LVI-ExC: A Target-free LiDAR-Visual-Inertial Extrinsic Calibration Framework

Zhong Wang
School of Software Engineering, Tongji University
Shanghai, China
2010194@tongji.edu.cn

Lin Zhang
School of Software Engineering, Tongji University
Shanghai, China
cslinzhang@tongji.edu.cn

Ying Shen*
School of Software Engineering, Tongji University
Shanghai, China
yingshen@tongji.edu.cn

Yicong Zhou
Department of Computer and Information Science,
University of Macau, China
yicongzhou@um.edu.mo

ABSTRACT

Recently, the multi-modal fusion with 3D LiDAR, camera, and IMU has shown great potential in applications of automation-related fields. Yet a prerequisite for a successful fusion is that the geometric relationships among the sensors are accurately determined, which is called an extrinsic calibration problem. To date, the existing target-based approaches to deal with this problem rely on sophisticated calibration objects (sites) and well-trained operators, which is time-consuming and inflexible in practical applications. Contrarily, a few target-free methods can overcome these shortcomings, while they only focus on the calibrations of two types of the sensors. Although it is possible to obtain LiDAR-visual-inertial extrinsics by chained calibrations, problems such as cumbersome operations, large cumulative errors, and weak geometric consistency still exist. To this end, we propose **LVI-ExC**, an integrated **L**iDAR-**V**isual-**I**nertial **Ex**trinsic **C**alibration framework, which takes natural multi-modal data as input and yields sensor-to-sensor extrinsics end-to-end without any auxiliary object (site) or manual assistance. To fuse multi-modal data, we formulate the LiDAR-visual-inertial extrinsic calibration as a continuous-time simultaneous localization and mapping problem, in which the extrinsics, trajectories, time differences, and map points are jointly estimated by establishing sensor-to-sensor and sensor-to-trajectory constraints. Extensive experiments show that LVI-ExC can produce precise results. With LVI-ExC's outputs, the LiDAR-visual reprojection results and the reconstructed environment map are all highly consistent with the actual natural scenes, demonstrating LVI-ExC's outstanding performance. To ensure that our results are fully reproducible, all the relevant data and codes have been released publicly[1].

*Corresponding author: Ying Shen
[1]https://cslinzhang.github.io/LVI-ExC/LVI-ExC.html

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547878

CCS CONCEPTS

• **Computing methodologies** → **Camera calibration**.

KEYWORDS

Extrinsic calibration, Multi-modal fusion, Target-free calibration

ACM Reference Format:
Zhong Wang, Lin Zhang, Ying Shen, and Yicong Zhou. 2022. LVI-ExC: A Target-free LiDAR-Visual-Inertial Extrinsic Calibration Framework. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547878

1 INTRODUCTION

Today, the sensor suite which consists of multi-beam LiDARs (light detection and ranging), optical cameras, and IMUs (inertial measurement unit) is a common configuration for many unmanned drones, robots, and vehicles. Likewise, perception, localization, and mapping via multi-sensor fusion also play prominent roles in these intelligence agents [18, 19, 32]. Yet before such multi-sensor data can be meaningfully fused, the extrinsics among the LiDAR, the camera, and the IMU must be precisely determined.

From the point of view of whether or not to resort to auxiliary calibrators, existing *inter-sensor* (LiDAR-camera, camera-IMU, or LiDAR-IMU) calibration schemes can be roughly classified into two categories, target-based and target-free. The target-based ones generally rely on precise calibration objects (sites), such as checkerboards [14], folding planes [2], plates with circular holes [10], ubiquitous cartons [30], etc. Although these approaches can conveniently establish geometric constraints and estimate the extrinsics with high precision, they suffer from the following defects in practical applications: 1) it is time-consuming, costly, and laborious to produce sophisticated calibration objects (sites); 2) this kind of methods requires operators to be highly professional; and 3) mechanical vibrations are inevitable during carriers' online movements, leading to extrinsic drifts, while the target-based methods are limited to offline calibration and thus can not rectify the extrinsics in time.

Contrarily, the target-free approaches try to estimate the extrinsics directly from natural scenes. This class of methods usually determine the extrinsics via "hand-eye calibration", which is based on the assumption that the rigidly attached sensors share the same motion over time [3]. However, the performance of hand-eye calibration is limited due to the reasons as follows. For one aspect, its

accuracy depends on the precision of sensor-independent odometry estimation. Up to now, however, the error of pose estimation with a single sensor still can not be ignored. For another aspect, the data collected by different sensors have time differences, which will contradict the essential premise of the equal relative motion in hand-eye calibration, resulting in extra errors. An elegant way to deal with these two problems is to model the extrinsic calibration under the CT-SLAM (continuous-time simultaneous localization and mapping) framework [6, 8]. With CT-SLAM, the sensor observations at any time can be conveniently fused. To date, few studies have employed the CT-SLAM techniques to conduct target-free LiDAR-IMU or LiDAR-camera calibrations [11, 16, 21]. Although the *sensor-to-sensor* (LiDAR-camera, camera-IMU, and LiDAR-IMU) extrinsics can be obtained by chained calibrations (e.g., the LiDAR-IMU extrinsics can be inferred via chained LiDAR-camera and camera-IMU calibrations), there are still some problems such as the cumbersome operations, large accumulated errors, and geometric inconsistency of the calibration results.

Taking aforementioned analysis into considerations, in this article, we attempt to integrate **L**iDAR, **V**isual, and **I**nertial information into a unified **Ex**trinsic **C**alibration framework, **LVI-ExC** for short, and jointly estimate the sensor-to-sensor extrinsics directly from the data collected in natural scenes. The features of LVI-ExC and our contributions can be summarized as follows:

(1) As far as we know, LVI-ExC is the first totally target-free integrated framework that can calibrate extrinsics among 3D LiDAR, camera and IMU jointly. Compared with the target-based ones, LVI-ExC gets rid of the dependence on the professional operators and the expensive sophisticated calibrators, which makes it possible to rectify the extrinsics online when the carrier is working. Compared with the chained calibration, LVI-ExC only takes the multi-modal data collected from natural scenes as the input and can output the extrinsics end-to-end, which avoids cumbersome operations and effectively fuses the multi-modal information, thus greatly improving the calibration efficiency and eliminating cumulative errors.
(2) Considering the time asynchrony among the three sensors, we formulate the LiDAR-visual-inertial calibration as a CT-SLAM problem. The loss terms with respect to LiDAR measurements, visual features, and IMU readings are constructed in a unified framework. These terms help LVI-ExC to fuse the multi-modal data in deep and produce results with strong geometric consistency. To guarantee a smooth optimization of these terms, a reasonable initialization approach of all the relevant variables is also presented, which can be seamlessly embedded in any LiDAR-inertial or visual-inertial extrinsic calibration system.
(3) To verify the effectiveness of LVI-ExC, we developed a hand-held device (Fig. 1 (a)) consisting of a LiDAR, a camera, and an IMU, gathered several data sequences from various natural scenes, and made comprehensive experimental analysis. The results showed that LVI-ExC can achieve comparable accuracy with state-of-the-art target-based approaches. With the calibrated extrinsics, the results of the LiDAR scan reprojected to the image and the image reprojected to the point cloud map were highly consistent with the actual scenes. To ensure a high degree of reproducibility of our results and to benefit the community, we have open-sourced all the relevant codes and data on an online website[1].

2 RELATED WORK

2.1 Target-based Approaches

To date, most of the visual-inertial and LiDAR-camera extrinsic calibrations are target-based. Often, the visual-inertial ones rely on checkerboards for camera pose estimation. In [24], based on the extended Kalman filter, Mirzaei and Roumeliotis combined the camera-IMU extrinsics, time difference, carrier pose, and IMU biases into a unified state vector, performed state recursion with IMU readings and updated the state with visual inputs. Similarly, Kelly and Sukhatme [14] attempted to extend the state vector defined in [24] with 3D visual map points. However, only the results with known map points were provided in their experiments. To take full advantage of multi-frame information, some studies modeled the camera-IMU extrinsic calibration as a CT-SLAM problem. For example, Fleps *et al.* [6] conducted camera-IMU extrinsic calibration under the CT-SLAM framework, which estimates the camera pose in the world frame from an auxiliary checkerboard and obtains the extrinsics by associating IMU measurements and camera poses. Later, Furgale *et al.* [8] constructed a complete continuous-time SLAM theory stemming from the Bayesian law and validated its effectiveness via camera-IMU calibration. Although the auxiliary checkerboards ease these visual-inertial calibrations, as Fleps *et al.* pointed out in [6], the dependence of the boards also limits the motion of the carrier, which in turn adversely affects the observability of the IMU.

For LiDAR-camera extrinsic calibration, the target-based methods are usually carried out in static states, which requires sufficient common view areas between the two sensors. For example, by putting a carton with a known side length in the co-visible area of the camera and LiDAR, Pusztai and Hajder [30] extracted the planes from LiDAR point clouds and the corresponding edges from images, and further formulated the extrinsic calibration as a perspective-n-point problem; In [34], with a checkerboard, by finding the plane where the board is located and its edge lines from the point cloud, Zhou *et al.* extracted the edge points of the board plane from the image and established point-to-line and point-to-plane constraints to estimate the extrinsic parameters; Guindel *et al.* [10] designed a calibration board with a checkerboard and four circular holes on it, and established constraints between the circles fitted from the point cloud and the known visual points to estimate the extrinsics. Although these methods simplify the design of the algorithm by using auxiliary calibrators, they also limit themselves to offline calibrations and the requirement for a co-visible area between the LiDAR and the camera is actually not always met.

2.2 Target-free Approaches

The target-free manner is usually encountered for the LiDAR-IMU calibration. In [9], Geiger *et al.* estimated LiDAR-IMU extrinsics resorting to the hand-eye calibration. Their approach expects that each sensor can estimate the trajectory accurately and that the sensors are precisely time-synchronized, which is, in fact, difficult to meet in practice. In [16], Gentil *et al.* extracted the geometric planes

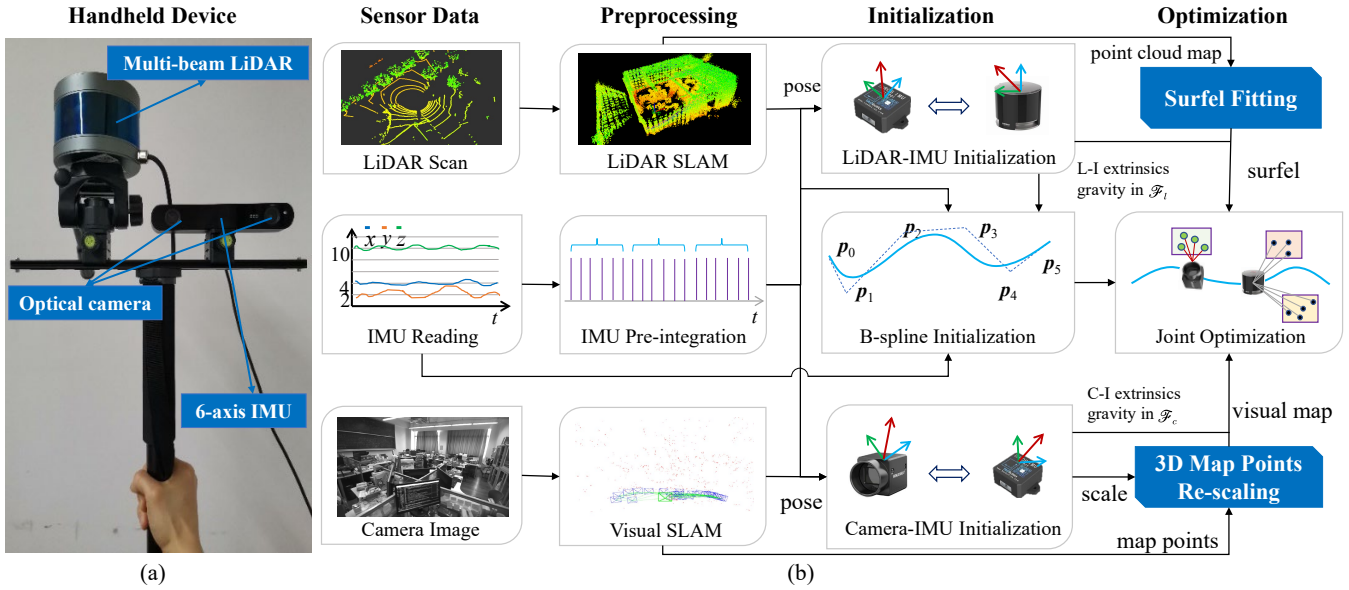


Figure 1: The self-developed handheld device (a) and the framework of LVI-ExC (b). LVI-ExC takes LiDAR-visual-inertial data as input, roughly estimates the sensor-to-sensor extrinsics via LiDAR-IMU, camera-IMU, and B-spline initialization, and jointly optimize them subsequently by establishing sensor-to-sensor and sensor-to-trajectory constraints.

from the first frame of point clouds, established point-to-plane constraints, and modeled the calibration as a graph optimization problem. Although their approach did not rely on specific calibration objects, fitting planes from a single frame alone was not stable enough and actually required a higher regularity of the planes in the scene. Under the framework of CT-SLAM, Lv *et al.* proposed LI-Calib [21] to calibrate the LiDAR-IMU extrinsics. LI-Calib fitted surfels from coarse-aligned point cloud maps, established point-to-surfel constraints, and optimized the extrinsics in the CT-SLAM framework. Further, its accuracy was improved by undistorting the point clouds and performing iterative calibrations.

In a similar manner, several LiDAR-camera calibration methods resort to the hand-eye calibration to estimate the extrinsics. For instance, Taylor and Nieto [31] calibrated the LiDAR-camera-GPS extrinsics simultaneously based on the motion constraints among the sensors; Ishikawa *et al.* [11] designed a two-stage scheme, which determined the approximate LiDAR-camera extrinsics by the hand-eye calibration, and skillfully established frame-to-frame constraints via Lucas–Kanade tracking [20]. Similar to [11], Park *et al.* [28] put forward a solution to calibrate the LiDAR-camera extrinsics from coarse to fine. Due to the inherent shortcomings of the hand-eye calibration described in Sect. 1, how to properly tune its output for target-free LiDAR-camrera calibration still remains as an open problem.

3 METHODOLOGY

3.1 Framework Overview

As illustrated in Fig. 1 (b), when the multi-sensor data sequence flows in, LVI-ExC performs the extrinsic calibration via three stages, i.e., a preprocessing stage, an initialization stage, and a joint optimization stage. In the preprocessing stage, the incoming data is processed by LiDAR SLAM, IMU pre-integration, and visual SLAM to obtain pre-integrations, LiDAR (camera) poses, and LiDAR (visual) maps. In the initialization stage, the preprocessed data is employed to roughly align the LiDAR and the camera to the IMU. Meanwhile, the carrier trajectory is also fitted from the IMU readings, LiDAR poses, and LiDAR-IMU extrinsics estimated. In the optimization stage, the coarsely estimated results are jointly optimized in a unified framework by establishing sensor-to-sensor and sensor-to-trajectory constraints.

3.2 Preliminary Knowledge and Preprocessing

3.2.1 Notation. We use $(\cdot)^w$, $(\cdot)^b$, $(\cdot)^c$, and $(\cdot)^l$ to denote a quantity in the world frame $\mathcal{F}_w$, the body frame (IMU frame) $\mathcal{F}_b$, the camera frame $\mathcal{F}_c$, and the LiDAR frame $\mathcal{F}_l$, respectively. The z-axis of $\mathcal{F}_w$ is assumed to be vertical to the horizontal plane. The right subscript stands for the owner or reference time of the state quantity. The right superscript denotes the reference frame. The left superscript implies some special attributes depending on the specific context. A rotation matrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ or a quaternion $\boldsymbol{q} = [q_w, \boldsymbol{q}_v^T]^T \in \mathbb{R}^4$ is indiscriminately utilized to denote a 3D rotation, where q_w ($\boldsymbol{q}_v^T$) is the real (imaginary) part of $\boldsymbol{q}$. $\boldsymbol{p}$ ($\boldsymbol{v}$) $\in \mathbb{R}^3$ denotes the 3D spatial position (velocity) of the carrier. $\boldsymbol{g} \in \mathbb{R}^3$ represents the gravity. For compactness of expression, we use $\boldsymbol{T}$ to denote the compound transformation of $\boldsymbol{R}$ and $\boldsymbol{p}$.

3.2.2 LiDAR/Visual SLAM and CT-SLAM. Since the constraint between an IMU and a camera (LiDAR) can only be established through their motion relationships, it is required to accurately estimate the camera's (LiDAR's) poses, which is generally achieved via the visual (LiDAR) SLAM technology. A SLAM system not only

estimates the camera (LiDAR) poses but also reconstructs the environmental map simultaneously. In the preprocessing stage, the input LiDAR scans and camera images are fed into two high-precision SLAM systems, LOAM [33] and ORB-SLAM2 [26], to generate poses and maps, respectively. Note that the scale of a visual SLAM system remains unknown. Hence, to restore the visual map points and the camera's physical trajectory, it's also necessary to determine the absolute scale of the visual SLAM in the subsequent stages.

Although the aforementioned SLAM systems can effectively process data from a single sensor, it becomes challenging to build data associations among sensors with different timestamps when the system input is multi-modal data. To make up for this shortcoming, Furgale *et al.* [8] generalized the traditional discrete-time SLAM to CT-SLAM whose trajectory is no longer discrete poses but continuous curves with time as the independent variable. To fully fuse the LiDAR-camera-IMU data, we consider combining all their measurements together in the CT-SLAM framework. Next, we will introduce the continuous-time trajectory representation in LVI-ExC beforehand and discuss how to formulate our calibration as a CT-SLAM problem in detail in the optimization stage.

3.2.3 Trajectory Representation. To ease state inference and update, we expect the mathematical representation of the trajectory to have the following properties, local controllability and analytic second-order derivability. Among the parametric curves with such properties, B-spline is an ideal choice since it is a $(k-1)$-th order piecewise polynomial and is C^{k-2} continuous (k is the degree of the B-spline). According to [4], the trajectory of a carrier in 3D space can be expressed with the B-spline,

$$\boldsymbol{p}(t) = \sum_{i=0}^{n} \boldsymbol{p}_i B_{i,k}(t), \tag{1}$$

where $\boldsymbol{p}_i \in \mathbb{R}^3$ is the i-th control point, n is the max index of the control points, and $B_{i,k}(t)$ is the corresponding basic function,

$$B_{i,1}(t) = \begin{cases} 1 & \text{if } t \in [t_i, t_{i+1}), \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

$$B_{i,k}(t) = \left(\frac{t - t_i}{t_{i+k-1} - t_i}\right) B_{i,k-1}(t) + \left(\frac{t_{i+k} - t}{t_{i+k} - t_{i+1}}\right) B_{i+1,k-1}(t). \tag{3}$$

This $\mathbb{R}^3$ trajectory can also be given in a cumulative form,

$$\boldsymbol{p}(t) = \boldsymbol{p}_0 \bar{B}_{0,k}(t) + \sum_{i=1}^{n} (\boldsymbol{p}_i - \boldsymbol{p}_{i-1}) \bar{B}_{i,k}(t), \bar{B}_{i,k}(t) = \sum_{j=i}^{n} B_{j,k}(t). \tag{4}$$

For the rotation trajectory in the special orthogonal group ($\mathbb{SO}(3)$), Kim *et al.* proposed the following representation [15],

$$\boldsymbol{q}(t) = \boldsymbol{q}_0^{\bar{B}_{0,k}(t)} \otimes \prod_{i=1}^{n} \mathrm{Exp}(\mathrm{Log}(\boldsymbol{q}_{i-1}^{*} \otimes \boldsymbol{q}_i) \bar{B}_{i,k}(t), \tag{5}$$

where $\boldsymbol{q}_{i-1}^{*}$ is the conjugate quaternion of $\boldsymbol{q}_{i-1}$, $\otimes$ means the quaternion multiplication, $\mathrm{Exp}(\cdot)$ maps an element in $\mathfrak{so}(3)$ (the Lie algebra of $\mathbb{SO}(3)$) to $\mathbb{SO}(3)$, and $\mathrm{Log}(\cdot)$ is the inverse operator of $\mathrm{Exp}(\cdot)$.

3.2.4 IMU Pre-integration. Since both the LiDAR-IMU and camera-IMU initializations have close connections with IMU pre-integration, we briefly introduce it here in advance. According to [7], the relationship among the pre-integration measurements, true states and noises conforms to,

$$\tilde{\boldsymbol{q}}_{b_j}^{b_i} \approx \boldsymbol{q}_{b_i}^{w\,T} \otimes \boldsymbol{q}_{b_j}^{w} \otimes \mathrm{Exp}(\delta \boldsymbol{\phi}_{b_j}^{b_i}), \tag{6}$$

$$\tilde{\boldsymbol{v}}_{b_j}^{b_i} \approx \boldsymbol{R}_{b_i}^{w\,T} (\boldsymbol{v}_{b_j}^{w} - \boldsymbol{v}_{b_i}^{w} - \boldsymbol{g}^{w} \Delta t_{ij}) + \delta \boldsymbol{v}_{b_j}^{b_i}, \tag{7}$$

$$\tilde{\boldsymbol{p}}_{b_j}^{b_i} \approx \boldsymbol{R}_{b_i}^{w\,T} (\boldsymbol{p}_{b_j}^{w} - \boldsymbol{p}_{b_i}^{w} - \boldsymbol{v}_{b_i}^{w} \Delta t_{ij} - \frac{1}{2} \boldsymbol{g}^{w} \Delta {t_{ij}}^2) + \delta \boldsymbol{p}_{b_j}^{b_i}, \tag{8}$$

where $\delta \boldsymbol{\phi}_{b_i}^{b_j}, \delta \boldsymbol{v}_{b_i}^{b_j}$ and $\delta \boldsymbol{p}_{b_i}^{b_j}$ are the Gaussian noises of the pre-integration measurements, and Δt_{ij} is the time difference between the i-th and the j-th IMU readings. How to obtain the pre-integration measurements from IMU raw readings is provided in the supplementary material.

3.3 Initialization

A reasonable initialization is a prerequisite to ensure that the joint optimization can be performed successfully. In LVI-ExC, the initialization involves the coarse estimations of the extrinsics of $\mathscr{F}_c$ in $\mathscr{F}_b$ ($\boldsymbol{R}_c^b$ and $\boldsymbol{p}_c^b$, or $\boldsymbol{T}_c^b$), the ones of $\mathscr{F}_l$ in $\mathscr{F}_b$ ($\boldsymbol{R}_l^b$ and $\boldsymbol{p}_l^b$, or $\boldsymbol{T}_l^b$), and the $\mathbb{SO}(3)$ ($\mathbb{R}^3$) trajectory. Furthermore, the gravity direction in $\mathscr{F}_w$ is also needed to be inferred to align the trajectories to $\mathscr{F}_w$. Next, we present how to properly perform the camera-IMU, LiDAR-IMU, and trajectory initializations.

3.3.1 Camera-IMU Initialization. Since the accelerometer measurement of an IMU is coupled with the gravity while the gyroscope is only involved with the carrier rotation, we first infer the relative rotation from the gyroscope readings. Then, the relative translation, the gravity and the visual map points will be further inferred with the estimated relative rotation.

Initialization of Rotation. Consider two frames of images with timestamp t_i and t_j respectively. According to the hand-eye calibration [3], an IMU and a camera that are rigidly attached shall meet the following equation,

$$\boldsymbol{R}_{b_j}^{b_i} \boldsymbol{R}_c^b = \boldsymbol{R}_c^b \boldsymbol{R}_{c_j}^{c_i}. \tag{9}$$

Further, with the law of quaternion-matrix multiplication,

$$[\boldsymbol{q}]_L = q_w \boldsymbol{I} + \begin{bmatrix} 0 & -\boldsymbol{q}_v^T \\ \boldsymbol{q}_v & [\boldsymbol{q}_v]_\times \end{bmatrix}, [\boldsymbol{q}]_R = q_w \boldsymbol{I} + \begin{bmatrix} 0 & -\boldsymbol{q}_v^T \\ \boldsymbol{q}_v & -[\boldsymbol{q}_v]_\times \end{bmatrix}, \tag{10}$$

where $[\cdot]_L$ ($[\cdot]_R$) denotes the left (right) quaternion multiplication, $[\cdot]_\times$ means the associated skew-symmetric matrix, and $\boldsymbol{I}$ is the identity matrix, Eq. 9 can be reformulated as,

$$\boldsymbol{q}_{b_j}^{b_i} \otimes \boldsymbol{q}_c^b - \boldsymbol{q}_c^b \otimes \boldsymbol{q}_{c_j}^{c_i} = ([\boldsymbol{q}_{b_j}^{b_i}]_L - [\boldsymbol{q}_{c_j}^{c_i}]_R) \boldsymbol{q}_c^b = \boldsymbol{0}. \tag{11}$$

According to Eq. 11, by employing the rotation measurements, $\tilde{\boldsymbol{q}}_{c_j}^{c_i}$ and $\tilde{\boldsymbol{q}}_{b_j}^{b_i}$, of the visual SLAM and IMU pre-integrations respectively, overdetermined linear equations can be constructed to solve the relative rotation $\boldsymbol{q}_c^b$ ($\boldsymbol{R}_c^b$) from the camera to IMU.

Initialization of Translation, Gravity and Visual Map Points. With a reasonable estimate of $\boldsymbol{R}_c^b$, it becomes possible to infer the relative translation $\boldsymbol{p}_c^b$, the gravity at the starting time $\boldsymbol{g}^{b_0}$, and the scale ambiguity (μ) of visual SLAM based on the velocity-related (Eq. 7) and translation-related (Eq. 8) equations for the IMU pre-integration. From Eq. 7 and Eq. 8, when taking the camera frame at the initial moment ($\mathscr{F}_{c_0}$) as the reference, we need to establish the association between the IMU frame at time t ($\mathscr{F}_{b_t}$) and $\mathscr{F}_{c_0}$ first.

According to the chain rule of 3D spatial coordinate transformation, the following equation holds,

$$\begin{bmatrix} \boldsymbol{R}_{b_t}^{c_0} & \mu\boldsymbol{p}_{b_t}^{c_0} \\ \boldsymbol{0} & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_{c_t}^{c_0} & \mu\boldsymbol{p}_{c_t}^{c_0} \\ \boldsymbol{0} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{b}^{c} & \boldsymbol{p}_{b}^{c} \\ \boldsymbol{0} & 1 \end{bmatrix}. \tag{12}$$

With Eq. 12, we can get the following relations,

$$\boldsymbol{R}_{b_t}^{c_0} = \boldsymbol{R}_{c_t}^{c_0}\boldsymbol{R}_{b}^{c}, \mu\boldsymbol{p}_{b_t}^{c_0} = \boldsymbol{R}_{c_t}^{c_0}\boldsymbol{p}_{b}^{c} + \mu\boldsymbol{p}_{c_t}^{c_0} = \mu\boldsymbol{p}_{c_t}^{c_0} - \boldsymbol{R}_{b_t}^{c_0}\boldsymbol{p}_{c}^{b}. \tag{13}$$

Substituting Eq. 13 into Eq. 7 produces,

$$\Delta\tilde{\boldsymbol{v}}_{b_j}^{b_i} \approx \boldsymbol{R}_{b_i}^{c_0\,T}(\boldsymbol{R}_{b_j}^{c_0}\boldsymbol{v}_{b_j}^{b_j} - \boldsymbol{R}_{b_i}^{c_0}\boldsymbol{v}_{b_i}^{b_i} - \boldsymbol{g}^{c_0}\Delta t_{ij}), \tag{14}$$

in which $\boldsymbol{R}_{b_i}^{c_0}$ and $\boldsymbol{R}_{b_j}^{c_0}$ can be obtained via Eq. 13, and $\boldsymbol{v}_{b_i}^{b_i}, \boldsymbol{v}_{b_j}^{b_j}$, and $\boldsymbol{g}^{c_0}$ are the velocities and the gravity to be estimated. Similarly, substituting Eq. 13 into Eq. 8 yields,

$$\begin{aligned}\Delta\tilde{\boldsymbol{p}}_{b_j}^{b_i} \approx \boldsymbol{R}_{b_i}^{c_0\,T}((&\boldsymbol{R}_{c_j}^{c_0}\boldsymbol{p}_{b}^{c} - \boldsymbol{R}_{c_i}^{c_0}\boldsymbol{p}_{b}^{c}) \\ &+ \mu(\boldsymbol{p}_{c_j}^{c_0} - \boldsymbol{p}_{c_i}^{c_0}) - \boldsymbol{v}_{b_i}^{c_0}\Delta t_{ij} - \frac{1}{2}\boldsymbol{g}^{c_0}\Delta {t_{ij}}^2).\end{aligned} \tag{15}$$

To this point, by computing the multiple keyframe poses and pre-integrating the associated IMU readings in a certain time window, the overdetermined equations can be established according to Eq. 14 and Eq. 15, so that the camera-IMU extrinsics and the scale ambiguity can be determined by linear least squares estimation. Furthermore, the depths of all the 3D visual points can be re-scaled with the estimated sale. In addition, with the estimated $\boldsymbol{g}^{c_0}$ and $\boldsymbol{R}_{c}^{b}$, $\boldsymbol{g}^{b_0}$ can be conveniently obtained and all the variables can be roughly transformed to the world frame by aligning $\boldsymbol{g}^{b_0}$ with $\boldsymbol{g}^{w}$. As a result, we get proper estimates of $\boldsymbol{R}_{c}^{b}$, $\boldsymbol{p}_{c}^{b}$ and all 3D visual points via camera-IMU initialization.

3.3.2 LiDAR-IMU Initialization. We employ a technique similar to the camera-IMU initialization to roughly align $\mathscr{F}_l$ with $\mathscr{F}_b$ by establishing the relationship between the IMU pre-integration and the LiDAR odometry. Specifically, we first resort to LOAM [33] to estimate the poses of the point clouds and build the environmental map concurrently. Meanwhile, with the timestamps of the point clouds, the frame-to-frame pre-integrations are propagated from the IMU readings. After that, by employing the technique presented in Sect. 3.3.1, rough estimates of the LiDAR-IMU extrinsics as well as the gravity in the first frame of the scans can be obtained.

As the environmental map constructed by LOAM still has non-negligible errors due to the motion distortion of the laser points [33], we consider establishing more accurate geometric constraints for a finer calibration. A natural idea is to extract surface features from the map. However, regular planes do not always exist stably in natural scenes. Therefore, inspired by [21], we fit surfels from point cloud maps to enhance the environmental adaptability of the framework. Specifically, we first divide the point cloud map constructed by LOAM into spatial voxels. Afterwards, we calculate the first-order and second-order moments of the point clouds within the voxels. According to [1], whether a point cloud constitutes a surfel can be detected by the following surfel likeliness coefficient,

$$\zeta = 2\frac{\lambda_1 - \lambda_0}{\lambda_0 + \lambda_1 + \lambda_2}, \tag{16}$$

where $\lambda_0 \le \lambda_1 \le \lambda_2$, and they are the eigenvalues of the second-order moment matrix. If the point clouds in the voxel are sampled from a surfel, the corresponding coefficient ζ should be close to 1. Therefore, if ζ of a voxel is greater than a certain threshold, we will perform a RANSAC [5] plane fitting for the points that belong to the voxel. As a result, the fitted surfel parameters are regarded as variables and will be updated in the joint optimization later.

3.3.3 Trajectory Initialization. As described in Sect. 3.2.3, B-splines of $\mathbb{SO}(3)$ and $\mathbb{R}^3$ are utilized to represent the carrier's trajectories. In the initialization of the $\mathbb{R}(3)$ trajectory, we first transform the LiDAR positional attitude estimated by LOAM to IMU frame employing the coarse LiDAR-IMU extrinsics. Then this position sequence is used as the control points for the $\mathbb{R}(3)$ trajectory fitting. As for the $\mathbb{SO}(3)$ trajectory, we straightforwardly fit it with the IMU gyroscope readings as its control points.

3.4 Joint Optimization

To refine the coarse results obtained from the initialization stage, we consider fusing all the measurements into a unified CT-SLAM framework via MAP (maximum a posteriori) estimation and jointly optimizing all the involved variables.

3.4.1 Problem Formulation. We first define all the observations and the variables to be optimized in the joint optimization. Denote the LiDAR data, the image features, and the IMU measurements by $\mathcal{L}, \mathcal{V}$, and $\mathcal{I}$, respectively. The variables to be optimized in LVI-ExC are the $\mathbb{R}(3)$ and $\mathbb{SO}(3)$ B-splines as well as the gravity (denoting these trajectory-related variables by C), the visual map points ($\mathcal{M}$), the LiDAR surfels ($\mathcal{S}$), and the extrinsics and time offsets along with the IMU biases (denoting these sensor-related parameters by $\mathcal{T}$). The optimization objective of LVI-ExC is to find the maximum value of the joint probability of the variables to be estimated conditioning on all the observations, i.e.,

$$\{C, \mathcal{M}, \mathcal{S}, \mathcal{T}\}^* = \arg\max_{C,\mathcal{M},\mathcal{S},\mathcal{T}} p(C, \mathcal{M}, \mathcal{S}, \mathcal{T} \mid \mathcal{L}, \mathcal{V}, \mathcal{I}). \tag{17}$$

According to the Bayes law, $p(\cdot)$ can be reformulated as,

$$\begin{aligned} p(C, \mathcal{M}, \mathcal{S}, \mathcal{T} \mid \mathcal{L}, \mathcal{V}, \mathcal{I}) &= \frac{p(C, \mathcal{M}, \mathcal{S}, \mathcal{T})p(\mathcal{L}, \mathcal{V}, \mathcal{I} \mid C, \mathcal{M}, \mathcal{S}, \mathcal{T})}{p(\mathcal{L}, \mathcal{V}, \mathcal{I})} \\ &\propto p(C, \mathcal{M}, \mathcal{S}, \mathcal{T})p(\mathcal{L}, \mathcal{V}, \mathcal{I} \mid C, \mathcal{M}, \mathcal{S}, \mathcal{T}). \end{aligned} \tag{18}$$

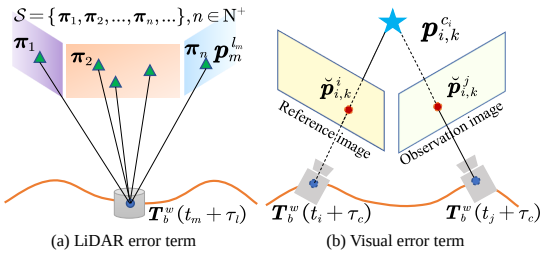
Since LiDAR observations are associated with the surfels, extrinsics and trajectories while IMU observations are only associated with the trajectories, the following equation should hold,

$$\begin{aligned} &p(\mathcal{L}, \mathcal{V}, \mathcal{I} \mid C, \mathcal{M}, \mathcal{S}, \mathcal{T}) = \\ &\qquad p(\mathcal{L} \mid C, \mathcal{S}, \mathcal{T})p(\mathcal{V} \mid C, \mathcal{M}, \mathcal{S}, \mathcal{T})p(\mathcal{I} \mid C). \end{aligned} \tag{19}$$

Further, via initialization, we already have reasonable estimates for the trajectories, visual points, surfels, and extrinsics. Therefore, they can be considered as independent prior terms in Eq. 18. Thus, using the results of Eq. 19, we can obtain,

$$\begin{aligned} &p(C, \mathcal{M}, \mathcal{S}, \mathcal{T} \mid \mathcal{L}, \mathcal{V}, \mathcal{I}) \propto p(C)p(\mathcal{M})p(\mathcal{S})p(\mathcal{T}) \cdot \\ &\qquad p(\mathcal{L} \mid C, \mathcal{S}, \mathcal{T})p(\mathcal{V} \mid C, \mathcal{M}, \mathcal{S}, \mathcal{T})(\mathcal{I} \mid C), \end{aligned} \tag{20}$$

where the first four terms are the prior terms, and the last three are the observed posteriors of the LiDAR, the camera and the IMU, respectively. In MAP estimation, these posterior terms can be modeled as the corresponding high-dimensional Gaussian distributions characterized by their means and variances. Thus, the optimization



(a) LiDAR error term (b) Visual error term

Figure 2: Illustrations of the LiDAR and visual error terms.

objective equates to minimizing the sum of the quadratic error terms, i.e.,

$$\{C, \mathcal{M}, \mathcal{S}, \mathcal{T}\}^* = \arg \min_{C,\mathcal{M},\mathcal{S},\mathcal{T}} {}^{L}\boldsymbol{e} + {}^{V}\boldsymbol{e} + {}^{I}\boldsymbol{e}, \tag{21}$$

where ${}^{L}\boldsymbol{e}$, ${}^{V}\boldsymbol{e}$, and ${}^{I}\boldsymbol{e}$ are the LiDAR, visual, and inertial error terms, respectively. As a result, to perform the joint optimization, we need to first construct the concrete forms of ${}^{L}\boldsymbol{e}$, ${}^{V}\boldsymbol{e}$, and ${}^{I}\boldsymbol{e}$.

3.4.2 LiDAR Error Term. A significant difference between a LiDAR scan and an image is the sparsity of the point cloud, which makes it difficult to directly establish frame-to-frame constraints among scans. Therefore, we consider building an environmental map by accumulating multi-frame point clouds and fitting surfels from the map to build point-to-surfel constraints. As introduced in Sect. 3.3.2, the scan poses are estimated by LOAM and the surfel set $\mathcal{S}$ is obtained via surfel fitting. According to the estimated scan poses, each laser point is firstly assigned to its associated surfel which has the nearest Euclidean distance and lies within a certain distance to the point. Assume that the surfel set $\mathcal{S}$ is established at time t_s and $\mathcal{S} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_n, \ldots\}$ where $n \in N^+$ is the index of a surfel and $\boldsymbol{\pi}_n \in \mathbb{R}^4$ is the parameter vector of the associated surfel. Thus, for a laser point $\boldsymbol{p}_m^{l_m}$ measured at time t_m and its associated surfel $\boldsymbol{\pi}_n$, the corresponding error term ${}^{L}\boldsymbol{e}_{m,n}$ is established as,

$$\mathring{\boldsymbol{p}}_m^{b_s} = \boldsymbol{T}_b^{wT}(t_s + \tau_l)\boldsymbol{T}_b^{w}(t_m + \tau_l)\boldsymbol{T}_l^{b}\mathring{\boldsymbol{p}}_m^{l_m}, \tag{22}$$

$${}^{L}\boldsymbol{e}_{m,n} = \boldsymbol{\pi}_n \cdot \mathring{\boldsymbol{p}}_m^{b_s} / \|\check{\boldsymbol{\pi}}_n\|_2, \tag{23}$$

in which $\mathring{(\cdot)}$ is the homogeneous coordinate of the associated point, $\boldsymbol{T}_b^w(\cdot)$ returns the pose of $\mathscr{F}_b$ in $\mathscr{F}_w$ at the given timestamp, τ_l is the LiDAR-to-IMU time difference, $\check{\boldsymbol{\pi}}_n$ is the vector with the first three elements of $\boldsymbol{\pi}_n$, and $\|\cdot\|_2$ is the l-2 norm of the operated vector. An illustration of this error term is given in Fig. 2 (a).

3.4.3 Visual Error Terms. **Reprojection Error Term.** For ease of management, we assign a reference frame to each visual feature point (the frame in which the point is first time observed). When the point is observed in subsequent frames, a visual constraint can be established between the observed frame and the reference frame with respect to the point by visual feature association. Specifically, when the k-th 2D visual feature point $\check{\boldsymbol{p}}_{i,k}^{i}$ of the i-th frame is observed in the j-th frame, its corresponding 3D spatial point $\boldsymbol{p}_{i,k}^{c_i}$ is first obtained by triangulation. Due to the large number of feature points, to reduce the computational complexity, we use a pair of the 2D point and its inverse depth, $(\check{\boldsymbol{p}}_{i,k}^{i}, d_{i,k})$, to identify each 3D visual point as Patrob-Perez *et al.* suggested [29]. Furthermore, according to the epipolar geometry, the reprojection error ${}^{V}\boldsymbol{e}_{i,j,k}$ between the i-th and the j-th frame with respect to $(\check{\boldsymbol{p}}_{i,k}^{i}, d_{i,k})$ is defined as,

$${}^{V}\boldsymbol{e}_{i,j,k} = \check{\boldsymbol{p}}_{i,k}^{j} - \phi(\boldsymbol{T}_b^{c}\boldsymbol{T}_b^{wT}(t_j + \tau_c)\boldsymbol{T}_b^{w}(t_i + \tau_c)(\boldsymbol{R}_c^{b}\mathring{\boldsymbol{p}}_{i,k}^{c_i}), \boldsymbol{K}), \tag{24}$$

where $\boldsymbol{K}$ is the camera intrinsic matrix which is regarded known, τ_c is the time difference between the camera and the IMU, ϕ represents the projection of the associated 3D spatial visual point to the 2D feature, and the inverse operator of ϕ conforms to,

$$\boldsymbol{p}_{i,k}^{c_i} = \phi^{-1}(\check{\boldsymbol{p}}_{i,k}^{i}, d_{i,k}, \boldsymbol{K}), \tag{25}$$

which lifts a 2D feature on an image to the 3D space. An intuitive example of this visual error term is provided in Fig. 2 (b).

Visual-Point to LiDAR-Surfel Error Term. Considering that in real scenarios, visual feature points may appear on spatial surfaces, such as paintings on a wall or text on a blackboard, we also append the constraints among the spatial visual points and the fitted surfels to the optimization problem. Starting from the initialization results, we first associate the re-scaled visual points and laser surfels by the Euclidean distances among them. Then, for a pair of a visual point and a LiDAR surfel $((\boldsymbol{p}_{i,k}^{i}, d_{i,k}), \boldsymbol{\pi}_n)$, the corresponding loss term ${}^{V}\boldsymbol{e}_{n,i,k}$ is defined as,

$$\mathring{\boldsymbol{p}}_{i,k}^{b_s} = \boldsymbol{T}_b^{wT}(t_s + \tau_l)\boldsymbol{T}_b^{w}(t_i + \tau_c)\boldsymbol{T}_l^{b}\mathring{\boldsymbol{p}}_{i,k}^{c_i}, \tag{26}$$

$${}^{V}\boldsymbol{e}_{n,i,k} = \boldsymbol{\pi}_n \cdot \mathring{\boldsymbol{p}}_{i,k}^{b_s} / \|\check{\boldsymbol{\pi}}_n\|_2, \tag{27}$$

where $\boldsymbol{p}_{i,k}^{c_i}$ is obtained by Eq. 25. With this term, the results yielded by the optimization are expected to be more geometric consistent.

3.4.4 IMU Error Terms. The constraints among the IMU measurements and the trajectories are straightforward. Assuming that the rotation and translation at the timestamp t on the $\mathbb{SO}(3)$ and $\mathbb{R}^3$ trajectories are $\boldsymbol{R}_b^w(t)$ and $\boldsymbol{p}_b^w(t)$, respectively, the associated estimates of the IMU measurements that can be extrapolated from the trajectories are,

$$\boldsymbol{a}^b(t) = \boldsymbol{R}_b^{wT}(t)(\ddot{\boldsymbol{p}}_b^w(t) - \boldsymbol{g}^w), \boldsymbol{\omega}^b(t) = \boldsymbol{R}_b^{wT}(t)(\dot{\boldsymbol{R}}_b^w(t)), \tag{28}$$

where the operator $\ddot{(\cdot)}$ ($\dot{(\cdot)}$) returns the associated second-order (first-order) derivative. Thus, the loss term between the IMU reading $\tilde{\boldsymbol{a}}_t$ ($\tilde{\boldsymbol{\omega}}_t$) at time t and the $\mathbb{R}^3$ ($\mathbb{SO}(3)$) trajectory can be defined as,

$${}^{I}\boldsymbol{e}_a = \tilde{\boldsymbol{a}}_t - \boldsymbol{a}^b(t) - {}^{a}\boldsymbol{b}, {}^{I}\boldsymbol{e}_\omega = \tilde{\boldsymbol{\omega}}_t - \boldsymbol{\omega}^b(t) - {}^{\omega}\boldsymbol{b}. \tag{29}$$

3.4.5 Bundle Adjustment. After the sensor-to-sensor and sensor-to-trajectory constraints are established, starting from the estimated initial values,we resort to the Levenberg-Marquardt algorithm [17, 22] to jointly optimize all the variables. More details about the bundle adjustment are provided in the supplementary material.

4 EXPERIMENT

4.1 Setup

To verify the effectiveness of LVI-ExC, a handheld data acquisition device was developed as shown in Fig. 1 (a). The device consists of a 16-beam ROBOSENSE LiDAR, a ZED-2 stereo camera, and a built-in 6-axis IMU which can report readings at 400 Hz. The three sensors are rigidly connected during all the experiments. With the handheld device, several sequences of experimental data were collected from various natural scenes. Each data sequence was about one minute

Table 1: Implicit errors. "Acc." and "Gyr." are the abbreviations of accelerator and gyroscope respectively.

Error Type	Number	Axis	Mean	Variance	Unit
Visual error	6223	u	-0.1495	0.1477	pixel
		v	-0.0099	0.0004	
Acc. error	18007	x	0.0157	8.9350*1e-6	m/s^2
		y	0.008	1.5099*1e-6	
		z	0.0146	1.7076*1e-6	
Gyr. error	18007	x	0.0033	4.7624*1e-7	rad/s
		y	0.0034	1.8057*1e-6	
		z	0.0054	8.5605*1e-6	
LiDAR error	24171		0.0178	8.2177*1e-5	m

long and contained LiDAR point clouds, 1280×720 grayscale images, and IMU readings with accelerations and angular velocities. During each sequence acquisition, the handheld device was fully panned and rotated in all directions to ensure the observabilities in all axes.

4.2 Implicit Error Analysis

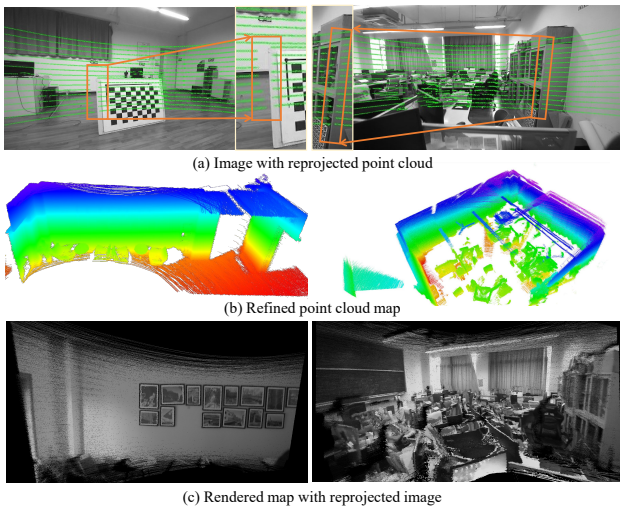
To evaluate LVI-ExC's inherent accuracy, we analyse the implicit errors which consists of, 1) visual reprojection errors, 2) the errors between the IMU measurements (acceleration and angular velocity) and those derived from the trajectories, and 3) the distance error between the LiDAR point and its associated surfel. Since these implicit errors are also the loss terms to be optimized in LVI-ExC, if these errors are small enough, it implies that the constraints among the LiDAR, the camera, and the IMU are well met.

All the errors involved in LVI-ExC after the convergence of the joint optimization are listed in Table 1. As seen, in terms of visual errors, the reprojection errors reach a sub-pixel level in both the horizontal and vertical directions, indicating that the spatial structure of the visual points is well recovered with the optimized inverse depths, the extrinsic parameters, and the trajectories. In terms of IMU errors, the differences between the measurements of both linear acceleration and angular velocity and those derived from the optimized trajectories are very small, which also verifies that the optimized trajectories are in good agreement with the actual measurements. In addition, the average error between the laser points and the fitted surfels is less than $2cm$, which is comparable to the measurement error of the original laser scan and also implies the high consistency between the trajectory and the built point cloud map with the optimized results.

4.3 Reprojection Result

In order to have an intuitive understanding of the accuracy of LVI-ExC's results, we perform three types of evaluation here.

Image with reprojected point cloud. With the calibration results, a LiDAR point cloud was reprojected into the camera image which had the same timestamp as the scan. Two scan-image pairs after reprojection are shown in Fig. 3 (a). It can be seen that the distribution of the point cloud at the corner of the cabinet and its distribution at the edge of the calibration board accord with the real scenes well.



(a) Image with reprojected point cloud

(b) Refined point cloud map

(c) Rendered map with reprojected image

Figure 3: Reprojection results and the constructed maps.

Refined point cloud map. After optimization, we first obtained the pose of each laser point when it was observed from the estimated B-spline trajectory according to its timestamp and then registered it in the global map in turn. Fig. 3 (b) shows the reconstructed point cloud maps of the calibration sites (the maps are gradually colored from red to blue according to the z values of the points). As shown, the objects such as tables, chairs, fans, and so on in the optimized maps are distinguishable, implying high-precision trajectories as well as refined maps are obtained.

Rendered map with reprojected image. At last, we took an image from a data sequence, retrieved its corresponding pose from the optimized trajectory according to its timestamp, reprojected its pixels into the point cloud map according to the estimated extrinsics and put color on the corresponding map points meanwhile. The rendered resulting maps are shown in Fig. 3 (c). As shown, the point cloud maps with reprojected images are highly consistent with the actual scenes, which once again verifies that the extrinsics calibrated by LVI-ExC have pleasing accuracy.

4.4 Comparison with State-of-the-art Methods

Table 2: Time cost to obtain LiDAR-visual-inertial extrinsics.

	Chained calibration		LVI-ExC
	Kalibr [8, 23, 27]	Autoware [12, 13]	
Time cost	94.62s	about 10min	208.13s
Total	about 694s		208.13s

In this subsection, we compare LVI-ExC with two other well-known calibration schemes. Since there are no target-free approaches in existing literature for integrated LiDAR-visual-inertial calibration, we choose two state-of-the-art target-based calibration methods, Kalibr [8, 23, 27] and Autoware [12, 13] for comparison, although it is a little unfair to LVI-ExC. Among the existing visual-inertial calibration schemes, Kalibr is well-known in academia and

Table 3: Calibration errors. Top three rows are the results of LVI-ExC and its counterparts. The bottom one is the result of LVI-ExC with random initialization (LVI-ExC$_{RI}$). The lowest error (runner-up) in a column is marked in blue (violet).

	$x(m)$	$y(m)$	$z(m)$	roll(°)	pitch(°)	yaw(°)
Kalibr [8, 23, 27]	0.0083±2.15e-5	-0.0018±4.94e-5	0.0030±1.37e-5	-0.2389±0.28	0.9392±0.47	0.0089±0.09
Autoware [12, 13]	0.0364±0.01	0.0188±4.8e-3	-0.0170±6.81e-3	-0.2820±7.37	0.0821±1.18	0.0224±0.37
LVI-ExC	0.0086±1.87e-4	0.0172±3.58e-3	-0.0147±6.32e-4	-0.0442±4.58e-2	-0.1784±0.20	-0.6129±0.13
LVI-ExC$_{RI}$	0.4615±0.1698	0.1365±0.2098	0.2767±0.3854	2.1938±25.0639	20.5690±973.6595	-23.6859±1125.9607

industry, which integrated many advanced techniques of camera-IMU calibration. As for the LiDAR-camera evaluation, we compare LVI-ExC with the calibration toolkit of a widely used autonomous driving software, Autoware.

4.4.1 Time cost. To obtain the extrinsics among LiDAR, camera and IMU, the two counterparts have to resort to a chained calibration. In other words, to infer LiDAR-IMU extrinsics, it is required to first calibrate camera-IMU by Kalibr and then calibrate camera-LiDAR by Autoware. To evaluate LVI-ExC's efficiency, we compared its average time cost with that of the chained calibration. Specifically, we conducted nine tests on LVI-ExC and the chained calibration respectively, and recorded their time costs in Table 2. As presented, the chained calibration takes three times as long as LVI-ExC. Although Kalibr performs camera-IMU calibration with a high efficiency, Autoware estimates the camera-LiDAR extrinsics with a lot of manual assistance, which significantly restricts its efficiency. By contrast, LVI-ExC only takes natural data as input and can automatically solve the extrinsics, thus having a much higher efficiency.

4.4.2 Calibration accuracy. Since it is difficult to directly obtain the ground truth of the LiDAR-visual-inertial extrinsics, while the extrinsics of the ZED-2 stereo cameras are factory-calibrated, we took it as the ground truth and conducted a chained calibration to evaluate LVI-ExC. Specifically, when evaluating LVI-ExC's calibration results, we obtained the extrinsics of the right camera frame relative to the left camera frame by calibrating the "LiDAR, left camera, IMU" and the "LiDAR, right camera, IMU" suites, respectively. Likewise, similar chained calibrations were performed to obtain the extrinsics between the left and right cameras when performing the calibrations of Kalibr and Autoware. For the evaluation of each method, nine independent experiments were conducted and the mean and variance of the translation errors and rotation errors (in Euler angle) of the results were listed in Table 3. As shown, the calibration accuracy of our LVI-ExC is comparable to that of the checkerboard-based methods, **despite the fact that it is free of any auxiliary calibration objects**. In terms of the translation errors, LVI-ExC and Kalibr achieve similar results, both being below $2cm$, while Autoware produces a larger deviation. In terms of the rotation error, the three schemes yield comparable results, all below 1° and with high stability. It should be noted that among the sensors available on today's market, the range error of a LiDAR is mostly $2 \sim 3cm$, while the highest accuracy of the monocular visual-SLAM is $6cm$ even after scale recovery [25]. Therefore, we believe that the accuracy of LVI-ExC is sufficient for most of the perception, mapping or localization applications.

4.5 Ablation Study on Initialization

Due to the large number of variables to be estimated in the joint optimization, intuitively, a successful initialization eases the estimation of initial values for the extrinsics, the gravity, the visual points, and the control points, which in turn helps LVI-ExC to find the optimal values smoothly. To verify the effectiveness of the proposed initialization scheme, we obtained the chained calibration results using LVI-ExC with the proposed initialization and LVI-ExC$_{RI}$ with random initialization on the same data sequences of the left and right cameras. The results are listed in the third and fourth rows of Table 3. As seen, the calibration results are with unacceptable errors for both translation and rotation when no reasonable initialization is performed. By contrast, our proposed initialization scheme helps LVI-ExC to find the optimal solution more easily and the converged results are much closer to the ground truth.

5 CONCLUSION

In this article, we present LVI-ExC, an integrated framework for LiDAR-visual-inertial extrinsic calibration without using any artificial auxiliary calibrators, which gets rid of the shortcomings like tedious operations, large accumulated errors, and poor geometric consistency of existing schemes. LVI-ExC formulates the LiDAR-visual-inertial extrinsic calibration as a CT-SLAM problem and estimates the trajectories and maps while estimating the sensor-to-sensor extrinsics. To make the joint optimization with ultra-high dimensional variables easy to be carried out, we also propose a reasonable initialization scheme for LVI-ExC with complete estimates of all the associated variables. The effectiveness of LVI-ExC is fully verified via comprehensive experimental analysis and comparisons with competing counterparts. In future work, we will devote our efforts to designing an integrated framework for the simultaneous calibration of multiple LiDARs, cameras, and IMUs.

6 ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grants 61972285 and 61973235, in part by the Natural Science Foundation of Shanghai under Grant 19ZR1461300, in part by the Shanghai Science and Technology Innovation Plan under Grant 20510760400, in part by the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission under Grant 21SG23, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100, and in part by the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Michael Bosse and Robert Zlot. 2009. Continuous 3D scan-matching with a spinning 2D laser. In *Proceedings of IEEE International Conference on Robotics and Automation*. Kobe, Japan, 4312–4319.

[2] Shoubin Chen, Jingbin Liu, Xinlian Liang, Shuming Zhang, Juha Hyyppä, and Ruizhi Chen. 2020. A novel calibration method between a camera and a 3D LiDAR with infrared images. In *Proceedings of IEEE International Conference on Robotics and Automation*. Paris, France, 4963–4969.

[3] Konstantinos Daniilidis. 1999. Hand-eye calibration using dual quaternions. *International Journal of Robotics Research* 18, 3 (1999), 286–298.

[4] Carl de Boor. 1978. *A Practical Guide to Spline*. Vol. 27. Springer, New York, NY, USA.

[5] Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *ACM Communications* 24, 6 (1981), 381–395.

[6] Michael Fleps, Elmar Mair, Oliver Ruepp, Michael Suppa, and Darius Burschka. 2011. Optimization based IMU camera calibration. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*. San Francisco, CA, USA, 3297–3304.

[7] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. 2017. On-manifold preintegration for real-time visual-inertial odometry. *IEEE Transactions on Robotics* 33, 1 (2017), 1–21.

[8] Paul Furgale, Timothy D. Barfoot, and Gabe Sibley. 2012. Continuous-time batch estimation using temporal basis functions. In *Proceedings of IEEE International Conference on Robotics and Automation*. Saint Paul, MN, USA, 2088–2095.

[9] A Geiger, P Lenz, C Stiller, and R Urtasun. 2013. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research* 32, 11 (Aug. 2013), 1231–1237.

[10] Carlos Guindel, Jorge Beltrán, David Martín, and Fernando García. 2017. Automatic extrinsic calibration for lidar-stereo vehicle sensor setups. In *Proceedings of IEEE International Conference on Intelligent Transportation Systems*. Yokohama, Japan, 1–6.

[11] Ryoichi Ishikawa, Takeshi Oishi, and Katsushi Ikeuchi. 2018. LiDAR and camera calibration using motions estimated by sensor fusion odometry. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*. Madrid, Spain, 7342–7349.

[12] Shinpei Kato, Eijiro Takeuchi, Yoshio Ishiguro, Yoshiki Ninomiya, Kazuya Takeda, and Tsuyoshi Hamada. 2015. An open approach to autonomous vehicles. *IEEE Micro* 35, 6 (Nov. 2015), 60–68.

[13] Shinpei Kato, Shota Tokunaga, Yuya Maruyama, Seiya Maeda, Manato Hirabayashi, Yuki Kitsukawa, Abraham Monrroy, Tomohito Ando, Yusuke Fujii, and Takuya Azumi. 2018. Autoware on board: Enabling autonomous vehicles with embedded systems. In *ACM/IEEE 9th International Conference on Cyber-Physical System*. Porto, Portugal, 287–296.

[14] Jonathan Kelly and Gaurav S. Sukhatme. 2009. Visual-inertial simultaneous localization, mapping and sensor-to-sensor self-calibration. In *Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation*. Daejeon, Korea (South), 360–368.

[15] Myoung-Jun Kim, Myung-Soo Kim, and Sung Yong Shin. 1995. A general construction scheme for unit quaternion curves with simple high order derivatives. In *Proceedings of 22nd Annual Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA.

[16] Cedric Le Gentil, Teresa Vidal-Calleja, and Shoudong Huang. 2018. 3D LiDAR-IMU calibration based on upsampled preintegrated measurements for motion distortion correction. In *Proceedings of IEEE International Conference on Robotics and Automation*. Brisbane, QLD, Australia, 2149–2155.

[17] Kenneth Levenberg. 1944. A method for the solution of certain problems in least square. *Quarterly Applied Mathematics* 2, 2 (1944), 164–168.

[18] Li Li, Zhu Li, Shan Liu, and Houqiang Li. Early Access, 2021. Motion estimation and coding structure for inter-prediction of LiDAR point cloud geometry. *IEEE Transactions on Multimedia* (Early Access, 2021).

[19] Jiarong Lin, Chunran Zheng, Wei Xu, and Fu Zhang. 2021. R^2LIVE: A robust, real-Time, LiDAR-inertial-visual tightly-coupled state estimator and mapping. *IEEE Robotics and Automation Letters* 6, 4 (2021), 7469–7476.

[20] Bruce D. Lucas and Takeo Kanade. 1981. "An iterative image registration technique with an application to stereo vision. In *Proceedings of International Joint Conference on Artificial Intelligence*. Vancouver, BC, Canada, 674–679.

[21] Jiajun Lv, Jinhong Xu, Kewei Hu, Yong Liu, and Xingxing Zuo. 2020. Targetless calibration of LiDAR-IMU system based on continuous-time batch estimation. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*. Las Vegas, NV, USA, 9968–9975.

[22] Donald W. Marquardt. 1963. An algorithm for least-squares estimation of nonlinear parameter. *Society for Industrial and Applied Mathematics* 11, 2 (1963), 431–441.

[23] Jérôme Maye, Paul Furgale, and Roland Siegwart. 2013. Self-supervised calibration for robotic systems. In *Proceedings of IEEE Intelligent Vehicles Symposium*. Gold Coast, QLD, Australia, 473–480.

[24] F.M. Mirzaei and S.I. Roumeliotis. 2008. A Kalman filter-based algorithm for IMU-camera calibration: Observability analysis and performance evaluation. *IEEE Transactions on Robotics* 24, 5 (Oct. 2008), 1143–1156.

[25] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. 2015. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics* 31, 5 (2015), 1147–1163.

[26] Raúl Mur-Artal and Juan D. Tardós. 2017. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics* 33, 5 (2017), 1255–1262.

[27] Luc Oth, Paul Furgale, Laurent Kneip, and Roland Siegwart. 2013. Rolling Shutter Camera Calibration. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Portland, ON, USA, 1360–1367.

[28] Chanoh Park, Peyman Moghadam, Soohwan Kim, Sridha Sridharan, and Clinton Fookes. 2020. Spatiotemporal camera-LiDAR calibration: A targetless and structureless approach. *IEEE Robotics and Automation Letters* 5, 2 (2020), 1556–1563.

[29] Alonso Patron-Perez, Steven Lovegrove, and Gabe Sibley. 2015. A spline-based trajectory representation for sensor fusion and rolling shutter cameras. *International Journal on Computer Vision* 113, 3 (Feb. 2015), 208–219.

[30] Zoltan Pusztai and Levente Hajder. 2017. Accurate calibration of LiDAR-camera systems using ordinary boxes. In *Proceedings of IEEE/CVF International Conference on Computer Vision Workshop*. Venice, Italy, 394–402.

[31] Zachary Taylor and Juan Nieto. 2016. Motion-Based calibration of multimodal sensor extrinsics and timing offset estimation. *IEEE Transactions on Robotics* 32, 5 (2016), 1215–1229.

[32] Wei Xu and Fu Zhang. 2021. FAST-LIO: A fast, robust LiDAR-inertial odometry package by tightly-coupled iterated Kalman filter. *IEEE Robotics and Automation Letters* 6, 2 (2021), 3317–3324.

[33] Ji Zhang and Sanjiv Singh. 2014. LOAM: LiDAR odometry and mapping in real-time. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*. California, USA.

[34] Lipu Zhou, Zimo Li, and Michael Kaess. 2018. Automatic extrinsic calibration of a camera and a 3D LiDAR using line and plane correspondences. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*. Madrid, Spain, 5562–5569.