LE-VINS: A Robust Solid-State-LiDAR-Enhanced Visual-Inertial Navigation System for Low-Speed Robots

Hailiang Tang, Xiaoji Niu, Tisheng Zhang, Liqiang Wang, and Jingnan Liu

Abstract—**Accurate and long-distance depth estimation for visual landmarks is challenging in visual-inertial navigation systems (VINS). In visual-degenerated scenes with illumination changes, moving objects, or weak texture, depth estimation may be more difficult, resulting in poor robustness and accuracy. For low-speed robot navigation, we present a solid-state-LiDAR-enhanced VINS (LE-VINS) to improve the system robustness and accuracy in challenging environments. The point clouds from the solid-state LiDAR are projected to the visual keyframe with the inertial navigation system (INS) pose for depth association while compensating for the motion distortion. A robust depth-association method with an effective plane-checking algorithm is proposed to estimate the landmark depth. With the estimated depth, we present a LiDAR depth factor to construct accurate depth measurements for visual landmarks in factor graph optimization (FGO). The visual feature, LiDAR depth, and IMU measurements are tightly fused within the FGO framework to achieve maximum-a-posterior estimation. Field tests were conducted on a low-speed robot in large-scale challenging environments. The results demonstrate that the proposed LE-VINS yields significantly improved robustness and accuracy compared to the original VINS. Besides, LE-VINS exhibits superior accuracy than the state-of-the-art LiDAR-visual-inertial navigation system. LE-VINS also outperforms the existing LiDAR-enhanced method, benefiting from the robust depth-association algorithm and the effective LiDAR depth factor.**

Index Terms—**Multi-sensor fusion navigation, LiDAR depth enhancement, visual-inertial navigation system, factor graph optimization, mobile robot localization.**

I. Introduction

Visual navigation system has been widely used in autonomous mobile robots. For visual navigation based on a monocular camera, the rotation can be recovered except for the translation scale [1]. An inertial measurement unit (IMU) can be incorporated to construct a visual-inertial navigation system (VINS) [2], [3] and retrieve the scale. However, low-cost micro-electro-mechanical system (MEMS) IMU suffers from various errors [4], especially the time-varying biases [5], making VINS challenging. Typically, the depth of the visual landmark can be triangulated by the tracked features using the camera pose. However, the camera pose is usually predicted by the MEMS inertial navigation system (INS), which might cause an inaccurate estimation of the depth. Besides, the landmark depth can be estimated accurately only when the parallax is enough. In addition, drastic illumination changes and moving objects may affect the depth-estimation accuracy. Multiple cameras can be used to estimate the depth without extra motion directly [3], [6], [7], but they may still be affected by external environmental factors. Besides, if the baselines of the cameras are too small, the estimated depth might also be inaccurate and distance-limited. In addition, precise calibration is required for multiple cameras to obtain an accurate depth. Hence, inaccurately estimated depth might frequently occur in visual navigation systems, resulting in poor robustness and accuracy, especially in challenging environments with illumination changes, moving objects, or weak texture.

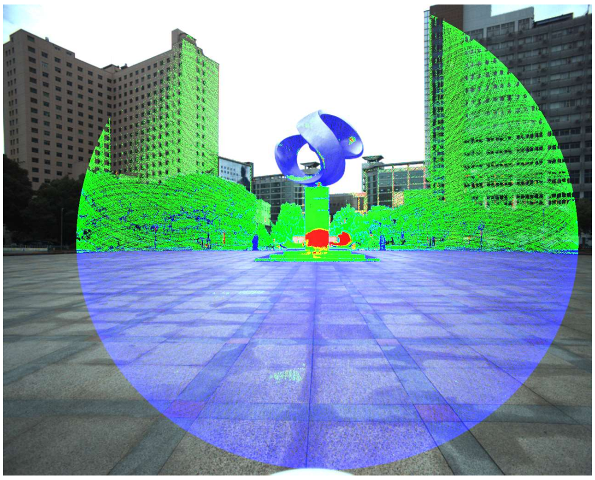

Fig. 1. An illustration of the FOV of the Solid-State LiDAR (Livox Mid-70 with 70° circular FOV) and the camera (80° horizontal FOV). The color is rendered by the reflection intensity of the point clouds. The point clouds are accumulated for 20 seconds. The misalignment is caused by the LiDAR-camera extrinsic parameters.

This research is partly funded by the National Key Research and Development Program of China (No. 2020YFB0505803) and the National Natural Science Foundation of China (No. 41974024). *(Corresponding author: Tisheng Zhang.)*

Hailiang Tang and Liqiang Wang are with the GNSS Research Center, Wuhan University, Wuhan 430079, China (e-mail: thl@whu.edu.cn; wlq@whu.edu.cn).

Xiaoji Niu is with the GNSS Research Center, Wuhan University, Wuhan 430079, China, the Hubei Luojia Laboratory, Wuhan 430079, China, and the Artificial Intelligence Institute of Wuhan University, Wuhan 430079, China (email: xjniu@whu.edu.cn).

Tisheng Zhang and Jingnan Liu are with the GNSS Research Center, Wuhan University, Wuhan 430079, China, and the Hubei Luojia Laboratory, Wuhan 430079, China (e-mail: zts@whu.edu.cn; jnliu@whu.edu.cn).

The RGB-D camera and light detection and ranging (LiDAR) have been widely used to enhance the visual navigation system, which can directly measure accurate depth. RGB-D camera is a particular sensor capable of simultaneously providing RGB and depth images [8], [9]. However, a standard RGB-D camera can only measure depth within several meters, and thus it is commonly used for indoor applications. LiDAR can directly obtain long-distance and centimeter-level depth measurements. It has been incorporated into the visual navigation system to provide depth estimation for visual landmarks in recent years [1], [10]–[16]. The spinning 3D LiDAR can achieve a 360° horizontal field of view (FOV) while a smaller vertical FOV, such as 30° for Velodyne VLP-16. In addition, the spinning LiDAR provides limited laser beams, such as 16, 32, 64, and 128, leading to small overlapping areas regarding the FOV of the camera and LiDAR. The sparse LiDAR makes it highly challenging to associate visual features or textures with LiDAR depth. Using LiDAR with more laser beams can mitigate this effect. In [1], [10]–[12], 64-beam LiDAR is adopted to provide depth for their feature-based visual navigation systems. However, the more laser beams the LiDAR has, the more expensive it is. In [15], [16], the landmark depth is associated with the map built by LiDAR, but the accuracy might be affected by the quality of the built map. Another solution to incorporate a sparse spinning LiDAR into visual navigation systems is using direct methods [13], [14].

In recent years, low-cost solid-state LiDAR based on the non-repetitive pattern, such as Livox AVIA and Mid-70, has been widely used in LiDAR navigation systems [17]–[21], and LiDAR-visual navigation systems [22]–[25]. Due to the non-repetitive pattern, solid-state LiDAR can share large overlapping areas with the camera, as shown in Fig. 1. The non-repetitive scanning pattern of solid-state LiDAR can maximize the coverage ratio. Hence, we can obtain a relatively dense map by accumulating point clouds for several scans, as depicted in Fig. 1. Due to this reason, solid-state LiDAR has been adopted to provide depth in recent LiDAR-visual navigation systems [22], [24], [25], including feature-based methods [22] and direct methods [24], [25].

However, the accurate LiDAR depth has not been fully used in most of the current LiDAR-enhanced methods. The LiDAR depth is only used as an initial value for visual landmarks in some methods [11], [22], which wastes the accurate depth to a certain extent. The depth is set to a constant and would not be optimized in both the feature-based methods [15], [16] and the direct methods [13], [14], [24], [25]. However, the LiDAR depth also contains noise; thus, it is unreasonable to set it as a constant. A cost function is adopted to punish the deviation of the landmark depth from the LiDAR depth in [10]. Still, this constraint is not directly applied to the landmark depth because the landmark is parameterized as a 3D position, which might introduce extra errors. Besides, this constraint cannot be reserved in their optimization problem [10] once the keyframe is removed from the optimization window, resulting in a loss of information. Hence, the LiDAR depth must be utilized more reasonably and thoroughly.

For low-speed robots with a speed of several meters per second, we propose a solid-state-LiDAR-enhanced visual-inertial navigation system (LE-VINS) to achieve a real-time, robust, and accurate positioning in large-scale challenging environments. The solid-state LiDAR with the non-repetitive pattern can generate relatively dense point clouds for low-speed robots, which is conducive to depth association. We propose a robust depth-association method to estimate accurate depths for visual landmarks. The associated depth is employed to construct a LiDAR depth factor to constrain the inverse-depth parameter of the landmark directly [26] in the proposed factor graph optimization (FGO) [27]. If the landmark is marginalized, the accurate depth constraint can be reserved by converting the LiDAR depth factor into the prior factor. The main contributions of our work are as follows:

- A robust solid-state-LiDAR-enhanced visual-inertial navigation system is proposed for low-speed robots. The solid-state-LiDAR with the non-repetitive scanning pattern is employed to provide accurate depths for visual landmarks. The visual feature, LiDAR depth, and IMU measurements are tightly fused within the FGO framework to achieve maximum-a-posterior estimation.
- The LiDAR points are projected to the visual keyframe for depth association with the accurate INS pose while compensating for the motion distortion. A robust depth-association method with an effective plane-checking algorithm is proposed to estimate and verify landmark depths, significantly improving the accuracy of the estimated depths and avoiding wrong associations.
- The estimated depth is not only adopted as the initial depth of the landmark but also employed to construct a LiDAR depth factor in the FGO to constrain the landmark depth directly. If the landmark is marginalized, the LiDAR depth factor can be converted into the prior factor, and thus the accurate constraints can be reserved.
- Field tests were conducted to evaluate the proposed LE-VINS in large-scale challenging environments using a low-speed robot. The results demonstrate that the proposed LE-VINS yields improved robustness and accuracy compared to the original VINS. Besides, the proposed LE-VINS outperforms the existing LiDAR-visual-inertial navigation system and LiDAR-enhanced method.

The remainder of this paper is organized as follows. The next section discusses a literature review on LiDAR-enhanced visual systems. We give an overview of the system pipeline in section III. The proposed solid-state-LiDAR-enhanced VINS is presented in section IV. The experiments and results are discussed in section V for quantitative evaluation. Finally, we conclude the proposed LE-VINS.

II. Related Works

In LiDAR-enhanced visual systems, LiDAR is commonly adopted to provide accurate depth for visual systems. According to the role of the visual subsystems, LiDAR-enhanced visual systems can be classified into feature-based methods [1], [10]–[12], [15], [22], [28], [29] and direct methods [13], [14], [24], [25]. Here, some tightly-coupled methods [12], [24], [25] are also involved due to the use of LiDAR enhancement. In feature-based methods, the depth from the LiDAR is associated with the feature points or feature lines to achieve state estimation by bundle adjustment. Direct methods use the LiDAR depth for the visual pixel, and the pose optimization is conducted by minimizing the photometric error.

1) *Feature-based Methods*

In feature-based methods, the visual features are first extracted, including feature points and feature lines. Different depth-association methods are employed to estimate the feature depth from the LiDAR. The estimated depth will be incorporated into the state estimator to improve the accuracy.

In DEMO [1], LiDAR is utilized to provide depth for their feature-based visual odometry. The visual features and point clouds are all projected to a sphere with a unit distance to the camera center for depth association. Besides, the measurement noise of the LiDAR is adopted to weigh the depth residual of the features in the bundle adjustment. The same depth-association method is employed in [15], [16]. However, the LiDAR depth is used as a constant in the FGO [15], [16], and the depth-association error has not been considered.

Unlike DEMO, in LIMO [10], the depth estimation is conducted in the image plane using the 64-beam spinning LiDAR in KITTI datasets [30]. This method is unsuitable for LiDAR with fewer laser beams, such as 16-beam LiDAR. In addition, the landmark depth from the LiDAR is added to the optimizer by punishing the deviation of the landmark depth from the measured depth, which can help to estimate the odometry scale. However, such constraints in LIMO can only be applied within the optimization window and cannot be reserved once the landmarks are removed, which results in a loss of information. The depth-association method in LIMO is adopted in [12] to construct a mono landmark factor with the LiDAR depth in Euclidean space. Similarly, in [11], feature lines are used together with feature points in the LiDAR-monocular visual odometry, in which depth estimation is conducted in the image plane using a 64-beam LiDAR. However, the accurate depth from the LiDAR is only treated as a depth prior in [11], which results in a loss of accuracy.

A novel voxel-map-based depth-association method was proposed in a vanishing point-aided LiDAR-visual-inertial system [28]. Besides, the work in [28] is further integrated into Super Odometry [29]. However, the accurate depth from the LiDAR is only used as a prior depth. Hence, the accurate depth information is wasted in the state estimation [28], [29].

CamVox [22] uses a non-repetitive solid-state LiDAR, Livox Horizon, to build a dense, accurate, and long-range depth image. The depth image with the RGB image is incorporated into ORB_SLAM2 [6] to achieve an RGB-D simultaneous localization and mapping (SLAM). However, Livox Horizon has a vertical FOV of 25°, which is far smaller than the typical FOV of a camera (70° or even larger). Hence, the visual information has been partially wasted.

In these feature-based LiDAR-enhanced visual navigation systems, the depth association and the use of the associated depth are the two critical parts. Some depth-association methods are platform-specific, e.g. 64-beam spinning LiDAR for [10]–[12], [28], [29], and non-repetitive solid-state LiDAR for [22]. The non-repetitive solid-state LiDAR has not been thoroughly studied in current methods. As for the use of the associated depth, it has been used insufficiently in these methods. The LiDAR depth is only employed as a depth prior in [11], [28], [29], or a constant in [15]. Though the measurement noise of the LiDAR has been partially considered in [1], [10], [22], these methods can still be improved to utilize the accurate LiDAR depth fully.

2) *Direct Methods*

The 64-beam spinning LiDAR is adopted in most feature-based LiDAR-enhanced visual navigation systems because such LiDAR can provide relatively dense point clouds. However, it is challenging to associate features using a sparse LiDAR, such as 16-beam spinning LiDAR, which is much cheaper than 64-beam spinning LiDAR. Consequently, LiDAR-enhanced direct visual navigation has been proposed to employ the sparse LiDAR [13], [14], [24], [25].

In [13], direct laser-visual odometry was proposed by building upon the photometric-image alignment with occlusion handling and plane detection, which can utilize a 16-beam LiDAR. However, it [13] relies much on the LiDAR, and the non-overlapping area of the image cannot be used, which may result in robustness and accuracy degradation in challenging environments due to the limited FOV of LiDAR. Similarly, DVL-SLAM [14] is a direct visual SLAM using the sparse depth of LiDAR with a narrow FOV. The window-based optimization is conducted by minimizing the photometric errors of the selected points in the image to estimate the pose [14]. However, the occlusion has not been considered in [14], which may destroy the constant image brightness assumption for the direct method and thus degrade the accuracy.

The direct method is also employed in some tightly coupled LiDAR-visual-inertial navigation systems. In R3LIVE [24], the direct method is employed in the VIO subsystem, and the LIO subsystem is based on the FAST_LIO2 [19]. Similarly, FAST-LIVO [25] is sparse-direct LiDAR-inertial-visual odometry, in which the VIO is based on SVO [31], and the LIO is adapted from FAST_LIO2 [19]. The LiDAR depth from the built global map is also used in R3LIVE and FAST-LIVO.

In these direct methods [13], [14], [24], [25], the depths from the LiDAR are used as constants. This may degrade the system accuracy because the measurement noise and the camera-LiDAR extrinsic error may introduce inaccurate depth measurement. Moreover, direct methods are susceptible to illumination changes and involve precise photometric calibration [32]. Hence, direct methods may degrade accuracy in complex environments with drastic illumination changes.

In this study, we aim at LiDAR-enhanced visual navigation to fully utilize the accurate and long-distance LiDAR depth and thus improve the robustness and accuracy in large-scale challenging environments. As there are typically rich visual textures in outdoor environments, the feature-based method is adopted in this study. Algorithm improvement is conducted in both the depth association and the use of the estimated depth. Specifically, a robust depth-association method with an effective plane-checking algorithm is proposed, which

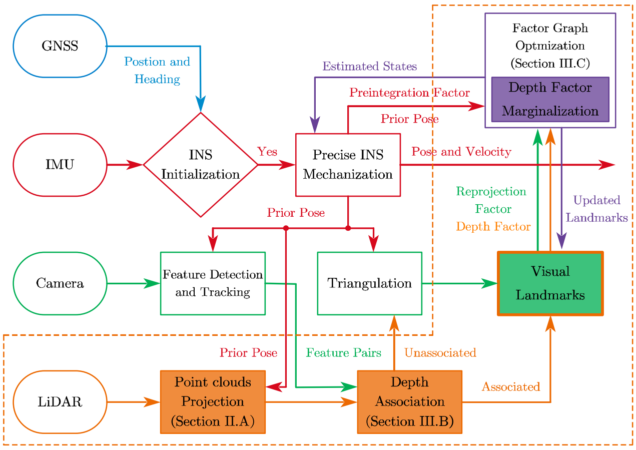

Fig. 2. System overview of LE-VINS. The parts within the orange area are the works in this study. The filled blocks denote the proposed methods in section IV.

significantly improves the accuracy of the estimated depth and avoids outliers. The estimated depth is not only adopted as the initial depth of the landmark but also employed to construct a LiDAR depth factor in the FGO, which can provide a direct and effective constraint for visual landmarks.

III. System Overview

The proposed LE-VINS is built upon our previous work IC-GVINS [33] by incorporating the solid-state LiDAR to provide accurate and long-distance depth for visual landmarks. The system framework of LE-VINS is depicted in Fig. 2. Here, the solid-state LiDAR with the non-repetitive pattern is employed because dense point clouds can be obtained by accumulating several LiDAR scans for low-speed robots, which is conducive for depth association. In addition, solid-state LiDAR, such as Livox Mid-70 and AVIA, can share large overlapping areas with the camera, which can associate more visual landmarks with accurate LiDAR depths.

Once the INS is initialized by the GNSS positioning, the INS mechanization is employed to provide an accurate prior pose for the visual subsystem and LiDAR subsystem. The visual subsystem is directly initialized with the INS pose, and the detected feature points (Shi-Tomasi) are tracked from frame to frame using the Lukas-Kanade optical flow algorithm. For the LiDAR subsystem, the point clouds from the solid-state LiDAR are projected to the visual keyframe using the INS pose and the LiDAR-camera extrinsic parameters.

Then, the tracked feature pairs in the keyframe are associated with the depth from the projected point clouds. If the feature pairs are associated, new visual landmarks will be added to the visual landmark map. Unassociated feature pairs will be triangulated to obtain the initial depth and will be further optimized by the FGO. For those landmarks with the LiDAR depth, LiDAR depth factors are constructed in the FGO to construct LiDAR depth measurements for visual landmarks.

Finally, the visual reprojection factors, the LiDAR depth factors, the IMU preintegration factors, and the prior factor from the marginalization are tightly fused within the FGO framework to achieve maximum-a-posterior (MAP) estimation. The estimated states will be used to update the INS mechanization and the newest INS states.

TABLE I
Notations and Symbols

Notations	Explanation
	Expressions
$\mathbf{q}, \mathbf{R}, \boldsymbol{\phi}$	The attitude quaternion, rotation matrix, and rotation vector
$\otimes$	The quaternion product
Log,Exp	The transformation between the quaternion and rotation vector
$\boldsymbol{p}$	A three-dimension position
$\boldsymbol{n}$	The normal vector of a plane
$\boldsymbol{\Sigma}$	The covariance matrix
d, δ	The depth and inverse-depth parameter of a visual landmark
	Variables
$\boldsymbol{p}_{\mathrm{wb}}^{\mathrm{w}}, \mathbf{q}_{\mathrm{b}}^{\mathrm{w}}$	The IMU pose w.r.t the world frame
$\boldsymbol{v}_{\mathrm{wb}}^{\mathrm{w}}$	The IMU velocity in the world frame
$\boldsymbol{b}_g, \boldsymbol{b}_a$	The gyroscope and accelerometer biases
$\boldsymbol{p}_{\mathrm{cl}}^{\mathrm{c}}, \mathbf{q}_{\mathrm{l}}^{\mathrm{c}}$	The LiDAR-camera extrinsic parameters
$\boldsymbol{p}_{\mathrm{bc}}^{\mathrm{b}}, \mathbf{q}_{\mathrm{c}}^{\mathrm{b}}$	The camera-IMU extrinsic parameters

IV. Solid-State-LiDAR-Enhanced Visual-Inertial Navigation System

This section presents the proposed solid-state-LiDAR-enhanced visual-inertial navigation system, as depicted in Fig. 2. The point clouds from the solid-state LiDAR are first projected and accumulated for depth association. A robust depth-association method is proposed to estimate the landmark depth, with an effective plane-checking algorithm for outlier culling. We propose a LiDAR depth factor with the estimated depth in the FGO to constrain the landmark depth. Finally, the visual feature, LiDAR depth, and IMU measurements are tightly fused using the FGO to achieve MAP estimation. The main notations involved in this section are shown in Table I.

A. Point clouds Projection

Dense point clouds can be obtained from the solid-state LiDAR due to its non-repetitive pattern, especially for low-speed robots, as depicted in Fig. 1. However, each LiDAR point is sampled at a different time for such solid-state LiDAR, which may cause motion distortion, degrading the accuracy of the depth association. In addition, the point clouds need to be accumulated to associate the visual landmark with the LiDAR depth. Hence, the sequentially sampled point clouds should be projected to the camera frame corresponding to the visual keyframe.

The LiDAR-camera extrinsic parameters should be obtained first to project the point clouds to the camera frame. By using the non-repetitive pattern of the solid-state LiDAR, the LiDAR-camera extrinsic parameters can be estimated precisely. Specifically, the LiDAR point clouds and images are sampled simultaneously during the stationary state for several seconds. Then the 3D points in the accumulated point clouds and the corresponding 2D pixels in the image can be obtained

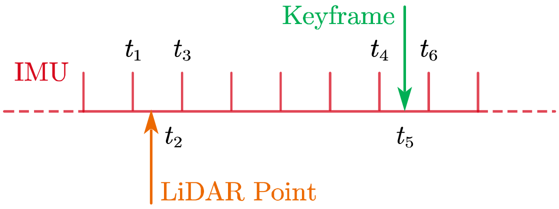

Fig. 3. The sample time of the IMU, LiDAR point, and visual keyframe.

simultaneously. A nonlinear optimization can be conducted by minimizing the reprojection error to estimate the extrinsic parameters. The estimated extrinsic parameters can be represented as $\{\boldsymbol{p}^{\mathrm{c}}_{\mathrm{cl}}, \mathbf{q}^{\mathrm{c}}_{\mathrm{l}}\}$, where c and l denote the camera frame (c-frame) and LiDAR frame (l-frame), respectively.

As depicted in Fig. 3, with the continuous INS pose and the extrinsic parameters, the LiDAR point sampled at t_2 can be projected into the camera frame at t_5. The INS pose between the two IMU samples can be obtained by linear interpolation, and the position at t_2 can be written as

$$s = (t_2 - t_1) / (t_3 - t_1), \tag{1}$$

$$\boldsymbol{p}^{\mathrm{w}}_{\mathrm{wb},t_2} \approx \boldsymbol{p}^{\mathrm{w}}_{\mathrm{wb},t_1} + s(\boldsymbol{p}^{\mathrm{w}}_{\mathrm{wb},t_3} - \boldsymbol{p}^{\mathrm{w}}_{\mathrm{wb},t_1}) \tag{2}$$

where s is a scale coefficient; w denotes the world frame (w-frame), which is defined at the initial position of the navigation frame (n-frame), i.e. the local geodetic north-east-down (NED) frame; b denotes the IMU body frame (b-frame). As for the attitude, we can interpolate the rotation vector as

$$\boldsymbol{\phi}^{\mathrm{b},t_2}_{\mathrm{b},t_1} \approx s\boldsymbol{\phi}^{\mathrm{b},t_3}_{\mathrm{b},t_1} \approx s\mathrm{Log}((\mathbf{q}^{\mathrm{w}}_{\mathrm{b},t_3})^{-1} \otimes \mathbf{q}^{\mathrm{w}}_{\mathrm{b},t_1}), \tag{3}$$

$$\mathbf{q}^{\mathrm{w}}_{\mathrm{b},t_2} = \mathbf{q}^{\mathrm{w}}_{\mathrm{b},t_1} \otimes (\mathrm{Exp}(\boldsymbol{\phi}^{\mathrm{b},t_2}_{\mathrm{b},t_1}))^{-1}, \tag{4}$$

where $\boldsymbol{\phi}$ denotes the rotation vector, as can be seen in [34]. Hence, we obtain the interpolated INS pose $\boldsymbol{p}^{\mathrm{w}}_{\mathrm{wb},t_2}, \mathbf{q}^{\mathrm{w}}_{\mathrm{b},t_2}\}$ at t_2 from (2) and (4). The INS pose can be converted to the camera pose by using the camera-IMU extrinsic parameters $\{\boldsymbol{p}^{\mathrm{b}}_{\mathrm{bc}}, \mathbf{q}^{\mathrm{b}}_{\mathrm{c}}\}$, as follow

$$\begin{aligned} \boldsymbol{p}^{\mathrm{w}}_{\mathrm{wc},t_2} &= \boldsymbol{p}^{\mathrm{w}}_{\mathrm{wb},t_2} + \mathbf{R}^{\mathrm{w}}_{\mathrm{b},t_2}\boldsymbol{p}^{\mathrm{b}}_{\mathrm{bc}}, \\ \mathbf{q}^{\mathrm{w}}_{\mathrm{c},t_2} &= \mathbf{q}^{\mathrm{w}}_{\mathrm{b},t_2} \otimes \mathbf{q}^{\mathrm{b}}_{\mathrm{c}}. \end{aligned} \tag{5}$$

The same processes can be conducted to obtain the camera pose $\boldsymbol{p}^{\mathrm{w}}_{\mathrm{wc},t_5}, \mathbf{q}^{\mathrm{w}}_{\mathrm{c},t_5}\}$ at t_5. Hence, for a LiDAR point m at t_2, denoted as $\boldsymbol{p}^{\mathrm{l}}_{\mathrm{lm},t_2}$, it can be transformed to the c-frame as

$$\boldsymbol{p}^{\mathrm{c}}_{\mathrm{cm},t_2} = \mathbf{R}^{\mathrm{c}}_{\mathrm{l}}\boldsymbol{p}^{\mathrm{l}}_{\mathrm{lm},t_2} + \boldsymbol{p}^{\mathrm{c}}_{\mathrm{cl}}. \tag{6}$$

With the camera pose at t_2 and t_5, the LiDAR point at t_2 can be transformed to the c-frame at t_5 as

$$\boldsymbol{p}^{\mathrm{c}}_{\mathrm{cm},t_5} = (\mathbf{R}^{\mathrm{w}}_{\mathrm{c},t_5})^T \mathbf{R}^{\mathrm{w}}_{\mathrm{c},t_2}\boldsymbol{p}^{\mathrm{c}}_{\mathrm{cm},t_2} + (\mathbf{R}^{\mathrm{w}}_{\mathrm{c},t_5})^T(\boldsymbol{p}^{\mathrm{w}}_{\mathrm{wc},t_2} - \boldsymbol{p}^{\mathrm{w}}_{\mathrm{wc},t_5}). \tag{7}$$

Finally, the LiDAR point is projected to the c-frame corresponding to the visual keyframe without motion distortion.

The sequentially sampled point clouds can be projected to the same c-frame to obtain relatively dense point clouds for depth association. However, more points consume more computation resources. Hence, the accumulation time should be limited to bound computational complexity. For solid-state LiDAR Livox Mid-70, it can measure 100,000 points in one second. For a low-speed robot with a speed of about 1.5 m/s, the projected point clouds in the image plane are depicted in Fig. 4. When the accumulation time is 0.5 seconds, the point clouds are dense enough for depth association. In contrast, the point clouds are sparse for the accumulation time of 0.25 seconds, and the point clouds are much denser for the accumulation time of 1 second. According to our experiments, the accuracy will not significantly improve if the accumulation time is longer than 0.5 seconds. Consequently, the accumulation time is set to 0.5 seconds to balance the accuracy and computational complexity. The projected point clouds will be further processed for depth association.

B. Robust Depth Association

We can retrieve depth for visual landmarks with the projected point clouds. However, the visual feature points are in the 2D image plane, while the point clouds are in the 3D space. To achieve depth association, we can project the point clouds into the image plane, like [10], or project the visual feature points into the c-frame, like [1]. Compared to a large number of point clouds, the number of visual features is usually within several hundred. Hence, we project the visual features into the c-frame for depth association, which can significantly reduce computational complexity. The projected visual features in the c-frame have no depth, and thus the association is conducted on a unit sphere, whose center is at the origin of the c-frame, as shown in Fig. 5.

The projected point clouds are normalized to be converted to the unit sphere. Besides, they are downsampled simultaneously for a constant-angle density on the sphere. In addition, only the points in the foreground will be reserved during the

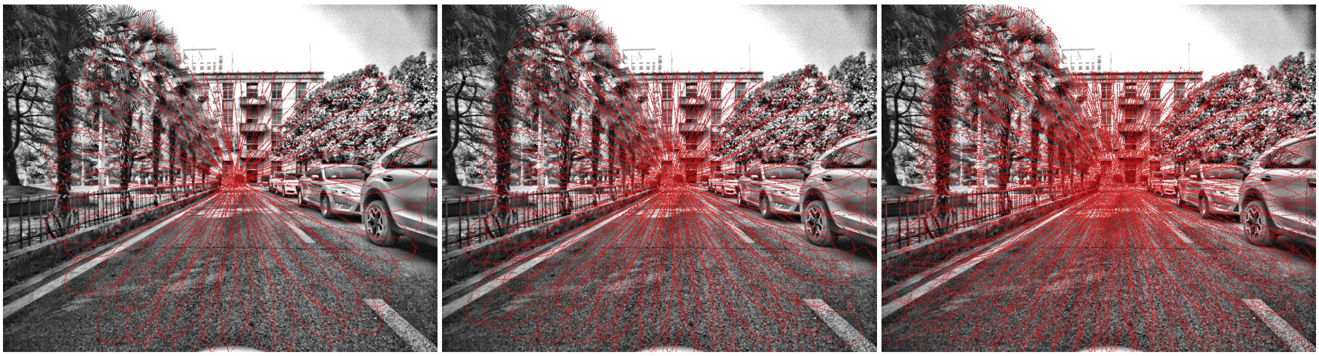

Fig. 4. The projected point clouds in the image plane (the robot is in dynamic condition with a speed of about 1.5 m/s). Each point is expressed as a red cross for better visualization. The accumulation time for the three images are 0.25 second (left), 0.5 second (middle), and 1.0 second (right), respectively.

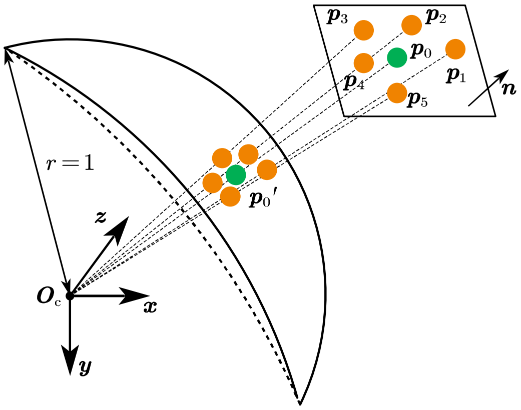

Fig. 5. An illustration of the depth association for visual feature. The orange points denote the LiDAR points, while the green points denote the visual feature point.

downsampling; thus, some occlusion points can be removed. The higher angle density denotes more computation, while the lower angle density denotes lower accuracy. The horizontal FOV of the used camera is about 80°. The FOV of the used solid-state LiDAR Mid-70 is about 70° (circular), but the FOV of the projected point clouds can be larger due to movement. Hence, the FOV of the projected point clouds is set to 90° for downsampling to ensure that more visual features can be associated. The noise of the feature tracking is usually within 1.5 pixels, which means the tracked feature can be expressed within a square with a width of 3 pixels. Hence, the angle density for the downsampling can be determined by considering the camera resolution of 1280×1024 as

$$\begin{aligned} \Delta a_h &= 90^\circ \div (1280 \div 3) \approx 0.21^\circ, \\ \Delta a_v &= 90^\circ \div (1024 \div 3) \approx 0.26^\circ, \end{aligned} \tag{8}$$

where h and v denote the horizontal and vertical direction, respectively. According to the results in (8), the angle density Δa is set to 0.2° to improve the depth-association accuracy. Hence, the downsampled point clouds are located on the unit sphere with the resolution of 450×450. In the meantime, the visual feature points are also projected to the unit sphere using the camera projection function.

A KD tree is constructed using the down-sampled point clouds on the unit sphere to find the corresponding points and retrieve depth for visual features. Typically, the plane fitting is adopted to estimate the depth for visual feature, using three selected LiDAR points, such as in [1], [10], [15]. However, we cannot ensure that the visual landmark is lying on a real plane using only three points; thus, inaccurate or wrong depth estimation may occur. Hence, a plane-checking algorithm can be employed to verify the estimated depth to avoid and decrease wrong associations. As depicted in Fig. 5, we find the nearest five points of the visual feature point by searching in the KD-tree. The found five points are employed to fit a local plane around the visual landmark. Suppose the furthest found point around the visual feature point on the unit sphere is not within $\pm 3\Delta a$; the association will not be continued to avoid introducing possible outliers. Specifically, the found five points can be approximately expressed within a circle with the radium of $r \approx 3 \times 3 = 9$ pixels in the image plane, according to the formulation in (8). The expected plane will not pass through the origin $\boldsymbol{O}_c$ in this depth-association problem. Hence, the LiDAR point $\boldsymbol{p}$ in a plane can be expressed as

$$\boldsymbol{p}^T \boldsymbol{n} + 1 = 0, \tag{9}$$

where $\boldsymbol{n}$ is the normal vector of the plane. Hence, an overdetermined linear equation can be constructed to solve this plane as

$$\begin{aligned} &\mathbf{A}\boldsymbol{n} = \boldsymbol{b}, \\ &\mathbf{A} = \begin{bmatrix} \boldsymbol{p}_1 & \boldsymbol{p}_2 & \boldsymbol{p}_3 & \boldsymbol{p}_4 & \boldsymbol{p}_5 \end{bmatrix}^T, \\ &\boldsymbol{b} = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 \end{bmatrix}^T, \end{aligned} \tag{10}$$

where $\boldsymbol{p}_e, e \in [1,5]$ are the found LiDAR points in the c-frame with 3D coordinates, as depicted in Fig. 5. The linear equation (10) can be solved by a method like QR decomposition, and thus the normal vector $\boldsymbol{n}$ can be obtained.

A plane-checking algorithm is employed to avoid wrong associations and ensure that the visual landmark lies on a real plane. More specifically, the plane checking is conducted by calculating the point-to-plane distance as

$$dis = \frac{\left| \boldsymbol{p}_e{}^T \boldsymbol{n} + 1 \right|}{\left\| \boldsymbol{n} \right\|}, e \in [1,5]. \tag{11}$$

If $dis < 0.1m$ for all the five LiDAR points, the fitted plane will be used to estimate the depth of the visual landmark; otherwise, the depth association will be failed. Here, the distance threshold of 0.1 m is set according to the point-to-plane metric in the LiDAR-inertial navigation system (LINS), such as in [17], [35]. The occlusion points are not explicitly processed in our method, but they can be easily detected by the plane-checking algorithm and will not be employed for depth association. The plane checking can significantly avoid wrong depth associations and thus improve the system robustness and accuracy.

With the normal vector $\boldsymbol{n}$, the landmark depth can be retrieved. We find the LiDAR point furthest to the visual feature on the unit sphere, i.e. the $\boldsymbol{p}_1$ in Fig. 5. With the visual feature point $\boldsymbol{p}_0'$ on the unit sphere, we want to retrieve its 3D coordinate $\boldsymbol{p}_0$ in the c-frame as

$$\boldsymbol{p}_0 = t\boldsymbol{p}_0', \tag{12}$$

where t is the distance to the c-frame center $\boldsymbol{O}_c$. Using the plane equation (9), t can be solved as follows

$$(\boldsymbol{p}_0 - \boldsymbol{p}_1)^T \boldsymbol{n} = (t\boldsymbol{p}_0' - \boldsymbol{p}_1)^T \boldsymbol{n} = 0, \tag{13}$$

$$t = \frac{\boldsymbol{p}_1{}^T \boldsymbol{n}}{(\boldsymbol{p}_0')^T \boldsymbol{n}}. \tag{14}$$

Finally, the 3D coordinate of the visual landmark in c-frame, denoted as $\boldsymbol{p}_0(p_{0,x}, p_{0,y}, p_{0,z})$, can be obtained from (12). The landmark depth in this visual keyframe can be expressed as

$$d_0 = p_{0,z}. \tag{15}$$

The estimated depth from the depth association can be employed to directly constrain the inverse-depth parameter of the visual landmark [26], which will be presented in section IV.C.3.

A robust depth-association method is presented in this section, which can estimate accurate depth for visual landmarks. More specifically, the landmark depth is obtained by fitting a local plane with an effective plane-checking algorithm rather than the rough processes as in [1], [15]. For those depth-unassociated visual features, triangulation will be conducted to retrieve the initial depth, and the depth will be further estimated in the FGO. Results will be presented to demonstrate the superior robustness and accuracy of the proposed depth-association method in section V. The estimated landmark depth from the LiDAR will be treated as an initial value and employed to construct the LiDAR depth factor in the FGO to constrain the landmark depth.

C. Factor Graph Optimization

The associated depth from LiDAR can be incorporated into the FGO to improve the state estimation accuracy. Specifically, when a new visual keyframe is selected, a new time node is created in the sliding-window optimizer. In the meantime, the LiDAR points are projected to the visual keyframe for depth association. For those visual features with LiDAR depths, new landmarks are created and added to the landmarks map, while other landmarks are created by triangulation. Hence, the LiDAR depth factors can be constructed in the FGO to constrain the inverse-depth parameter of the landmarks directly. The IMU preintegration is employed to provide relative constraints between each consecutive time node. Finally, the visual reprojection factors, the LiDAR depth factors, the IMU preintegration factors, and the prior factors are tightly fused under the framework of FGO to achieve MAP estimation. Factor graph optimization is equivalent to nonlinear optimization and is conducted by solving a nonlinear least square problem. The FGO framework of LE-VINS is depicted in Fig. 6.

1) Formulation

The state vector $\boldsymbol{X}$ in the sliding-window optimizer of LE-VINS can be defined as follow:

$$
\begin{aligned}
\boldsymbol{X} &= \left[\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_{\mathrm{c}}^{\mathrm{b}}, \delta_1, \delta_2, \ldots, \delta_s\right], \\
\boldsymbol{x}_k &= \left[\boldsymbol{p}_{\mathrm{wb}_k}^{\mathrm{w}}, \mathbf{q}_{\mathrm{b}_k}^{\mathrm{w}}, \boldsymbol{v}_{\mathrm{wb}_k}^{\mathrm{w}}, \boldsymbol{b}_{g_k}, \boldsymbol{b}_{a_k}\right], k \in [0, n], \\
\boldsymbol{x}_{\mathrm{c}}^{\mathrm{b}} &= \left[\boldsymbol{p}_{\mathrm{bc}}^{\mathrm{b}}, \mathbf{q}_{\mathrm{c}}^{\mathrm{b}}\right],
\end{aligned}
\tag{16}
$$

where $\boldsymbol{x}_k$ is the IMU state at each time node, including the position, the attitude quaternion, and the velocity in the w-frame, and the gyroscope biases $\boldsymbol{b}_g$ and the accelerometer biases $\boldsymbol{b}_a$; δ is the inverse-depth parameter of the visual landmark in the reference keyframe, i.e. the first observed keyframe for triangulated landmarks, or the keyframe associated with the LiDAR depth; n is the number of the IMU preintegration in the sliding window; $\boldsymbol{x}_{\mathrm{c}}^{\mathrm{b}}$ is the camera-IMU extrinsic parameters.

The following nonlinear optimization problem can be solved by minimizing the sum of the Mahalanobis norm of all measurements and the prior as

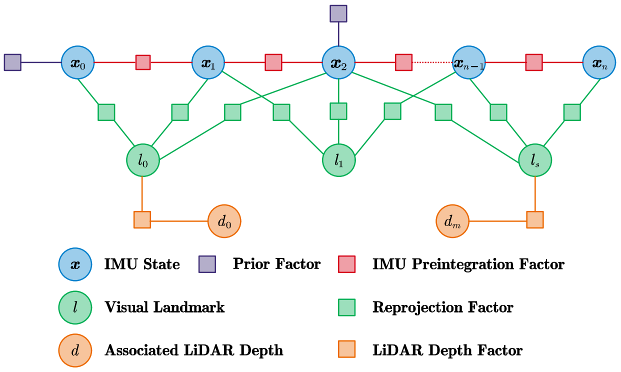

Fig. 6. The FGO framework of the LE-VINS.

$$
\min_{\boldsymbol{X}} \left\{
\begin{aligned}
&\left\|\mathbf{r}_p - \mathbf{H}_p \boldsymbol{X}\right\|^2 + \sum_{k \in [1,n]} \left\|\mathbf{r}_{Pre}\left(\tilde{\boldsymbol{z}}_{k-1,k}^{Pre}, \boldsymbol{X}\right)\right\|_{\boldsymbol{\Sigma}_{k-1,k}^{Pre}}^2 \\
&+\sum_{l \in L} \left\|\mathbf{r}_V\left(\tilde{\boldsymbol{z}}_l^{V_{i,j}}, \boldsymbol{X}\right)\right\|_{\boldsymbol{\Sigma}_l^{V_{i,j}}}^2 + \sum_{h \in [0,m]} \left\|\mathbf{r}_D\left(\tilde{\boldsymbol{z}}_h^D, \boldsymbol{X}\right)\right\|_{\boldsymbol{\Sigma}_h^D}^2
\end{aligned}
\right\}, \tag{17}
$$

where $\mathbf{r}_V$ are the residuals for the visual measurements; L is the landmark map in the sliding window, and l is the landmark in the map; i denotes the reference keyframe of the landmark l, and j is another observed keyframe; $\mathbf{r}_D$ are the residuals for the LiDAR depth measurements, which directly constrain the inverse-depth parameters of the visual landmarks; $\mathbf{r}_{Pre}$ are the residuals for the IMU preintegration measurements; $\{\mathbf{r}_p, \mathbf{H}_p\}$ denotes the prior information from the marginalization. The Ceres solver [36], an open-sourced library for modelling and solving large optimization problems, is adopted in LE-VINS. Specifically, the Levenberg-Marquardt algorithm [36] is employed to solve the nonlinear least squares problem in (17).

2) Visual Reprojection Factor

The visual reprojection residual is defined on a unit sphere. For the landmark l with its inverse-depth parameter δ_l in its reference frame i and another observed keyframe j, the visual reprojection residuals can be defined as

$$
\begin{aligned}
\mathbf{r}_V\left(\tilde{\boldsymbol{z}}_l^{V_{i,j}}, \boldsymbol{X}\right) = \left[\boldsymbol{b}_1 \quad \boldsymbol{b}_2\right]^T \cdot \left(\frac{\hat{\boldsymbol{p}}_{\mathrm{c}_j}}{\left\|\hat{\boldsymbol{p}}_{\mathrm{c}_j}\right\|} - \pi_{\mathrm{c}}^{-1}\left(\tilde{\boldsymbol{p}}_{\mathrm{p}_j}\right)\right)&, \\
\hat{\boldsymbol{p}}_{\mathrm{c}_j} = (\mathbf{R}_{\mathrm{c}}^{\mathrm{b}})^T \left(\hat{\boldsymbol{p}}_{\mathrm{b}_j} - \boldsymbol{p}_{\mathrm{bc}}^{\mathrm{b}}\right)&, \\
\hat{\boldsymbol{p}}_{\mathrm{b}_j} = (\mathbf{R}_{\mathrm{b}_j}^{\mathrm{w}})^T \left(\mathbf{R}_{\mathrm{b}_i}^{\mathrm{w}} \hat{\boldsymbol{p}}_{\mathrm{b}_i} + \boldsymbol{p}_{\mathrm{wb}_i}^{\mathrm{w}} - \boldsymbol{p}_{\mathrm{wb}_j}^{\mathrm{w}}\right)&, \\
\hat{\boldsymbol{p}}_{\mathrm{b}_i} = \mathbf{R}_{\mathrm{c}}^{\mathrm{b}} \frac{1}{\delta_l} \pi_{\mathrm{c}}^{-1}\left(\tilde{\boldsymbol{p}}_{\mathrm{p}_i}\right) + \boldsymbol{p}_{\mathrm{bc}}^{\mathrm{b}}&,
\end{aligned}
\tag{18}
$$

where $\tilde{\boldsymbol{p}}_{\mathrm{p}_i}$ and $\tilde{\boldsymbol{p}}_{\mathrm{p}_j}$ are the observed visual features in the pixel plane; $\hat{\boldsymbol{p}}_{\mathrm{b}_i}$ and $\hat{\boldsymbol{p}}_{\mathrm{b}_j}$ are the coordinates of the landmark l in the b-frame corresponding to the keyframes i and j; $\hat{\boldsymbol{p}}_{\mathrm{c}_j}$ is the calculated coordinate of the landmark l in the c-frame of the keyframe j; $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are two orthogonal bases that span the tangent plane of $\hat{\boldsymbol{p}}_{\mathrm{c}_j}$; π_{c}^{-1} is the back camera projection function that transforms the visual feature in pixel plane $\tilde{\boldsymbol{p}}_{\mathrm{p}}$ to a unit vector using the camera intrinsic parameters; $\boldsymbol{p}_{\mathrm{bc}}^{\mathrm{b}}$ and $\mathbf{R}_{\mathrm{c}}^{\mathrm{b}}$ are the extrinsic parameters in (16); $\boldsymbol{p}_{\mathrm{wb}}^{\mathrm{w}}$ and $\mathbf{R}_{\mathrm{b}}^{\mathrm{w}}$ represent the pose of the IMU in the w-frame, as in (16). The covariance $\boldsymbol{\Sigma}_{l}^{V_{i,j}}$ in (17) are also propagated from the pixel plane (the standard deviation of 1.5 pixels) onto the unit sphere.

3) LiDAR Depth Factor

With the depth-association algorithm in section IV.B, accurate depth estimation for visual landmarks can be obtained. However, the estimated depth also contains noise, mainly because the depth-association algorithm may introduce error. Hence, it is improper to set the landmark depth as a constant without being optimized in the FGO, such as in LVI-SAM [15]. To fully utilize the accurate depth from the LiDAR, we propose a LiDAR depth factor, which can directly constrain the landmark depth while considering the depth-association noise. Specifically, the LiDAR depth factor utilizes the estimated LiDAR depth to impose a constraint on the inverse-depth parameter of the landmark, and the residual can be written as

$$\mathbf{r}_{D}\left(\tilde{\boldsymbol{z}}_{h}^{D}, \boldsymbol{X}\right) = d_{h} - \frac{1}{\delta_{h}}, \tag{19}$$

where δ_{h} is the inverse-depth parameter of the landmark l_{h} as in (16), and d_{h} is the associated depth for the landmark l_{h} from (15). The standard deviation for the covariance $\boldsymbol{\Sigma}_{h}^{D}$ is set to 0.1 m according to the distance threshold for the plane-checking algorithm (11). In conclusion, the accurate depth from the LiDAR is employed to construct the LiDAR depth factor, and the depth-measurement noise is also considered. Moreover, the LiDAR depth factor can be converted to the prior factor if the landmark is marginalized, and thus the constraint can be reserved.

4) IMU Preintegration Factor

We follow our refined IMU preintegration [37] that compensates for the Earth rotation to improve the accuracy of industrial-grade MEMS IMU. The residuals of the employed IMU preintegration factor can be expressed as

$$\mathbf{r}_{Pre}\left(\tilde{\mathbf{z}}_{k-1,k}^{Pre}, \boldsymbol{X}\right) = \begin{bmatrix} \Delta \boldsymbol{p}_{k-1,k}^{Pre} - \Delta \hat{\boldsymbol{p}}_{k-1,k}^{Pre} \\ \Delta \boldsymbol{v}_{k-1,k}^{Pre} - \Delta \hat{\boldsymbol{v}}_{k-1,k}^{Pre} \\ 2\left[\left(\mathbf{q}_{\mathrm{b}_k}^{\mathrm{w}}\right)^{-1} \otimes \mathbf{q}_{\mathrm{w}_{\mathrm{i}(k-1)}}^{\mathrm{w}}\left(t_k\right) \otimes \mathbf{q}_{\mathrm{b}_{k-1}}^{\mathrm{w}} \otimes \hat{\mathbf{q}}_{k-1,k}^{Pre}\right]_{\boldsymbol{v}} \\ \boldsymbol{b}_{g_k} - \boldsymbol{b}_{g_{k-1}} \\ \boldsymbol{b}_{a_k} - \boldsymbol{b}_{a_{k-1}} \end{bmatrix}, \tag{20}$$

where $\Delta \boldsymbol{p}_{g/cor,k-1,k}^{\mathrm{w}}$ and $\Delta \boldsymbol{v}_{g/cor,k-1,k}^{\mathrm{w}}$ are the Coriolis correction term [37]; $\Delta \hat{\boldsymbol{p}}_{k-1,k}^{Pre}$, $\Delta \hat{\boldsymbol{v}}_{k-1,k}^{Pre}$, and $\hat{\mathbf{q}}_{k-1,k}^{Pre}$ are the position, velocity and attitude preintegration measurements, respectively; $\boldsymbol{g}^{\mathrm{w}}$ are the gravity in the w-frame; $\mathbf{q}_{\mathrm{w}_{\mathrm{i}(k-1)}}^{\mathrm{w}}\left(t_k\right)$ is caused by the Earth rotation. The gyroscope biases $\boldsymbol{b}_{g}$ and accelerometer biases $\boldsymbol{b}_{a}$ in (16) are also included in the residuals for online estimation and correction. The calculated position preintegration $\Delta \boldsymbol{p}_{k-1,k}^{Pre}$ and velocity preintegration $\Delta \boldsymbol{v}_{k-1,k}^{Pre}$ can be expressed as follows:

$$\begin{aligned} \Delta \boldsymbol{p}_{k-1,k}^{Pre} &= \left(\mathbf{R}_{\mathrm{b}_{k-1}}^{\mathrm{w}}\right)^{T}\left(\boldsymbol{p}_{\mathrm{wb}_k}^{\mathrm{w}} - \boldsymbol{p}_{\mathrm{wb}_{k-1}}^{\mathrm{w}} - \boldsymbol{v}_{\mathrm{wb}_{k-1}}^{\mathrm{w}} \Delta t_{k-1,k}\right. \\ &\quad \left. -0.5\boldsymbol{g}^{\mathrm{w}} \Delta t_{k-1,k}^{2} + \Delta \boldsymbol{p}_{g/cor,k-1,k}^{\mathrm{w}}\right), \\ \Delta \boldsymbol{v}_{k-1,k}^{Pre} &= \left(\mathbf{R}_{\mathrm{b}_{k-1}}^{\mathrm{w}}\right)^{T}\left(\boldsymbol{v}_{\mathrm{wb}_k}^{\mathrm{w}} - \boldsymbol{v}_{\mathrm{wb}_{k-1}}^{\mathrm{w}} - \boldsymbol{g}^{\mathrm{w}} \Delta t_{k-1,k}\right. \\ &\quad \left. +\Delta \boldsymbol{v}_{g/cor,k-1,k}^{\mathrm{w}}\right). \end{aligned} \tag{21}$$

The covariance $\boldsymbol{\Sigma}_{k-1,k}^{Pre}$ of the IMU preintegration factor is derived from the noise propagation [37].

5) Marginalization

For real-time positioning, only several visual keyframes can be reserved in the optimization window to bound the computational complexity. Hence, when a new keyframe is added to the window, an old keyframe with its landmarks will be removed. However, suppose the landmarks are removed from the window directly. In that case, the accurate LiDAR depth cannot be transformed or reserved in most existing methods, such as DEMO [1] and LIMO [10], which results in a loss of information. Hence, marginalization [32], [38] is adopted to convert all the measurements corresponding to the removed state into a prior. Specifically, the marginalization is conducted using the Schur complement operation [38], and the prior is constructed based on all marginalized measurements corresponding to the removed state. Without constructing the LiDAR depth factor, the depth information in LVI-SAM [15] cannot be converted into the prior. In contrast, with the employed LiDAR depth factor in the proposed LE-VINS, the LiDAR depth measurements can also be converted into the prior during the marginalization. Hence, the accurate depth constraints can be reserved, which can improve the system accuracy.

V. Experiments and Results

This section presents the experiments and results to evaluate the robustness and accuracy of the proposed LE-VINS. The equipment setup of the employed robot and the configurations of the experiments are described first. Then, the field tests were conducted in various environments for quantitative evaluation to examine the accuracy and robustness of LE-VINS. Finally, the running time of LE-VINS is presented.

A. Equipment Setup and Configurations

The proposed LE-VINS is implemented using C++ and the robot operating system (ROS). A low-speed wheeled robot, with an average speed of around 1.5 m/s, is employed for the field tests, as depicted in Fig. 7. The sensors used in this study include a camera with a resolution of 1280x1024 and a frame rate of 20 Hz (Allied Vision Mako-G131), a solid-state LiDAR with the frame rate of 10 Hz (Livox Mid-70), an industrial-grade MEMS IMU (ADI ADIS16465 with the gyroscope bias instability of 2 °/hr and a frame rate of 200 Hz), and a dual-antenna GNSS receiver with the frame rate of 1 Hz (NovAtel OEM-718D). The GNSS real-time kinematic (RTK) is adopted to achieve high-accuracy positioning [39]. These sensors are all synchronized through hardware triggers to the GNSS time. An

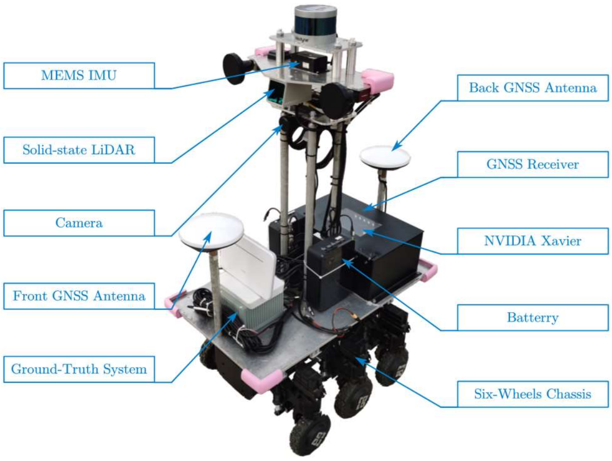

Fig. 7. Equipment setup of the robot.

onboard ARM computer (NVIDIA Xavier and 32GB RAM) is adopted for data acquisition and is utilized to achieve real-time navigation. The ground-truth system is a high-accuracy GNSS/INS integrated navigation system using the GNSS-RTK and a navigation-grade IMU. The ground truth (0.02 m for position and 0.01 deg for attitude) is generated by a post-processing software. Here, the GNSS positioning (5 seconds) is only used for initialization, as shown in Fig. 2.

The intrinsic parameters of the camera and the camera-IMU extrinsic parameters $\{\boldsymbol{p}^{\mathrm{b}}_{\mathrm{bc}}, \mathbf{q}^{\mathrm{b}}_{\mathrm{c}}\}$ are all calibrated offline using the Kalibr [40]. The camera-IMU extrinsic parameters are also estimated and compensated online in LE-VINS. The LiDAR-camera extrinsic parameters $\{\boldsymbol{p}^{\mathrm{c}}_{\mathrm{cl}}, \mathbf{q}^{\mathrm{c}}_{\mathrm{l}}\}$ are calibrated using the non-repetitive pattern of the solid-state LiDAR, as mentioned in section IV.A. The LiDAR-camera extrinsic parameters are not estimated online in LE-VINS.

We compared LE-VINS with its original VINS, IC-VINS, in [33] to evaluate the improvement by incorporating LiDAR depth. We also compared LE-VINS with the tightly-coupled LiDAR-visual-inertial navigation system R2LIVE [23]. Here, R2LIVE is adopted because it is feature-based and supports the solid-state LiDAR. However, LiDAR is not utilized to provide depth for visual landmarks in R2LIVE. As LVI-SAM [15] does not support the solid-state LiDAR, we implemented the LiDAR-enhanced method in LVI-SAM [15] to replace the proposed method, denoted as LE-VINS-LS ("LS" represents LVI-SAM). The difference between LE-VINS and LE-VINS-LS is only the LiDAR-enhanced method, including the depth association and the usage of the LiDAR depth. Specifically, LE-VINS-LS uses only three nearest points to associate depths with visual landmarks without using the plane-checking algorithm. Besides, the associated depths are set to constants during the optimization in LE-VINS-LS without using the proposed lidar-depth factor. As feature-based visual processes are employed, we use a max of 120 features for these systems.

The absolute and relative pose errors [41] were adopted for the quantitative evaluation. Specifically, the relative error over the sub-sequences of the length of 25m, 50m, 100m, and 200m are employed to evaluate the short-term and long-term accuracy. It should be noted that the results for IC-VINS, LE-VINS, and LE-VINS-LS are all determinant in each run. All the systems are run in real-time on a desktop PC (AMD R7-3700X and 32GB RAM) under the framework of ROS.

B. *Evaluation of the Accuracy*

To quantitatively evaluate the accuracy of the proposed LE-VINS, two field tests were conducted in large-scale challenging environments, all on the Wuhan University campus. The trajectory lengths in experiment-1 and experiment-2 are 2554 meters (1801 seconds) and 2533 meters (1778 seconds), respectively. There are various challenging scenes in the two experiments, including severe illumination changes, weak textures, and moving objects. The illumination changes are caused by the sunshine, mainly in experiment-1. The weak-texture scenes mostly happened in experiment-2, where there are some open-sky scenes. In addition, there are many moving objects in both experiments, including pedestrians, bicycles, and vehicles. These challenging scenes may significantly affect the robustness and accuracy of the visual navigation system.

1) Comparison of the Trajectory

The test trajectories in experiment-1 and experiment-2 are depicted in Fig. 8 and Fig. 9. The proposed LE-VINS is well aligned with the ground truth. There are no notable differences between LE-VINS and IC-VINS in terms of the trajectory in the two experiments. The better rotation accuracy denotes a more similar trajectory to the ground truth. Hence, it demonstrates that LE-VINS and IC-VINS yield almost similar long-term accuracy and rotation accuracy. For LE-VINS-LS, it occurs a large deviation in scene *S1* in experiment-1, as shown in Fig. 8. Besides, LE-VINS-LS exhibits a similar trajectory to LE-VINS in experiment-2, as depicted in Fig. 9. In contrast, R2LIVE deviates from the ground truth notably, showing worse long-term accuracy, especially in experiment-1. The LiDAR subsystem in R2LIVE plays a more critical role than the visual subsystem. However, the two experiments have rich visual textures with less structured scenes. The unstructured scenes are challenging for the LiDAR data association, especially for the solid-state LiDAR with a small horizontal FOV. Hence, R2LIVE exhibits a worse trajectory in experiment-1, as shown in Fig. 8. In addition, the LiDAR systems in R2LIVE are based on scan-to-map matching, and thus R2LIVE can match its previous trajectories with the help of the visual system in experiment-2, as shown in Fig. 9.

The absolute rotation error (ARE) and the absolute translation error (ATE) in the two experiments are exhibited in Table II. Compared to IC-VINS, LE-VINS demonstrates improved accuracy in the two experiments, and the improvement is more than 10% in absolute translation. In contrast, LE-VINS-LS degrades the accuracy in experiment-1, though it exhibits slightly higher translation accuracy than LE-VINS in experiment-2. As IC-VINS has already achieved superior accuracy, we cannot find notable differences between LE-VINS and IC-VINS in trajectory, as mentioned above. Besides, R2LIVE exhibits the worst absolute accuracy, especially for the rotation accuracy, which corresponds to the trajectory results.

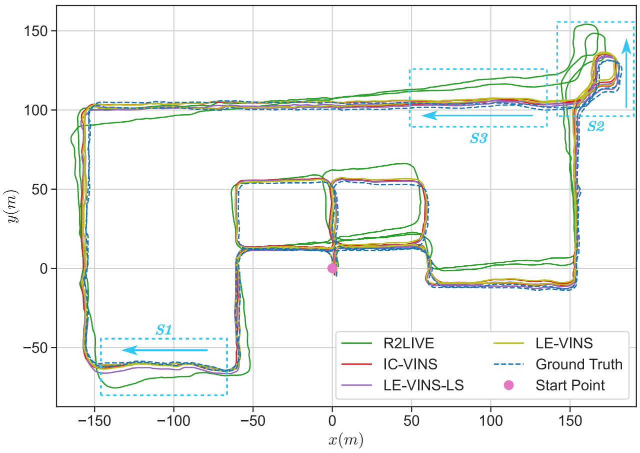

Fig. 8. The test trajectories in experiment-1. The trajectory length is 2554 meters. scenes *S1*, *S2*, and *S3* denote the degenerated scenes in section V.C.1.

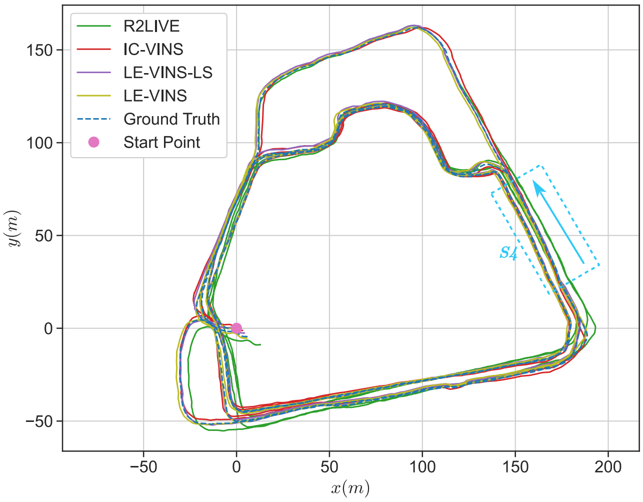

Fig. 9. The test trajectories in experiment-2. The trajectory length is 2533 meters. *S4* denotes the degenerated scene in section V.C.1.

2) *Comparison of the Relative Error*

The relative rotation error (RRE) and the relative translation error (RTE) are exhibited in Table III. Compared to IC-VINS, the proposed LE-VINS indicates a significant improvement in terms of short-term accuracy, i.e. the sub-sequences of the length of 25 m and 50 m. Specifically, the RTEs over 25 m decrease by more than 30% for LE-VINS in the two experiments. Such improvements for LE-VINS benefit from the robust depth-association method, which significantly avoids wrong associations. Besides, the proposed LiDAR depth factor in the FGO can fully utilize the accurate depth information. As for the long-term accuracy, i.e. the sub-sequences of the length of 100 m and 200 m, the improvements for LE-VINS are not significant, as IC-VINS has already achieved comparable long-term accuracy. In contrast, LE-VINS-LS exhibits few improvements compared to IC-VINS and even degrades accuracy in experiment-1. It demonstrates that the existing LiDAR-enhanced method is not sufficiently robust and is susceptible to environments.

Table III shows that R2LIVE achieves satisfied short-term accuracy and exhibits slightly higher translation accuracy than IC-VINS and LE-VINS in experiment-1, benefiting from the tightly-coupled design. However, it illustrates worse long-term accuracy and rotation accuracy than the proposed LE-VINS. The testing environments with fewer structured scenes are not conducive for the solid-state LiDAR with a small FOV, while the LiDAR subsystem plays a more critical role in R2LIVE. Hence, R2LIVE exhibits worse long-term accuracy, corresponding to the absolute-error results.

TABLE II
THE ABSOLUTE ROTATION AND TRANSLATION ERROR

ARE / ATE (deg / m)	Experiment-1	Experiment-2
R2LIVE	2.57 / 4.54	1.02 / 1.96
IC-VINS	**0.43** / 1.67	0.59 / 1.20
LE-VINS-LS	0.80 / 1.99	0.38 / **0.99**
LE-VINS	0.46 / **1.37**	**0.37** / 1.06

The bold results denote the best within different methods.

TABLE III
THE RELATIVE ROTATION AND TRANSLATION ERROR

RRE / RTE (deg / %)	25 m	50 m	100 m	200 m
Experiment-1				
R2LIVE	0.45 / **0.71**	0.57 / **0.64**	0.81 / 0.67	1.27 / 0.81
IC-VINS	0.16 / 1.05	0.21 / 0.83	0.29 / 0.71	0.41 / 0.60
LE-VINS-LS	0.17 / 1.00	0.21 / 0.89	0.28 / 0.81	0.46 / 0.72
LE-VINS	**0.15** / 0.73	**0.19** / 0.67	**0.25** / **0.60**	**0.37** / **0.52**
Experiment-2				
R2LIVE	0.40 / 0.97	0.53 / 0.76	0.76 / 0.68	1.10 / 0.71
IC-VINS	0.14 / 0.79	0.19 / 0.57	0.27 / 0.48	0.38 / 0.41
LE-VINS-LS	**0.13** / 0.64	**0.15** / 0.50	**0.19** / 0.39	**0.28** / **0.32**
LE-VINS	**0.13** / **0.49**	0.16 / **0.41**	0.21 / **0.35**	0.29 / **0.32**

The bold results denote the best within different methods.

In conclusion, LE-VINS exhibits notably improved accuracy compared to the original VINS, IC-VINS. Besides, LE-VINS yields better robustness and consistency than the existing LiDAR-enhanced method in different challenging environments. Furthermore, the proposed LE-VINS exhibits superior accuracy than the state-of-the-art LiDAR-visual-inertial navigation system. It demonstrates that the proposed LiDAR-enhanced method can significantly improve system accuracy.

C. *Evaluation of the Robustness*

1) *Comparison of the Short-term Error*

Table III indicates that the differences between IC-VINS, LE-VINS and LE-VINS-LS are mainly in the short-term translation error in the two experiments. Besides, the short-term accuracy can reflect the local consistency or robustness [42].

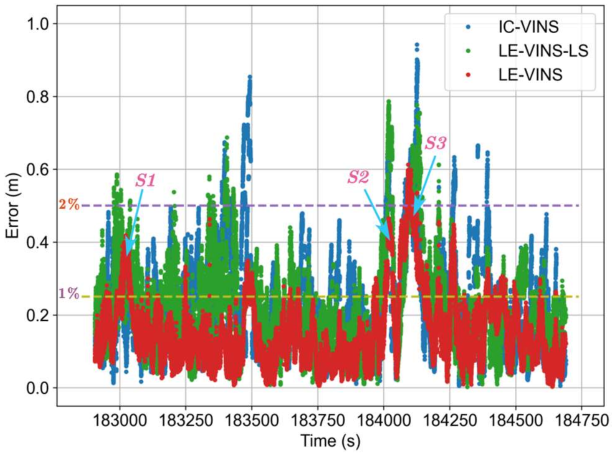

Fig. 10. The relative translation error over 25 m in experiment-1. *S1*, *S2*, and *S3* correspond the degenerated scenes in Fig. 8.

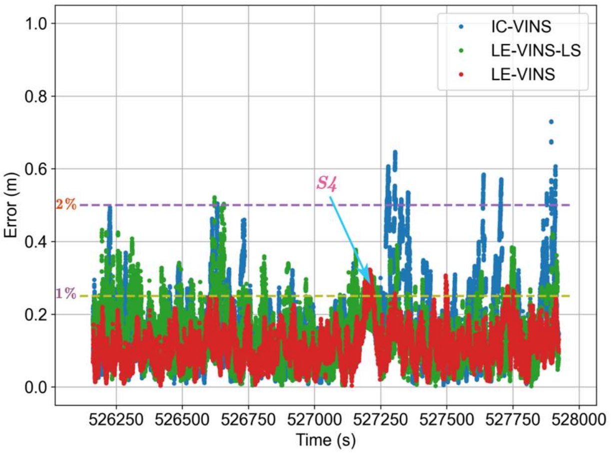

Fig. 11. The relative translation error over 25m in experiment-2. *S4* corresponds to the degenerated scene in Fig. 9.

Hence, the detailed results of the short-term translation error are presented in this part.

The relative translation error over the sub-sequence length of 25 m in experiment-1 and experiment-2 are depicted in Fig. 10 and Fig. 11. For IC-VINS, there are many bad cases where the RTEs are larger than 2%. These bad cases denote the visual-degenerated scenes in the two experiments, as the errors are more notable in these cases. The results indicate that IC-VINS is not accurate enough in these degenerated scenes, exhibiting insufficient robustness.

Compared to IC-VINS, the proposed LE-VINS demonstrates significantly improved accuracy. Almost all the relative translation errors are minor than 2%, except for one extremely challenging scene in experiment-1, as depicted in Fig. 10. Besides, most of the relative translation errors are smaller than 1%, which illustrates the superior accuracy of LE-VINS, benefiting from the robust depth-association method and the employed LiDAR depth factor in the FGO. In contrast, there are still many bad cases for LE-VINS-LS, where the relative translation errors are more significant than 2%. The results demonstrate that the existing LiDAR-enhanced algorithm is not accurate enough because of the rough depth-association method and the inefficient use of the associated LiDAR depth. It can be concluded that the proposed LE-VINS can significantly improve the system accuracy by fully using the accurate LiDAR depth.

However, there are several degenerated scenes for LE-VINS where the relative translation errors are more significant than 1%. We picked up four degenerated scenes for LE-VINS, denoted as *S1*, *S2*, *S3*, and *S4*, as shown in Fig. 8, Fig. 9, Fig. 10, and Fig. 11. According to our analysis, scene *S1* is located in a narrow passage several large walls parallel to the road, and the error of the extrinsic parameters may result in wrong or inaccurate depth estimation. High dynamic (continuous rotation) happens in scene *S2*, and thus the error of the MEMS IMU is significant. In scene *S3*, severe illumination changes occur, resulting in fewer usable areas in the image. Scene *S4* is mainly caused by moving objects, especially motor bicycles and vehicles. As LE-VINS is a visual-based system, it may still be affected by these visual-degenerated scenes to a certain extent. Nevertheless, the accuracy is notably improved in *S2* and *S3*, and without notable degradation in *S1* and *S4*, compared to IC-VINS and LE-VINS-LS. Hence, the results in these scenes are acceptable for LE-VINS.

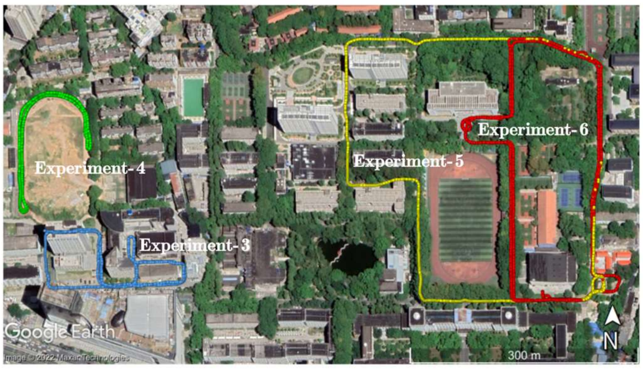

Fig. 12. The test scenes in various environments. Here, different colors represent different experiments.

2) Evaluation in Various Environments

To further evaluate the robustness of LE-VINS, we conducted different experiments in various environments. Specifically, four experiments were conducted on the Wuhan University campus, as depicted in Fig. 12. Experiment-3 (1151 meters) was conducted around the Xinghu building group, where there are drastic illumination changes, repetitive textures, and lots of moving objects. Experiment-4 (1657 meters) was conducted in an abandoned playground with a massive mound. Experiment-5 (2321 meters) and experiment-6 (1539 meters) were all carried on in complex campus scenes, where there are quantities of trees and moving objects, including pedestrians, bicycles, and vehicles.

The trajectory shape might have more impact on the absolute error than the relative error. Hence, the relative pose error is employed to evaluate the robustness of LE-VINS in various environments, as shown in Table IV. The results indicate that

TABLE IV
THE RELATIVE ERROR OF LE-VINS IN VARIOUS ENVIRONMENTS

RRE / RTE (deg / %)	25 m	50 m	100 m	200 m
Experiment-3	0.14 / 0.73	0.17 / 0.67	0.25 / 0.60	0.38 / 0.54
Experiment-4	0.13 / 0.49	0.16 / 0.44	0.20 / 0.38	0.29 / 0.29
Experiment-5	0.16 / 0.64	0.23 / 0.58	0.36 / 0.59	0.54 / 0.67
Experiment-6	0.18 / 0.64	0.22 / 0.58	0.27 / 0.54	0.42 / 0.57

The test scenes for these four experiments are depicted in Fig. 12.

TABLE V
AVERAGE RUNNING TIMES OF LE-VINS

PC / On-board (ms)	Front-end	Depth Association	FGO
Experiment-1	19.6 / 51.9	8.0 / 18.8	23.6 / 147.6
Experiment-2	19.6 / 50.5	7.7 / 18.2	22.2 / 137.9

The front-end includes the feature detection, feature tracking, depth association, and triangulation.

LE-VINS demonstrates very similar accuracy in various environments, which is identical to the results in Table III. As experiment-3 is less challenging, LE-VINS achieves the best accuracy. Nevertheless, LE-VINS achieves similar accuracy in various challenging environments with visual-degenerated scenes, including illumination changes, repetitive textures, and moving objects. The results demonstrate that LE-VINS yield superior robustness in various environments.

D. Running time analysis

The average running times of LE-VINS in experiment-1 and experiment-2 are exhibited in Table V. Non-keyframes are not included in the running-time statistic, and the average interval of the keyframes is around 200 ms. The depth association and FGO are only implemented when a keyframe is selected. Hence, LE-VINS can run in real-time on both the desktop PC (AMD R7-3700X and 32GB RAM) and onboard ARM computer (NVIDIA Xavier and 32GB RAM).

VI. CONCLUSIONS

This study proposes a robust solid-state-LiDAR-enhanced visual-inertial navigation system for low-speed robots. The solid-state LiDAR with the non-repetitive scanning pattern is employed to provide accurate and long-distance depth for visual landmarks. With the estimated depth by the robust depth-association method, the visual feature, LiDAR depth, and IMU measurements are tightly fused within the FGO framework to achieve MAP estimation. Field tests were conducted on a low-speed robot in various large-scale challenging environments for quantitative evaluation. The results demonstrate that the proposed LE-VINS yields superior robustness and accuracy compared to the state-of-the-art navigation systems. Besides, LE-VINS achieves improved robustness compared to the existing LiDAR-enhanced method, benefiting from the robust depth-association algorithm and the LiDAR depth factor in the FGO.

The system accuracy is improved by incorporating the solid-state LiDAR to provide accurate depth for visual landmarks. However, there are still some challenging scenes for LE-VINS, where the relative error (over 25 meters) is larger than 1% and even 2%. In addition, the solid-state LiDAR is only employed to provide depth for visual landmarks. Hence, future work is to implement a tightly-coupled LiDAR-visual-inertial navigation system to utilize all the measurements fully and achieve a more robust and accurate pose estimation.

ACKNOWLEDGEMENT

This research is partly funded by the National Key Research and Development Program of China (No. 2020YFB0505803) and the National Natural Science Foundation of China (No. 41974024). The authors would also like to thank Zaixing Zhang, Linfu Wei, and Guan Wang in our research group for their help in collecting the datasets.

REFERENCES

[1] J. Zhang, M. Kaess, and S. Singh, "A real-time method for depth enhanced visual odometry," *Auton. Robots*, vol. 41, no. 1, pp. 31–43, Jan. 2017.

[2] T. Qin, P. Li, and S. Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[3] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, 2021.

[4] N. El-Sheimy and A. Youssef, "Inertial sensors technologies for navigation applications: state of the art and future trends," *Satell. Navig.*, vol. 1, no. 1, p. 2, Jan. 2020.

[5] H. Tang, X. Niu, T. Zhang, Y. Li, and J. Liu, "OdoNet: Untethered Speed Aiding for Vehicle Navigation Without Hardware Wheeled Odometer," *IEEE Sens. J.*, vol. 22, no. 12, pp. 12197–12208, Jun. 2022.

[6] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[7] J. H. Jung, S. Heo, and C. G. Park, "Observability Analysis of IMU Intrinsic Parameters in Stereo Visual–Inertial Odometry," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 7530–7541, 2020.

[8] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel(R) RealSense(TM) Stereoscopic Depth Cameras," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1267–1276.

[9] J. Yuan, S. Zhu, K. Tang, and Q. Sun, "ORB-TEDM: An RGB-D SLAM Approach Fusing ORB Triangulation Estimates and Depth Measurements," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[10] J. Graeter, A. Wilczynski, and M. Lauer, "LIMO: Lidar-Monocular Visual Odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Oct. 2018, pp. 7872–7879.

[11] S.-S. Huang, Z.-Y. Ma, T.-J. Mu, H. Fu, and S.-M. Hu, "Lidar-Monocular Visual Odometry using Point and Line Features," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, Paris, France, May 2020, pp. 1091–1097.

[12] D. Wisth, M. Camurri, S. Das, and M. Fallon, "Unified Multi-Modal Landmark Tracking for Tightly Coupled Lidar-Visual-Inertial Odometry," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1004–1011, Apr. 2021.

[13] K. Huang, J. Xiao, and C. Stachniss, "Accurate Direct Visual-Laser Odometry with Explicit Occlusion Handling and Plane Detection," in *2019 International Conference on Robotics and Automation (ICRA)*, Montreal, QC, Canada, May 2019, pp. 1295–1301.

[14] Y.-S. Shin, Y. S. Park, and A. Kim, "DVL-SLAM: sparse depth enhanced direct visual-LiDAR SLAM," *Auton. Robots*, vol. 44, no. 2, pp. 115–130, Jan. 2020.

[15] 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), 2021, pp. 5692–5698.

[16] S. Chiodini, R. Giubilato, M. Pertile, and S. Debei, “Retrieving Scale on Monocular Visual Odometry Using Low-Resolution Range Sensors,” *IEEE Trans. Instrum. Meas.*, vol. 69, no. 8, pp. 5875–5889, 2020.

[17] K. Li, M. Li, and U. D. Hanebeck, “Towards High-Performance Solid-State-LiDAR-Inertial Odometry and Mapping,” *IEEE Robot. Autom. Lett.*, pp. 1–1, 2021.

[18] W. Xu and F. Zhang, “FAST-LIO: A Fast, Robust LiDAR-Inertial Odometry Package by Tightly-Coupled Iterated Kalman Filter,” *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 3317–3324, Apr. 2021.

[19] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, “FAST-LIO2: Fast Direct LiDAR-Inertial Odometry,” *IEEE Trans. Robot.*, pp. 1–21, 2022.

[20] 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)*, May 2020, pp. 3126–3131.

[21] X. Liu, C. Yuan, and F. Zhang, “Targetless Extrinsic Calibration of Multiple Small FoV LiDARs and Cameras Using Adaptive Voxelization,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[22] Y. Zhu, C. Zheng, C. Yuan, X. Huang, and X. Hong, “CamVox: A Low-cost and Accurate Lidar-assisted Visual SLAM System,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 5049–5055.

[23] J. Lin, C. Zheng, W. Xu, and F. Zhang, “R ^{2} LIVE: A Robust, Real-Time, LiDAR-Inertial-Visual Tightly-Coupled State Estimator and Mapping,” *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7469–7476, Oct. 2021.

[24] J. Lin and F. Zhang, “R3LIVE: A Robust, Real-time, RGB-colored, LiDAR-Inertial-Visual tightly-coupled state Estimation and mapping package,” in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 10672–10678.

[25] C. Zheng, Q. Zhu, W. Xu, X. Liu, Q. Guo, and F. Zhang, “FAST-LIVO: Fast and Tightly-coupled Sparse-Direct LiDAR-Inertial-Visual Odometry,” *ArXiv220300893 Cs*, Mar. 2022. [Online]. Available: http://arxiv.org/abs/2203.00893

[26] J. Civera, A. J. Davison, and J. M. M. Montiel, “Inverse Depth Parametrization for Monocular SLAM,” *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 932–945, Oct. 2008.

[27] F. Dellaert and M. Kaess, “Factor Graphs for Robot Perception,” *Found. Trends Robot.*, vol. 6, no. 1–2, pp. 1–139, 2017.

[28] P. Wang, Z. Fang, S. Zhao, Y. Chen, M. Zhou, and S. An, “Vanishing Point Aided LiDAR-Visual-Inertial Estimator,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13120–13126.

[29] S. Zhao, H. Zhang, P. Wang, L. Nogueira, and S. Scherer, “Super Odometry: IMU-centric LiDAR-Visual-Inertial Estimator for Challenging Environments,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2021, pp. 8729–8736.

[30] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[31] C. Forster, M. Pizzoli, and D. Scaramuzza, “SVO: Fast semi-direct monocular visual odometry,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, May 2014, pp. 15–22.

[32] J. Engel, V. Koltun, and D. Cremers, “Direct Sparse Odometry,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[33] X. Niu, H. Tang, T. Zhang, J. Fan, and J. Liu, “IC-GVINS: A Robust, Real-Time, INS-Centric GNSS-Visual-Inertial Navigation System,” *IEEE Robot. Autom. Lett.*, vol. 8, no. 1, pp. 216–223, Jan. 2023.

[34] J. Sola, “Quaternion kinematics for the error-state Kalman filter,” *ArXiv Prepr. ArXiv171102508*, 2017.

[35] J. Zhang and S. Singh, “LOAM: Lidar Odometry and Mapping in Real-time,” in *Robotics: Science and Systems X*, Jul. 2014.

[36] Agarwal, Sameer, Mierle, and Keir, “Ceres Solver — A Large Scale Nonlinear Optimization Library,” 2022. [Online]. Available: http://ceres-solver.org/

[37] H. Tang, T. Zhang, X. Niu, J. Fan, and J. Liu, “Impact of the Earth Rotation Compensation on MEMS-IMU Preintegration of Factor Graph Optimization,” *IEEE Sens. J.*, vol. 22, no. 17, pp. 17194–17204, Sep. 2022.

[38] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, “Keyframe-based visual–inertial odometry using nonlinear optimization,” *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, Mar. 2015.

[39] W. Liu, M. Wu, X. Zhang, W. Wang, W. Ke, and Z. Zhu, “Single-epoch RTK performance assessment of tightly combined BDS-2 and newly complete BDS-3,” *Satell. Navig.*, vol. 2, no. 1, p. 6, Apr. 2021.

[40] 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*, 2013, pp. 1280–1286.

[41] M. Grupp, “evo.” Jul. 21, 2022. [Online]. Available: https://github.com/MichaelGrupp/evo

[42] Z. Zhang and D. Scaramuzza, “A Tutorial on Quantitative Trajectory Evaluation for Visual(-Inertial) Odometry,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Oct. 2018, pp. 7244–7251.

Hailiang Tang received the B.E. and M.E. degrees from Wuhan University, China, in 2017 and 2020, respectively. He is pursuing a Ph.D. in communication and information systems with the GNSS Research Center, Wuhan University. His current research interests include GNSS/INS integration technology, deep learning, visual SLAM, and autonomous robotics system.

Xiaoji Niu received his bachelor’s and Ph.D. degrees from the Department of Precision Instruments, Tsinghua University, in 1997 and 2002, respectively.

He did post-doctoral research with the University of Calgary and worked as a Senior Scientist in SiRF Technology Inc. He is currently a Professor with the GNSS Research Center, Wuhan University, China. He has published more than 90 academic papers and own 28 patents. He leads a multi-sensor navigation group focusing on GNSS/INS integration, low-cost navigation sensor fusion, and its new applications.

Tisheng Zhang is an associate professor in GNSS Research Center at Wuhan University, China. He holds a B.SC. and a Ph.D. in Communication and Information Systems from Wuhan University, Wuhan, China, in 2008 and 2013, respectively. From 2018 to 2019, he was a Post Doctor at the Hong Kong Polytechnic University.

His research interests focus on the fields of GNSS receiver and multi-sensor deep integration.

Liqiang Wang received the B.Eng. (with honors) degree in electronic information engineering from Wuhan University, Wuhan, China, in 2020, where he is pursuing a master's degree in navigation, guidance, and control with the GNSS Research Center. His primary research interests include GNSS/INS integrations and visual-based navigation.

Jingnan Liu, member of Chinese Academy of Engineering, professor, Ph. D supervisor. He is an expert in geodesy and surveying engineering with the specialty of GNSS technology and applications. He has been engaged in the research of geodetic theories and applications, including national coordinate system establishment, GNSS technology and software development, as well as large project implementation. Over the past few decades, he has been engaged in the research of geodetic theories and applications. He has published more than 150 academic papers and supervised more than 100 postgraduates.