework,” in *Int’l Conf. on Ubiquitous Robots (UR) 2020*. Korea Robot Society, 2020.
- [10] A. Kramer, C. Stahoviak, A. Santamaria-Navarro, A.-A. Agha-Mohammadi, and C. Heckman, “Radar-inertial ego-velocity estimation for visually degraded environments,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 5739–5746.
- [11] J. Lin, C. Zheng, W. Xu, and F. Zhang, “R²live: A robust, real-time, lidar-inertial-visual tightly-coupled state estimator and mapping,” *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7469–7476, 2021.
- [12] T. Shan, B. Englot, C. Ratti, and D. Rus, “Lvi-sam: Tightly-coupled lidar-visual-inertial odometry via smoothing and mapping,” in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 5692–5698.
- [13] T.-M. Nguyen, S. Yuan, M. Cao, Y. Lyu, T. H. Nguyen, and L. Xie, “Ntu viral: A visual-inertial-ranging-lidar dataset, from an aerial vehicle viewpoint,” *International Journal of Robotics Research (accepted, to appear)*, 2021.
- [14] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, “Orb-slam: a versatile and accurate monocular slam system,” *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [15] M. Irani and P. Anandan, “All about direct methods,” in *Proc. Workshop Vis. Algorithms, Theory Pract*, 1999, pp. 267–277.
- [16] C. Forster, M. Pizzoli, and D. Scaramuzza, “Svo: Fast semi-direct monocular visual odometry,” in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 15–22.
- [17] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, “Fast-lio2: Fast direct lidar-inertial odometry,” *IEEE Transactions on Robotics*, pp. 1–21, 2022.
- [18] M. Meilland, A. I. Comport, and P. Rives, “Real-time dense visual tracking under large lighting variations,” in *British Machine Vision Conference*. British Machine Vision Association, 2011, pp. 45–1.
- [19] T. Tykkälä, C. Audras, and A. I. Comport, “Direct iterative closest point for real-time visual odometry,” in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 2050–2056.
- [20] C. Kerl, J. Sturm, and D. Cremers, “Robust odometry estimation for rgb-d cameras,” in *2013 IEEE international conference on robotics and automation*. IEEE, 2013, pp. 3748–3754.
- [21] J. Graeter, A. Wilczynski, and M. Lauer, “Limo: Lidar-monocular visual odometry,” in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2018, pp. 7872–7879.
- [22] X. Zuo, P. Geneva, W. Lee, Y. Liu, and G. Huang, “Lic-fusion: Lidar-inertial-camera odometry,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5848–5854.
- [23] K. Sun, K. Mohta, B. Pfrommer, M. Watterson, S. Liu, Y. Mulgaonkar, C. J. Taylor, and V. Kumar, “Robust stereo visual inertial odometry for fast autonomous flight,” *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 965–972, 2018.
- [24] X. Zuo, Y. Yang, P. Geneva, J. Lv, Y. Liu, G. Huang, and M. Pollefeys, “Lic-fusion 2.0: Lidar-inertial-camera odometry with sliding-window plane-feature tracking,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5112–5119.
- [25] W. Shao, S. Vijayarangan, C. Li, and G. Kantor, “Stereo visual inertial lidar simultaneous localization and mapping,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 370–377.
- [26] D. Wisth, M. Camurri, S. Das, and M. Fallon, “Unified multi-modal landmark tracking for tightly coupled lidar-visual-inertial odometry,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1004–1011, 2021.
- [27] C. F. W. Bell B.M., “The iterated kalman filter update as a gauss-newton method,” *Automatic Control IEEE Transactions*, vol. 38, no. 2, pp. 294–297, 1993.
- [28] Y.-S. Shin, Y. S. Park, and A. Kim, “Direct visual slam using sparse depth for camera-lidar system,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5144–5151.
- [29] —, “Dvl-slam: sparse depth enhanced direct visual-lidar slam,” *Autonomous Robots*, vol. 44, no. 2, pp. 115–130, 2020.
- [30] W. Xu and F. Zhang, “Fast-lio: A fast, robust lidar-inertial odometry package by tightly-coupled iterated kalman filter,” *IEEE Robotics and Automation Letters*, pp. 1–1, 2021.
- [31] D. He, W. Xu, and F. Zhang, “Embedding manifold structures into kalman filters,” *arXiv preprint arXiv:2102.03804*, 2021.
- [32] Y. Cai, W. Xu, and F. Zhang, “ikd-tree: An incremental kd tree for robotic applications,” *arXiv preprint arXiv:2102.10808*, 2021.
- [33] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, “Svo: Semidirect visual odometry for monocular and multicamera systems,” *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2016.