2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10609982

MAVIS: Multi-Camera Augmented Visual-Inertial SLAM using $\mathrm{SE}_2(3)$ Based Exact IMU Pre-integration

Yifu Wang*, Yonhon Ng*, Inkyu Sa, Álvaro Parra, Cristian Rodriguez-Opazo, Taojun Lin†, Hongdong Li†

***Abstract*— We present a novel optimization-based Visual-Inertial SLAM system designed for multiple partially overlapped camera systems, named MAVIS. Our framework fully exploits the benefits of wide field-of-view from multi-camera systems, and the metric scale measurements provided by an inertial measurement unit (IMU). We introduce an improved IMU pre-integration formulation based on the exponential function of an automorphism of $\mathbf{SE_2(3)}$, which can effectively enhance tracking performance under fast rotational motion and extended integration time. Furthermore, we extend conventional front-end tracking and back-end optimization module designed for monocular or stereo setup towards multi-camera systems, and introduce implementation details that contribute to the performance of our system in challenging scenarios. The practical validity of our approach is supported by our experiments on public datasets. Our MAVIS won the first place in all the vision-IMU tracks (single and multi-session SLAM) on Hilti SLAM Challenge 2023 with 1.7 times the score compared to the second place[1].**

I. Introduction

Robust and real-time Simultaneous Localization And Mapping (SLAM) is a long-standing problem within the computer vision and robotics communities. Pure vision-based solutions often lack the robustness and accuracy present in lidar-based solutions. Consequently, they are frequently enhanced with additional sensors, such as a low-cost IMU measuring angular velocity and acceleration, particularly on XR (VR/AR) virtual and augmented reality devices. While existing monocular or stereo visual-inertial solutions [1]–[9] have demonstrated their potential to enhance robustness in degenerate scenarios such as textureless environments or agile motion by integrating IMU measurements, there are still existing challenges such as limited camera field-of-view, and a restricted ability to handle feature tracking failures for long durations. These challenges can lead to rapid system divergence, even when using IMUs, causing a significant degradation in positioning accuracy.

The present paper focuses on yet another type of sensor systems, namely multi-camera systems. As shown in Figure 1, the forward-facing stereo cameras offer a broader co-visibility area compared to the left, right, and upwards cameras, which have limited overlap with the forward-facing stereo pair. Such systems offer the advantage of a larger fields-of-view, omni-directional observation of the environment that improves motion estimation accuracy and robustness against failures due to texture-poor environments.

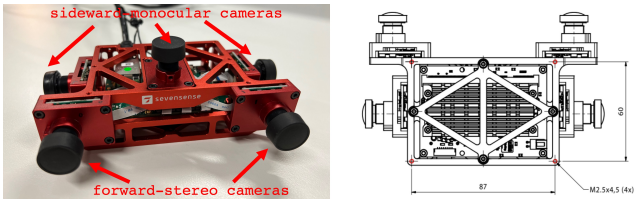

Fig. 1: AlphaSense multi-camera module as an example of multi-camera system analyzed in this paper, with a forward-facing stereo cameras and multiple sideward monocular cameras.

However, an inherent limitation of such setups is that the introduction of additional cameras directly leads to an increase in computational cost. The proper handling of measurements from all cameras is crucial for balancing accuracy, robustness, and computational efficiency.

Moreover, in order to improve the computational efficiency of optimization-based visual-inertial navigation methods without compromising accuracy, [10] has introduced an IMU pre-integration method, which can combine hundreds of inertial measurements into a single relative motion constraint by pre-integrating measurements between selected keyframes. This formulation plays a vital role in enhancing the effectiveness of front-end feature tracking across cameras and overall performance. However, existing methods [10], [11] rely on imprecise integration of the position and velocity where the IMU is assumed to be non-rotating between IMU measurements, such approximation can negatively impact the accuracy of pre-integrated poses, especially for fast rotational motion and extended integration time.

Our contributions are as follows:

- We present MAVIS, the state-of-the-art optimization-based visual-inertial SLAM framework specifically designed for multiple partially overlapped camera system.
- We introduce a new IMU pre-integration formulation based on the exponential function of an automorphism of $\mathbf{SE_2(3)}$. Our approach ensures highly accurate integration of IMU data, which directly contributes to the improved tracking performance of our SLAM system.
- We demonstrate a substantial advantage of MAVIS in terms of robustness and accuracy through an extensive experimental evaluation. Our method attains the first place in both the vision-only single-session and multi-session tracks of the Hilti SLAM Challenge 2023.

II. Related Work

The advantages and challenges of monocular or stereo visual-inertial SLAM have been discussed extensively in previous frameworks [1]–[9]. For a comprehensive survey,

* equal contribution, XR Vision Labs, Tencent.
† ANU, work completed while with Tencent.
[1] https://hilti-challenge.com/leader-board-2023.html

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:57:13 UTC from IEEE Xplore. Restrictions apply.

please refer to [12] and the latest research [13]. Here, we mainly focus on vision-based solutions for multi-camera systems. [14] extended ORB-SLAM2 [15] to multi-camera setups, supporting various rigidly coupled multi-camera systems. [16] introduced an adaptive SLAM system design for arbitrary multi-camera setups, requiring no sensor-specific tuning. Several works [17]–[21] focus on utilizing a surround-view camera system, often with multiple non-overlapping monocular cameras, or specializing in motion estimation for ground vehicles. While demonstrating advantages in robustness in complex environments, these methods exhibit limited performance in highly dynamic scenarios and minor accuracy improvements in real-world experiments.

While many multi-camera visual-inertial solutions have been presented, none achieve a perfect balance among accuracy, robustness, and computational efficiency, especially in challenging scenarios. VILENS-MC [22] presents a multi-camera visual-inertial odometry system based on factor graph optimization. It improves tracking efficiency through cross-camera feature selection. However, it lacks a local map tracking module and loop closure optimization, leading to reduced performance in revisited locations compared to ORB-SLAM3 [9]. BAMF-SLAM [23] introduces a multi-fisheye VI-SLAM system that relies on dense pixel-wise correspondences in a tightly-coupled semi-pose-graph bundle adjustment. This approach delivers exceptional accuracy but demands a high-end GPU for near real-time performance.

A. Inertial Preintegration

The theory of IMU preintegration was first proposed by [24], [25]. This work involves the discrete integration of the inertial measurement dynamics in a local frame of reference, such that the bias of state dynamics can be efficiently corrected at each optimization step. [26] presents a singularity-free orientation representation on $\mathbf{SO(3)}$ manifold, incorporating the IMU preintegration into optimization-based VINS, significantly improving on the stability of [25]. Moreover, [1], [27] introduced preintegration in the continuous form using quaternions, in order to overcome the discretization effect and improve the accuracy. There are also several approaches solving this problem by using analytical solution [28], [29] or a switched linear system [30], [31]. Another work that is closely related to ours is introduced by [32]. It extended on-manifold pre-integration of [10] to the Lie group $\mathbf{SE_2(3)}$. However, their method is still limited to Euler integration for position and velocity where the orientation is assumed non-rotating between IMU measurements.

III. Methodology

In this work, we present our multi-camera VI-SLAM system with a novel automorphism of $\mathbf{SE_2(3)}$ exponential-based exact IMU pre-integration formulation, alongside an extended front-end tracking and back-end optimization modules for multi-camera setups. Figure 2 shows the main system components.

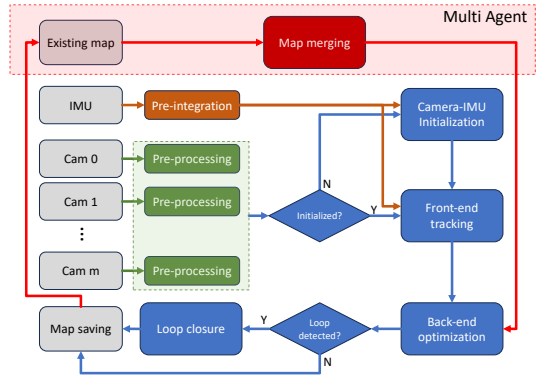

Fig. 2: Overview of our proposed visual-inertial localization and mapping pipeline for multi-camera systems.

A. IMU intrinsic compensation

We start by adding an IMU *intrinsic compensation* prior to the IMU pre-integration stage. When we use IMU information in a visual-inertial SLAM system, it is common to model raw accelerometer and gyroscope measurements as

$$\breve{\mathbf{a}}_t = \mathbf{a}_t + \mathbf{b}_{a_t} + \mathbf{n}_{a_t} \tag{1a}$$

$$\breve{\omega}_t = \omega_t + \mathbf{b}_{\omega_t} + \mathbf{n}_{\omega_t}, \tag{1b}$$

which only considers measurement noise density (i.e., $\mathbf{n}$) and bias $\mathbf{b}$. Affected by acceleration bias $\mathbf{b}_{a_t}$ and gyroscope bias $\mathbf{b}_{\omega_t}$, as well as additive noises $\mathbf{n}_a$ and $\mathbf{n}_\omega$, model (1) is simple and useful, resulting in a good approximation for devices with factory calibrated intrinsics. However, it may produce impaired calibration results for low-cost, consumer grade inertial sensors which exhibit significant axis misalignment and scale factor errors. Inspired by [33], we extend the IMU model (1) by introducing the IMU intrinsics modelling

$$\breve{\mathbf{a}}_t = \mathbf{S}_\alpha \mathbf{M}_\alpha \mathbf{a}_t + \mathbf{b}_{a_t} + \mathbf{n}_{a_t} \tag{2a}$$

$$\breve{\omega}_t = \mathbf{S}_\omega \mathbf{M}_\omega \omega'_t + \mathbf{A}_\omega \mathbf{a}_t + \mathbf{b}_{\omega_t} + \mathbf{n}_{\omega_t}, \tag{2b}$$

whereas $\omega'_t = \mathbf{C}_\omega \omega_t$ is the rotation matrix between accelerometer and gyroscope. We define $\mathbf{S}_\alpha$ and $\mathbf{S}_\omega$ as the diagonal matrix of scaling effects, $\mathbf{M}_\alpha$ and $\mathbf{M}_\omega$ are lower unitriangular matrix corresponding to misalignment small angles, and let $\mathbf{A}_\omega$ be a fully populated matrix. All characteristics above can be obtained by performing the widely-adopted Kalibr [34].

B. IMU pre-integration

Following the previous section, we compensate IMU axis misalignment and scale factor, and use the notation $\bar{\omega}_t$ and $\bar{a}_t$ to denote the biased, but noise/skew-free and scale-correct IMU measurements. In the following sections, we drop or retain the subscript t to simplify the notation, or highlight the time dependency of the variable.

We now proceed to our core contribution: a novel, exact IMU pre-integration formulation based on the exponential function of an automorphism of the matrix Lie group $\mathbf{SE_2(3)}$. We first model the noise-free kinematics of the system as follow:

$$\begin{cases} \dot{\mathbf{R}} = \mathbf{R}(\bar{\omega} - \mathbf{b}_\omega)^\wedge \\ \dot{\mathbf{p}} = \mathbf{v} \\ \dot{\mathbf{v}} = \mathbf{R}(\bar{\mathbf{a}} - \mathbf{b}_a) + \mathbf{g} \end{cases} \tag{3}$$

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:57:13 UTC from IEEE Xplore. Restrictions apply.

where $(\dot{\mathbf{R}}, \dot{\mathbf{p}}, \dot{\mathbf{v}})$ are the first order derivative of rotation, position and velocity of the IMU frame and $\mathbf{g}$ denotes the gravity vector, both expressed with respect to a world fixed frame. We also define the random walk of the biases $\dot{\mathbf{b}}_\omega = \tau_\omega$ and $\dot{\mathbf{b}}_a = \tau_a$. We can represent the IMU pose using the extended pose in $\mathbf{SE_2(3)}$ Lie group, such that the state is $\xi = \begin{bmatrix} \mathbf{R} & \mathbf{p} & \mathbf{v} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and the kinematics of the state ξ can be represented using an automorphism of $\mathbf{SE_2(3)}$ [35]

$$\dot{\xi} = (\mathbf{G} - \mathbf{D})\xi + \xi(\mathbf{U} - \mathbf{B} + \mathbf{D}) \tag{4}$$

where

$$\mathbf{U} = \begin{bmatrix} \omega^\wedge & 0 & \mathbf{a} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 0 & 0 & \mathbf{g} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} (\mathbf{b}_\omega)^\wedge & 0 & \mathbf{b}_a \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Let us assume t_{i-1} is the start time of pre-integration at previous image frame $\mathcal{F}_{i-1}$. We define t_j be the timestamp of an arbitrary IMU measurement between frames $\mathcal{F}_{i-1}$ and $\mathcal{F}_i$, and let $t_{j'}$ be the timestamp of its subsequent IMU measurement, such that the small integration time $\delta = t_{j'} - t_j$. Let ξ_{t_j} be the extended pose at time instant t_j. The exact integration given (4), and assuming $\mathbf{U}$ and $\mathbf{B}$ are constant within the integration time δ is

$$\xi_{t_{j'}} = \exp\left(\delta(\mathbf{G} - \mathbf{D})\right) \xi_{t_j} \exp(\delta(\mathbf{U} - \mathbf{B} + \mathbf{D})) \tag{5}$$

Assuming there are N sets of IMU measurements between time t_j and t_{i-1}, we have

$$\xi_{t_j} = \exp\left((\sum_{s=0}^{N} \delta_s)(\mathbf{G} - \mathbf{D}) \right) \xi_{t_{i-1}} \prod_{s=0}^{N} \exp(\delta_s(\mathbf{U}_s - \mathbf{B}_s + \mathbf{D})) \tag{6}$$

We define $\mathcal{T} = \sum_{s=0}^{N} \delta_s$ and then rearrange (6), we obtain

$$\xi_{t_{i-1}}^{-1} \exp\left(-\mathcal{T}(\mathbf{G} - \mathbf{D})\right) \xi_{t_j} = \prod_{s=0}^{N} \exp(\delta_s(\mathbf{U}_s - \mathbf{B}_s + \mathbf{D})). \tag{7}$$

The exponential on the left equals

$$\exp\left(-\mathcal{T}(\mathbf{G} - \mathbf{D})\right) = \begin{bmatrix} \mathbf{I} & \frac{1}{2}\mathcal{T}^2\mathbf{g} & -\mathcal{T}\mathbf{g} \\ 0 & 1 & 0 \\ 0 & \mathcal{T} & 1 \end{bmatrix}. \tag{8}$$

Substituting (8) into (7), the left hand side components of (7) is exactly the same as the middle of equation (33) in [10], the right hand side is our newly derived pre-integration terms, where each of the exponential can be expanded as

$$\exp(\delta_s(\mathbf{U}_s - \mathbf{B}_s + \mathbf{D})) = \begin{bmatrix} \exp(\delta_s(\omega_s - \mathbf{b}_{\omega_s})^\wedge) & \mathbf{J}_2(\mathbf{a}_s - \mathbf{b}_{a_s}) & \mathbf{J}_1(\mathbf{a}_s - \mathbf{b}_{a_s}) \\ 0 & 1 & 0 \\ 0 & \delta_s & 1 \end{bmatrix}, \tag{9}$$

where

$$\theta = \|\tilde{\omega}\| \tag{10}$$

$$\mathbf{J}_1 = \delta_s \mathbf{I} + \frac{1}{\theta^2}(1 - \cos(\delta_s\theta))\tilde{\omega}^\wedge + \frac{1}{\theta^3}(\delta_s\theta - \sin(\delta_s\theta))(\tilde{\omega}^\wedge)^2 \tag{11}$$

$$\mathbf{J}_2 = \frac{1}{2}{\delta_s}^2\mathbf{I} + \frac{1}{\theta^3}(\delta_s\theta - \sin(\delta_s\theta))\tilde{\omega}^\wedge + \frac{1}{\theta^4}(\frac{1}{2}{\delta_s}^2\theta^2 + \cos(\delta_s\theta) - 1)(\tilde{\omega}^\wedge)^2 \tag{12}$$

Here, we use $\tilde{\omega} = \omega_s - \mathbf{b}_{\omega_s}$. The iterative pre-integration is finally given as

$$\begin{cases} \Delta\mathbf{R}_{t_{j'}}^{t_i} = \Delta\mathbf{R}_{t_j}^{t_i} \exp(\delta_j(\omega_j - \mathbf{b}_{\omega_j})^\wedge) \\ \Delta\mathbf{p}_{t_{j'}}^{t_i} = \Delta\mathbf{p}_{t_j}^{t_i} + \delta_j \Delta\mathbf{v}_{t_j}^{t_i} + \Delta\mathbf{R}_{t_j}^{t_i}\mathbf{J}_2(\mathbf{a}_j - \mathbf{b}_{a_j}) \\ \Delta\mathbf{v}_{t_{j'}}^{t_i} = \Delta\mathbf{v}_{t_j}^{t_i} + \Delta\mathbf{R}_{t_j}^{t_i}\mathbf{J}_1(\mathbf{a}_j - \mathbf{b}_{a_j}) \end{cases} \tag{13}$$

Note that under the Euler integration scheme, Jacobian terms $\mathbf{J}_1$ and $\mathbf{J}_2$ are simplified to be $\mathbf{J}_1 = \delta_s\mathbf{I}$ and $\mathbf{J}_2 = \frac{1}{2}{\delta_s}^2\mathbf{I}$.

In practical implementation, the noise-free terms are not available, and are thus substituted by their corresponding estimate denoted with the $\hat{(\cdot)}$ notation. We then deal with the covariance propagation given the uncertainty of the previous estimate and measurement noise. We define the error terms as follows:

$$e_{ij'} = [(e_{ij'}^{\Delta\mathbf{R}})^\top, (e_{ij'}^{\Delta\mathbf{p}})^\top, (e_{ij'}^{\Delta\mathbf{v}})^\top, (e_{j'}^{\mathbf{b}_\omega})^\top, (e_{j'}^{\mathbf{b}_a})^\top]^\top \tag{14}$$

$$e_{ij'}^{\Delta\mathbf{R}} = \log(\Delta\mathbf{R}_{t_{j'}}^{t_i}{}^\top \Delta\hat{\mathbf{R}}_{t_{j'}}^{t_i})^\vee \sim \mathcal{N}(0, \Sigma_{ij'}^{\mathbf{R}}) \in \mathbb{R}^3 \tag{15a}$$

$$e_{ij'}^{\Delta\mathbf{p}} = \Delta\hat{\mathbf{p}}_{t_{j'}}^{t_i} - \Delta\mathbf{p}_{t_{j'}}^{t_i} \sim \mathcal{N}(0, \Sigma_{ij'}^{\mathbf{p}}) \in \mathbb{R}^3 \tag{15b}$$

$$e_{ij'}^{\Delta\mathbf{v}} = \Delta\hat{\mathbf{v}}_{t_{j'}}^{t_i} - \Delta\mathbf{v}_{t_{j'}}^{t_i} \sim \mathcal{N}(0, \Sigma_{ij'}^{\mathbf{v}}) \in \mathbb{R}^3 \tag{15c}$$

$$e_{j'}^{\mathbf{b}_\omega} = \hat{\mathbf{b}}_{\omega_j} - \mathbf{b}_{\omega_j} \sim \mathcal{N}(0, \Sigma_{ij'}^{\mathbf{b}_\omega}) \in \mathbb{R}^3 \tag{15d}$$

$$e_{j'}^{\mathbf{b}_a} = \hat{\mathbf{b}}_{a_j} - \mathbf{b}_{a_j} \sim \mathcal{N}(0, \Sigma_{ij'}^{\mathbf{b}_a}) \in \mathbb{R}^3 \tag{15e}$$

We derive the matrix representation of the pre-integration terms' evolution

$$e_{ij'} = \mathbf{A}_{j'} e_{ij} + \mathbf{B}_{j'} \mathbf{n}_{j'} \tag{16}$$

where

$$\mathbf{A} = \begin{bmatrix} \exp(-\delta_{j'}(\hat{\omega}_{j'} - \hat{\mathbf{b}}_{\omega_{j'}})^\wedge) & 0 & 0 & -\delta_{j'}\mathbf{I} & 0 \\ -\Delta\hat{\mathbf{R}}_{t_j}^{t_i}(\hat{\mathbf{J}}_2(\hat{\mathbf{a}}_{j'} - \hat{\mathbf{b}}_{a_{j'}}))^\wedge & \mathbf{I} & \delta_{j'}\mathbf{I} & \mathbf{D}_1 & -\Delta\hat{\mathbf{R}}_{t_j}^{t_i}\hat{\mathbf{J}}_2 \\ -\Delta\hat{\mathbf{R}}_{t_j}^{t_i}(\hat{\mathbf{J}}_1(\hat{\mathbf{a}}_{j'} - \hat{\mathbf{b}}_{a_{j'}}))^\wedge & 0 & \mathbf{I} & -\mathbf{D}_2 & -\Delta\hat{\mathbf{R}}_{t_j}^{t_i}\hat{\mathbf{J}}_1 \\ 0 & 0 & 0 & \mathbf{I} & 0 \\ 0 & 0 & 0 & 0 & \mathbf{I} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} \delta_{j'}\mathbf{I} & 0 & 0 & 0 \\ -\mathbf{D}_1 & \Delta\hat{\mathbf{R}}_{t_j}^{t_i}\hat{\mathbf{J}}_2 & 0 & 0 \\ \mathbf{D}_2 & \Delta\hat{\mathbf{R}}_{t_j}^{t_i}\hat{\mathbf{J}}_1 & 0 & 0 \\ 0 & 0 & -\delta_{j'}\mathbf{I} & 0 \\ 0 & 0 & 0 & -\delta_{j'}\mathbf{I} \end{bmatrix}, \mathbf{n}_{j'} = \begin{bmatrix} \mathbf{n}_{\omega_{j'}} \\ \mathbf{n}_{\mathbf{a}_{j'}} \\ \tau_{\omega_{j'}} \\ \tau_{a_{j'}} \end{bmatrix} \tag{17}$$

and

$$\begin{aligned} \mathbf{D}_1 = \Delta\hat{\mathbf{R}}_{t_j}^{t_i} \Bigg(& \frac{{\delta_{j'}}^2}{\|\hat{\omega}_{j'} - \hat{\mathbf{b}}_{\omega_{j'}}\|^2} \Big(((\hat{\omega}_{j'} - \hat{\mathbf{b}}_{\omega_{j'}})^\wedge (\hat{\mathbf{a}}_{j'} - \hat{\mathbf{b}}_{a_{j'}}))^\wedge \\ & + (\hat{\omega}_{j'} - \hat{\mathbf{b}}_{\omega_{j'}})^\wedge (\hat{\mathbf{a}}_{j'} - \hat{\mathbf{b}}_{a_{j'}})^\wedge \Big) \\ & + \frac{2{\delta_{j'}}^2}{\|\hat{\omega}_{j'} - \hat{\mathbf{b}}_{\omega_{j'}}\|^4} \Big(((\hat{\omega}_{j'} - \hat{\mathbf{b}}_{\omega_{j'}})^\wedge)^2 (\hat{\mathbf{a}} - \hat{\mathbf{b}}_{a_{j'}})(\hat{\omega}_{j'} - \hat{\mathbf{b}}_{\omega_{j'}})^\top \Big) \Bigg) \end{aligned} \tag{18}$$

$$\mathbf{D}_2 = \Delta\hat{\mathbf{R}}_{t_{j'}}^{t_i} \|\hat{\omega}_{j'} - \hat{\mathbf{b}}_{\omega_{j'}}\|^4 {\delta_{j'}}^2 (\hat{\omega}_{j'} - \hat{\mathbf{b}}_{\omega_{j'}})^\wedge (\hat{\mathbf{a}}_{j'} - \hat{\mathbf{b}}_{a_{j'}})(\hat{\omega}_{j'} - \hat{\mathbf{b}}_{\omega_{j'}})^\top \tag{19}$$

where $\hat{\omega}$ and $\hat{a}$ are the biased, noisy, and intrinsic compensated IMU measurement.

C. Front-end tracking

We aim to design a SLAM solution for sensor suites that are equipped with multiple partially overlapped cameras (cf. illustrated in Figure 1). Inspired by [22] and [9], we employ a localization strategy based on a local map, utilizing 2D extracted features and local 3D map points for precise pose estimation. Such local map matching mechanism can

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:57:13 UTC from IEEE Xplore. Restrictions apply.

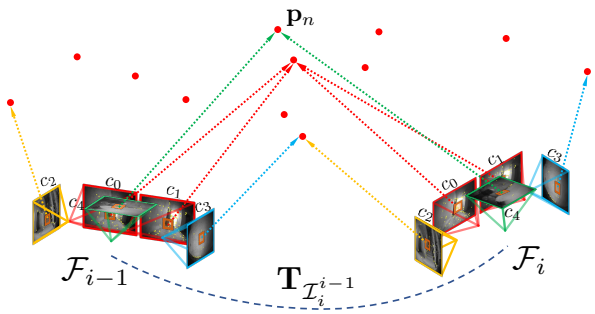

Fig. 3: Geometric relationships of local feature matching across multiple cameras.

significantly improve the tracking performance when the camera revisited a previous location. We define the body frame to be the same as the IMU frame. We first project all local map points onto the multi-camera image at the current time by using the predicted relative pose obtained from IMU pre-integration. As shown in Fig 3, feature matching is done for both intra-cameras and inter-cameras to enhance the co-visibility relationships. Given the multi-camera systems are precisely calibrated, the projected landmarks on an arbitrary camera can be formulated by:

$$(u_{c_k}^n, v_{c_k}^n) = \pi_{c_k}(\mathbf{T}_{bc_k}^{-1}\mathbf{T}_{\mathcal{I}_i^{i-1}}^{-1}\mathbf{T}_{i-1}^{-1}\mathbf{p}_n), \tag{20}$$

whereas $\mathbf{p}_n$ denotes the position of landmark n in world coordinate, and π_{c_k} be the projection function which turns $\mathbf{p}_n$ into a pixel location $(u_{c_k}^n, v_{c_k}^n)$ using intrinsic parameters of camera c_k. We define $\mathbf{T}_{bc_k}$ as the extrinsic parameters of camera c_k with respect to the IMU coordinates. Let $\mathbf{T}_{i-1}$ be the absolute pose of reference frame $\mathcal{F}_{i-1}$ in world coordinate and $\mathbf{T}_{\mathcal{I}_i^{i-1}}$ be the estimated relative pose between current frame $\mathcal{F}_i$ and reference frame using IMU pre-integration. For those 2D features that are not associated with landmarks, we perform feature matching between the current frame and keyframes, and create new local map points through triangulation. All these relationships are then used in the back-end optimization of MAVIS to augment co-visibility edges and improve the positioning accuracy. In addition, we utilize the distance between descriptors for further validation and employ a robust cost function in the back-end optimization to eliminate all incorrectly matched feature points.

D. Back-end optimization

Similar to many other visual-inertial SLAM/VIO systems based on optimization scheme [2], [9], [22], our MAVIS updates all the state vector by minimizing both visual reprojection errors based on observations from all cameras and error terms from pre-integrated IMU measurement using a sliding-window bundle adjustment scheme. We define the state vector $\mathcal{X}$ as the combination of motion states $\{\mathbf{x}\} = [\mathbf{x}_1 \cdots \mathbf{x}_v]$ and landmarks states $\{\mathbf{p}\} = [\mathbf{p}_0 \cdots \mathbf{p}_l]$, whereas $\mathbf{x}_i$ contains 6 DoF body poses $\mathbf{R}_i$ and $\mathbf{t}_i$, linear velocity $\mathbf{v}_i$, and IMU biases $\mathbf{b}_{\omega_i}$ and $\mathbf{b}_{a_i}$. $\mathbf{p}_n$ denotes the position of the landmark n. The visual-inertial BA can be formulated as:

$$\min_{\mathcal{X}} \sum_{i=1}^{v} \left(\|\mathbf{r}_{\mathbf{x}_i,\mathcal{I}_i^{i-1}}\|^2_{\Sigma_{\mathbf{x}_i,\mathcal{I}_i^{i-1}}} + \sum_{k=0}^{m} \sum_{n\in\mathcal{L}_i} \rho\|\mathbf{r}_{\mathbf{c}_k,\mathbf{x}_i,\mathbf{p}_n}\|_{\Sigma_{\mathbf{c}_k,\mathbf{x}_i,\mathbf{p}_n}} \right) \tag{21}$$

where

- $\mathcal{L}_i$ is the set of landmarks observed in keyframe $\mathcal{F}_i$.
- $\mathbf{r}_{\mathbf{x}_i,\mathcal{I}_i^{i-1}}$ is the residual for IMU, and $\mathbf{r}_{\mathbf{c}_k,\mathbf{x}_i,\mathbf{p}_n}$ is the residual for visual measurements of camera c_k.
- $\rho(\cdot)$ is the robust kernel used to eliminate outliers.

IV. Application to Multi-camera VI-SLAM

An overview of our system architecture is shown in Figure 2. After introducing IMU pre-integration, front-end tracking, and back-end optimization modules in Sec III, we now proceed to the implementation particulars in this section, which directly contribute to the overall precision and robustness of our SLAM system.

A. Visual measurement pre-processing

To address the challenges in real-world application scenarios and released datasets from [36], such as dark scenes, frame drops, and data discontinuities, we conduct data pre-processing prior to the feature extraction step. Specifically, we apply histogram equalization to compensate for dark frames. This technique significantly enhances both the quantity and distribution of the extracted feature points. Additionally, as multi-camera devices require increased bandwidth for transmitting image data, they are more likely to encounter frame drops and synchronization issues. We therefore leverage the remaining images in feature tracking to avoid integrating IMU data for long duration. We obtain the time delay between each camera and the IMU during the camera-IMU extrinsic parameter calibration process, and add a mid-exposure time compensation parameter to further reduce the impact of synchronization errors. By adopting these approaches, we achieve visible improvements in the performance of our SLAM solution.

B. Camera-IMU initialization

Inspired by [9], we employ a robust and accurate multi-camera IMU initialization method in our MAVIS. We firstly use stereo information from the first frame to generate an initial map. Once feature depth is known in the first keyframe, we project all landmarks in adjacent keyframes using the propagated motion model. Notably, all projections in subsequent stages are performed for both intra-camera and inter-camera scenarios. Please refer to Section III-C for detailed front-end tracking. Subsequently, we execute bundle adjustment for pure visual MAP estimation within 2 seconds while calculating IMU pre-integration and covariance between adjacent keyframes. Following the pure IMU optimization method, we jointly optimize map point positions, IMU parameters, and camera poses through multi-visual-inertial bundle adjustment. In order to avoid trajectory jumps in real-time outputs, we fix the last frame's pose during optimization. More advanced algorithms such as [39] can also be employed, but is not integrated into MAVIS.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:57:13 UTC from IEEE Xplore. Restrictions apply.

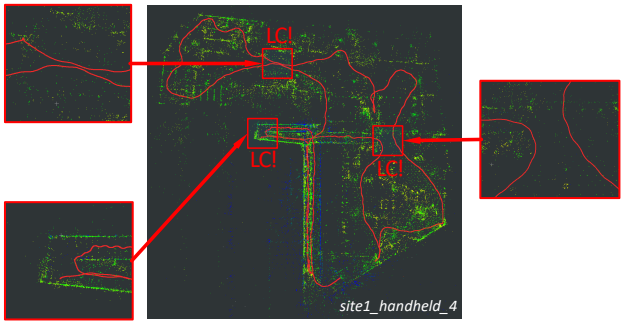

Fig. 4: Multiple detected closed loops from *site1_handheld_4* sequence in Hilti SLAM Challenge 2023. Accumulated drifts are significantly reduced, ensuring enhanced accuracy and robustness.

C. Loop closure

The loop-closure module utilizes the DBoW2 library [40] for candidate frame detection. To fully exploit the multi-camera system's wide field-of-view, we detect putative loop closures from both intra-cameras and inter-camera, which enables our MAVIS system to correctly detect loops—like u-turn motions—that regular monocular or stereo VIO systems fail to detect (cf. illustrated in Figure 4). Similar to [1], [7], [9], a geometric verification is performed to remove outlier loop closures. Upon detecting a correct closed loop, we execute a global bundle adjustment, optimizing the entire trajectory with all available information, substantially reducing drift across most sequences.

D. Multi-Agent SLAM

To participate in Hilti SLAM Challenges 2023 [41] multi-session, we adopt a similar approach for merging maps via our loop-closure correction module (refer to Figure 5). We employ the Bag-of-Words (BoW) model [40] to identify potential overlapping keyframes, and utilizing local maps to assist geometric alignment. By fusing multiple sub-maps and conducting successful verification checks, we finally generate a globally-consistent map. In the multi-session challenge context, we designate one single-session sequence as the base for the global map. Upon the completion of SLAM processing for a sequence, the map is locally saved, and pre-loaded prior to processing subsequent data sequences. With systematic processing of all sequences and map fusion, we can integrate all submaps into a global map.

V. Experiments

We evaluate the performance of our method on diverse datasets. We firstly compare it to most state-of-the-art visual-inertial approaches in stereo-inertial setup using the EuRoC datasets [38]. We furthermore evaluate our system under multi-camera setup on the Hilti SLAM Challenge 2023 Datasets [41], featuring challenging sequences from handheld devices and ground robots. Both qualitative and quantitative results highlight the effectiveness of our system. Our framework is implemented in C++ and evaluated on an Ubuntu 20.04 desktop, equipped with an AMD Ryzen 9 5950X 16-Core Processor.

A. Performance on EuRoC datasets

Our first experiments utilize the widely-used EuRoC datasets [38], featuring sequences captured by a drone flying inside the room, equipped with synchronized stereo cameras and an IMU. We benchmark our **MAVIS** against state-of-the-art methods, including VINS-MONO [1], OKVIS [2], SVOGTSAM [3], EqVIO [4], OpenVINS [6], VINS-FUSION [5], Kimera [7], BASALT [8], and ORB-SLAM3 [9]. We employ the EuRoC dataset's calibration results and exclude our IMU intrinsic compensation for a fair comparison. We quantitatively evaluate Absolute Trajectory Error (RMSE in meters) using EVO [42] and summarize the results in Table I. Best results are in **bold**, while "-" indicates a method's failure to complete the sequence.

As shown in the table above, our method outperforms in most sequences. Among 11 sequences in ATE evaluations, our approach achieves the best results in 6 of them, with only ORB-SLAM3 [9] and BASALT [8] approaching our method. We also provide the standard deviation (std) of RMSE for all sequences, for measuring the robustness of different algorithms on the same datasets. Our method again surpasses all alternatives with a 0.017 std error. To summarize, our stereo-inertial setup demonstrates state-of-the-art performance in terms of accuracy and robustness. This could be attributed to our improved IMU pre-integration formulation, which provides more precise motion modeling and robustness in handling rapid rotations and extended integration times.

B. Performance on Hilti SLAM Challenge 2023

To thoroughly evaluate the robustness and accuracy of our multi-camera VI-SLAM system in challenging conditions, we conducted experiments on the Hilti SLAM Challenge

		MH-01	MH-02	MH-03	MH-04	MH-05	V-101	V-102	V-103	V-201	V-202	V-203	Std.
Monocular Inertial	VINS-MONO [1]	0.070	0.050	0.080	0.120	0.090	0.040	0.060	0.110	0.060	0.060	0.090	0.025
	OKVIS [2]	0.160	0.220	0.240	0.340	0.470	0.090	0.200	0.240	0.130	0.160	0.290	0.106
	SVOGTSAM [3]	0.050	0.030	0.120	0.130	0.160	0.070	0.110	-	0.070	-	-	0.044
	EqVIO [4]	0.176	0.236	0.112	0.165	0.238	0.063	0.128	0.216	0.058	0.158	0.176	0.062
Stereo Inertial	OpenVINS [6]	0.183	0.129	0.170	0.172	0.212	0.055	0.044	0.069	0.058	0.045	0.147	0.063
	VINS-FUSION [5]	0.181	0.092	0.167	0.203	0.416	0.064	0.270	0.157	0.065	-	0.160	0.105
	Kimera [7]	0.080	0.090	0.110	0.150	0.240	0.050	0.110	0.120	0.070	0.100	0.190	0.055
	BASALT [8]	0.080	0.060	0.050	0.100	0.080	0.040	0.020	0.030	**0.030**	0.020	-	0.028
	ORB-SLAM3 [37]	0.035	0.033	0.035	**0.051**	0.082	0.038	**0.014**	0.024	0.032	**0.014**	**0.024**	0.019
	MAVIS(Ours)	**0.024**	**0.025**	**0.032**	0.053	**0.075**	**0.034**	0.016	**0.021**	0.031	0.021	0.039	**0.017**

TABLE I: Performance comparison on the EuRoC [38] datasets, RMSE of Absolute trajectory Error (ATE) in meters.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:57:13 UTC from IEEE Xplore. Restrictions apply.

Sequence name	Difficulties	Sequence length	Processing time	MAVIS(Ours)	BAMF-SLAM	Maplab2.0
site1_handheld_1	dark around stairs going to Floor 1	204.71s	224.59s	**32.5**	10.0	15.0
site1_handheld_2	dark around stairs going to Floor 2	167.11s	211.98s	**23.75**	5.0	12.5
site1_handheld_3	insufficient overlap for multi-session	170.63s	204.09s	**22.5**	17.5	5.0
site1_handheld_4	-	295.42s	364.41s	**30.0**	5.0	8.33
site1_handheld_5	-	159.29s	196.86s	**26.67**	11.67	13.33
site1_multi_session				4.17	-	**5.28**
site2_robot_1	unsynchronised cameras, long, no loop closure	699.31s	531.89s	**15.71**	6.43	7.86
site2_robot_2	unsynchronised cameras	305.79s	194.40s	**53.33**	28.33	15.0
site2_robot_3	dark, insufficient overlap for multi-session	359.00s	187.30s	**19.0**	13.0	5.0
site2_multi_session				**3.33**	-	2.33
site3_handheld_1	-	97.18s	124.07s	**105.0**	45.0	10.0
site3_handheld_2	dropped data	148.13s	182.31s	35.0	**56.67**	11.67
site3_handheld_3	dropped data	189.60s	243.46s	23.75	**31.25**	7.5
site3_handheld_4	-	106.88s	130.34s	**65.0**	37.5	10.0
site3_multi_session				**19.55**	-	7.73
			Total score	**452.21 / 27.0**	267.35 / -	121.19 / 15.3

TABLE II: Difficulties, timing and score information for test sequences

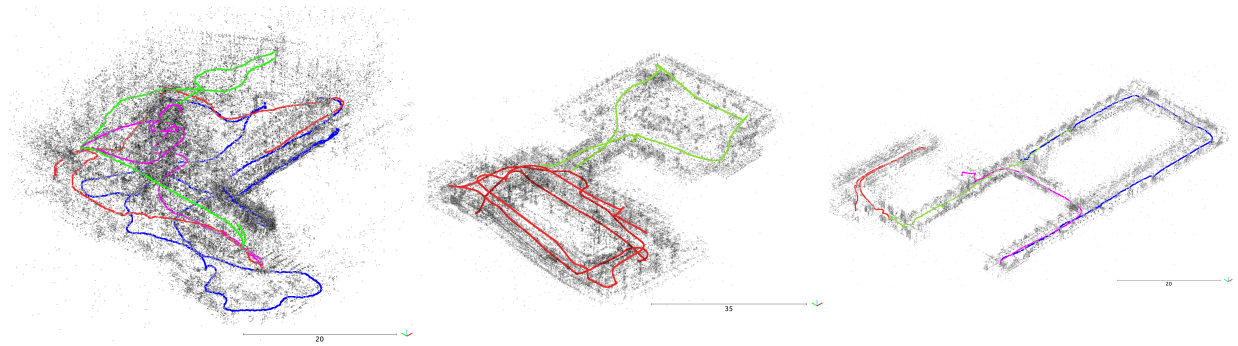

Fig. 5: Perspective view of reconstructed scenes. From left to right: site_1, site_2, site_3. Trajectory is coloured in sequence from red, green, blue, magenta.

2023 datasets [41]. This datasets involve handheld sequences using an Alphasense multi-camera development kit from Sevensense Robotics AG, which synchronizes an IMU with four grayscale fisheye cameras for data collection. For robot sequences, it uses four stereo OAK-D cameras and a high-end Xsens MTi-670 IMU mounted on a ground robot. We maintain consistent parameters within each dataset for our experiments. However, for the robot sequences in the Hilti Challenge 2023 datasets, we encountered inter-stereo-pair time synchronization issues in *site2_robot_1*. Consequently, we used only a pair of stereo cameras with IMU for this sequence. For the other two robot sequences, we selected the best-synchronized four cameras in each sequence. The datasets provide millimeter-accurate ground truth on multiple control points for ATE evaluation. We compared our approach to BAMF-SLAM [23], ranked 2nd in the single-session vision/IMU-only track, and Maplab2.0 [43], ranked 2nd in the multi-session track. Results are illustrated in Table II, the following are worth noting:

- We provide the difficulties, timing and score information for each sequence in Table II. The datasets suffer from challenges such as low-light, textureless environment, unsynchronized cameras, data loss, and the absence of closed-loop exploration scenarios. While all methods successfully process the entire datasets without any gross errors.
- Our method achieves superior performace, clearly outperforms in both single-session and multi-session. We achieve close to 2 times the score compared to the 2nd place. More detailed analysis and further quantitative results can be found on our technical report in [41].
- Our system's runtime performance also demonstrates its practical potential. It runs in real-time on a standard desktop using only CPU. However, BAMF-SLAM [23] requires a Nvidia GeForce RTX 4090 GPU for processing and still runs 1.6 times slower than ours.

We also test our method in the Hilti SLAM Challenge 2022 [36], achieving best results with a score of 130.2, which is three times higher than the 2nd place's score of 40.9. Please refer to the live leaderboard on [41] for more details.

VI. Conclusion

In this paper, we introduce MAVIS, an optimization-based visual-inertial SLAM system for multi-camera systems. Compared to alternatives, we present an exact IMU pre-integration formulation based on the $\mathbf{SE_2(3)}$ exponential, effectively improving tracking performance, especially during rapid rotations and extended integration times. We also extend front-end tracking and back-end optimization modules for multi-camera systems and introduce implementation particulars to enhance system performance in challenging scenarios. Extensive experiments across multiple datasets validate the superior performance of our method. We believe this robust and versatile SLAM system holds significant practical value for the community.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:57:13 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

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

[2] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[3] J. Delmerico and D. Scaramuzza, "A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 2502–2509.

[4] P. van Goor and R. Mahony, "Eqvio: An equivariant filter for visual-inertial odometry," *IEEE Transactions on Robotics*, 2023.

[5] T. Qin, J. Pan, S. Cao, and S. Shen, "A general optimization-based framework for local odometry estimation with multiple sensors," *arXiv preprint arXiv:1901.03638*, 2019.

[6] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "Openvins: A research platform for visual-inertial estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4666–4672.

[7] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an open-source library for real-time metric-semantic localization and mapping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1689–1696.

[8] V. Usenko, N. Demmel, D. Schubert, J. Stückler, and D. Cremers, "Visual-inertial mapping with non-linear factor recovery," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 422–429, 2019.

[9] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[10] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration theory for fast and accurate visual-inertial navigation," *IEEE Transactions on Robotics*, pp. 1–18, 2015.

[11] J.-s. Cui, F.-r. Zhang, D.-z. Feng, C. Li, F. Li, and Q.-c. Tian, "An improved slam based on rk-vif: Vision and inertial information fusion via runge-kutta method," *Defence Technology*, 2021.

[12] D. Scaramuzza, "Tutorial on visual odometry," 2016. [Online]. Available: http://mrsl.grasp.upenn.edu/loiannog/tutorial_ICRA2016/VO_Tutorial.pdf

[13] G. Huang, "Visual-inertial navigation: A concise review," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 9572–9582.

[14] S. Urban and S. Hinz, "Multicol-slam-a modular real-time multi-camera slam system," *arXiv preprint arXiv:1610.07336*, 2016.

[15] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[16] J. Kuo, M. Muglikar, Z. Zhang, and D. Scaramuzza, "Redesigning slam for arbitrary multi-camera systems," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2116–2122.

[17] P. Furgale, U. Schwesinger, M. Rufli, W. Derendarz, H. Grimmett, P. Mühlfellner, S. Wonneberger, J. Timpner, S. Rottmann, B. Li, *et al.*, "Toward automated driving in cities using close-to-market sensors: An overview of the v-charge project," in *2013 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2013, pp. 809–816.

[18] Y. Wang and L. Kneip, "On scale initialization in non-overlapping multi-perspective visual odometry," in *International Conference on Computer Vision Systems*. Springer, 2017, pp. 144–157.

[19] L. Heng, B. Choi, Z. Cui, M. Geppert, S. Hu, B. Kuan, P. Liu, R. Nguyen, Y. C. Yeo, A. Geiger, G. H. Lee, M. Pollefeys, and T. Sattler, "Project AutoVision: Localization and 3D scene perception for an autonomous vehicle with a multi-camera system," *arXiv*, vol. 1809.05477, 2018.

[20] P. Liu, M. Geppert, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys, "Towards robust visual odometry with a multi-camera system," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1154–1161.

[21] Y. Wang, K. Huang, X. Peng, H. Li, and L. Kneip, "Reliable frame-to-frame motion estimation for vehicle-mounted surround-view camera systems," in *2020 IEEE International conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 1660–1666.

[22] L. Zhang, D. Wisth, M. Camurri, and M. Fallon, "Balancing the budget: Feature selection and tracking for multi-camera visual-inertial odometry," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1182–1189, 2022.

[23] W. Zhang, S. Wang, X. Dong, R. Guo, and N. Haala, "Bamf-slam: Bundle adjusted multi-fisheye visual-inertial slam using recurrent field transforms," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 6232–6238.

[24] T. Lupton and S. Sukkarieh, "Efficient integration of inertial observations into visual slam without initialization," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 1547–1552.

[25] ——, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 61–76, 2012.

[26] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation," in *Robotics: Science and Systems*, 2015.

[27] S. Shen, N. Michael, and V. Kumar, "Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft mavs," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 5303–5310.

[28] K. Eckenhoff, P. Geneva, and G. Huang, "Direct visual-inertial navigation with analytical preintegration," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1429–1435.

[29] ——, "Closed-form preintegration methods for graph-based visual–inertial navigation," *The International Journal of Robotics Research*, vol. 38, no. 5, pp. 563–586, 2019.

[30] J. Henawy, Z. Li, W. Y. Yau, G. Seet, and K. W. Wan, "Accurate imu preintegration using switched linear systems for autonomous systems," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 3839–3844.

[31] J. Henawy, Z. Li, W.-Y. Yau, and G. Seet, "Accurate imu factor using switched linear systems for vio," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 8, pp. 7199–7208, 2020.

[32] A. Barrau and S. Bonnabel, "A mathematical framework for imu error propagation with applications to preintegration," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 5732–5738.

[33] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple imus and of individual axes," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 4304–4311.

[34] 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.

[35] P. van Goor, T. Hamel, and R. Mahony, "Constructive equivariant observer design for inertial navigation," *arXiv preprint arXiv:2308.11124*, 2023.

[36] L. Zhang, M. Helmberger, L. F. T. Fu, D. Wisth, M. Camurri, D. Scaramuzza, and M. Fallon, "Hilti-oxford dataset: A millimeter-accurate benchmark for simultaneous localization and mapping," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 408–415, 2023.

[37] C. Campos, J. M. Montiel, and J. D. Tardós, "Inertial-only optimization for visual-inertial initialization," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 51–57.

[38] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[39] Y. He, B. Xu, Z. Ouyang, and H. Li, "A rotation-translation-decoupled solution for robust and efficient visual-inertial initialization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 739–748.

[40] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[41] "Hilti slam challenge 2023," https://hilti-challenge.com/dataset-2023.html, accessed: 2023-08-23.

[42] M. Grupp, "evo: Python package for the evaluation of odometry and slam." https://github.com/MichaelGrupp/evo, 2017.

[43] A. Cramariuc, L. Bernreiter, F. Tschopp, M. Fehr, V. Reijgwart, J. Nieto, R. Siegwart, and C. Cadena, "maplab 2.0–a modular and multi-modal mapping framework," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 520–527, 2022.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:57:13 UTC from IEEE Xplore. Restrictions apply.