Toward Consistent and Efficient Map-Based Visual-Inertial Localization: Theory Framework and Filter Design

Zhuqing Zhang, Yang Song, Graduate Student Member, IEEE, Shoudong Huang, Senior Member, IEEE, Rong Xiong, Member, IEEE, and Yue Wang, Member, IEEE

Abstract—This article focuses on designing a consistent and efficient filter for visual-inertial localization given a prebuilt map. First, we propose a new Lie group with its algebra based on which a novel invariant extended Kalman filter (invariant EKF) is designed. We theoretically prove that, when we do not consider the uncertainty of map information, the proposed invariant EKF is able to naturally preserve the correct observability properties of the system. To consider the uncertainty of map information, we introduce a Schmidt filter. With the Schmidt filter, the uncertainty of map information can be taken into consideration to avoid overconfident estimation while the computation cost only increases linearly with the size of the map keyframes. In addition, we introduce an easily implemented observability-constrained technique because directly combining the invariant EKF with the Schmidt filter cannot maintain the correct observability properties of the system that considers the uncertainty of map information. Finally, we validate our proposed system’s high consistency, accuracy, and efficiency via extensive simulations and real-world experiments.

Index Terms—Consistency, invariant extended Kalman filter (EKF), visual-inertial localization (VIL).

I. INTRODUCTION

LOCALIZATION is a basic module for intelligent robots. Recently, cameras and inertial measurement units (IMUs) have been widely used in localization for their lightweight and low cost, giving birth to visual-inertial odometry (VIO). Under the efforts of researchers from different groups, there have been lots of mature VIO algorithms with properties of high accuracy and high efficiency [1], [2], [3], [4], [5], [6], [7], [8]. However, VIO inevitably suffers from drift for long-term running. In practice, it is more appreciated to perform localization with global information so that, on the one hand, the drift of the estimator can be bounded; on the other hand, further tasks, such as navigation, can be implemented by acquiring the global position. A typical implementation of drift-free localization is employing a prebuilt map.

In general, a good map-based visual-inertial localization (VIL) algorithm should not only be error bounded but be consistent and computationally efficient while fusing the prebuilt map information. An inconsistent algorithm will usually underestimate the uncertainty, leading to unreliable results [31]. According to the previous works, there are two main reasons for the inconsistency of the system: one is that the estimator fails to maintain the correct observability properties of the system [13], [30]; the other is that the uncertainty of the fused information (e.g., the prebuilt map) is not properly used by the system [12].

A theoretically sound method to consistently fuse a prebuilt map with VIO is to regard the problem as an extension of visual-inertial simultaneous localization and mapping (VI-SLAM), which takes all measurements from both the local odometry session and the mapping session to formulate a batch optimization problem. Unfortunately, even if this scheme yields consistent localization, the ever-growing size of the state prohibits online computation. A popular solution makes a compromise to improve the computational efficiency by only maintaining local odometry and the relative transformation between the odometry frame and the map frame, without considering the uncertainty of the map, i.e., assuming the map is perfect [9], [11]. However, as mentioned in [12], the perfect map assumption will lead to overconfident estimation. In particular, Lee et al. [10] point out that the formulation, such as [9] and [11], may result in the observability-deficient issue, causing inconsistency. Therefore, they propose a solution by directly representing the local odometry state in the global frame, discarding the relative transformation between the local odometry frame and the global frame, which is only suitable for map agnostic gravity-aligned global measurements, e.g., global positioning system (GPS).

In this article, we consider a more general situation: First, the prebuilt map has uncertainties, i.e., imperfect map; second, the map-based measurements (i.e., global measurements. For example, the observation of the matched map feature points) have uncertainties; third, the transformation between the map frame and the inertial frame is six degrees of freedom (DoF), i.e., gravity unaligned. Under this formulation, the problem in [10]...
can be regarded as a special case, where GPS acts like a perfect map providing global measurements with uncertainties, and the transformation between the map and the inertial frame is 4-DoF. We argue that the considered general situation is common in practice, where the maps are obtained indoors or via visual-based structure from motion (SFM) [39]. How to consistently and efficiently fuse such general map information is a nontrivial problem, and the theoretical consistency analysis remains vague, naturally leaving the corresponding solution very limited.

This work aims to fill the gap by proposing observability-based consistency theories and developing an observability-constrained (OC) filter for consistent and efficient map-based VIL. Specifically, we employ a visual-based prebuilt map, which contains map keyframe poses and map feature positions with uncertainties. We regard the relative transformation that connects the local inertial frame and the map (global) frame as the augmented variable and regard the whole VIL system that maintains a local VIO state and the augmented variable simultaneously as the augmented system. Furthermore, two kinds of augmented systems are analyzed: the perfect augmented system that assumes the map is perfect and the imperfect augmented system that considers the uncertainty of the map. According to the theoretical analyses toward the observability properties of the (im)perfect augmented system, we get the following conclusions. First, the ideal perfect augmented system for our map-based VIL problem has four unobservable dimensions. Second, the ideal imperfect augmented system for our map-based VIL problem has ten (four plus six) dimensions that are unobservable. Third, the ever-changing estimation of the augmented variable is the root of inconsistency issue, only three unobservable dimensions of the (im)perfect augmented system being preserved.

Based on the above conclusions, we design a consistent and efficient map-based VIL algorithm. First of all, based on our proposed novel Lie group and Lie algebra, an invariant EKF is designed, which is able to naturally preserve the correct observability properties of the perfect augmented system. Then, inspired by [12] and [14], we introduce the Schmidt EKF to take the uncertainty of the map into consideration while keeping the computation at a low level. Moreover, to keep the correct observability properties of the imperfect augmented system, we follow the idea of the article presented in [31] to constrain the updating of the augmented variable. Finally, a multistate constraint is introduced to improve the accuracy of the localization. With our final multistate observability-constrained Schmidt-invariant Kalman filter (MSOC-S-IKF), the system can correctly maintain the ten unobservable dimensions of the imperfect augmented system while satisfying the real-time demand.

In summary, the contributions of this article are as follows that have been verified by extensive simulations and real-world experiments.

1) Theoretically analyze the observability properties of the perfect/imperfect augmented system and ground these theories to physical explanations.

2) Propose an invariant EKF by introducing a novel Lie group structure, which is proven to maintain the consistency of the perfect augmented system naturally.

3) Propose a consistent and efficient filter that combines the invariant EKF, Schmidt filter, multistate constraint, and observability constraint such that the system has the abilities to consider the uncertainty of map information, keep the cost of computation and storage at a low level, maintain the correct observability properties of the imperfect augmented system, and perform accurate localization.

The source code is available online: https://github.com/zhuqingzhang/MSOC-S-IKF.

II. RELATED WORKS

A. Visual-Inertial Odometry

VIO can be divided into filter-based and optimization-based in general. For the filter-based VIO, the IMU information is modeled in the propagation step, while the relative motion calculated by the vision sensor is employed to update the state [17]. One famous framework is loosely coupled multisensor fusion [15], which fuses the results from IMU and the visual odometry in an outer EKF framework, i.e., the estimations from the IMU are independent of those from the visual odometry. A more popular tightly coupled framework is the multistate-constrained Kalman filter (MSCKF) [1], [2], [3], [4], which maintains a sliding window so that multiple states are able to constrain the update and, therefore, results in more accurate estimation. Besides, benefiting from the multistate constraint, MSCKF does not maintain any feature points in the state vector, realizing a very efficient VIO. Another tightly coupled filter-based VIO employs a robocentric formulation [5]. It keeps features in the state and utilizes iterative EKF to obtain the accurate estimation results, but it is less efficient than MSCKF.

For the optimization-based VIO, the measurements are formalized as a graph and optimized by iterative algorithms. OKVIS [8] is a famous early implementation of the optimization-based VIO, where IMU and visual measurements are tightly coupled in a factor graph. To reduce the number of variables in the optimization, preintegration [18] is proposed as a factor between the two keyframes. Following this framework, a popular VIO, VINS-mono [6], is proposed and utilized in many aerial and ground vehicles. Forster et al. [7] propose a VIO system where both pixel intensities and features are used to perform robust and efficient estimation.

Compared with the filter-based VIO, the optimization-based VIO requires much more computation. Considering accuracy and efficiency, we use Open-VINS [2], a variation of MSCKF, as the VIO part of our proposed system.

B. Drift-Free VIL

Although the studies on VIO have gained much progress, VIO inevitably suffers from drift. For the filter-based VIL, to bound the drift in VIO, GPS [10] and map-based measurements [11], [12] are introduced using the augmented state formulation for fusion. These works make tradeoffs between the computation cost and the consideration of map uncertainty. One challenge is to design a filter that is both consistent and efficient. In [12], the Schmidt filter is introduced to account for the uncertainty of the map. However, the system maintains all the map features, whose size (denoted as n) could be tens of thousands, in the state vector, unsuitable for large scenes. In this article, we propose to only retain the map keyframe poses in the state, whose size (denoted as m) is much smaller than n, improving the computational efficiency. Moreover, as mentioned in [10] and [13], the system with the formulation, such as [12], i.e., the augmented system, will suffer from inconsistency due to the linearization toward nonlinear systems, which is not addressed in [12].
To reduce the drift of the VIO, the global measurements can also be integrated into the system using an optimization-based VIL algorithm. In VINS fusion [6], [9], [33], [34], which is the extension of VINS-mono [6], map-based measurements are fused in an outer-loop pose-graph optimization, i.e., loosely coupled. In [29], a tightly coupled approach is proposed to optimize all measurements together with respect to all states. When both odometry and map-based measurements are integrated, a degeneracy analysis is performed in [19]. In general, optimization-based VIL methods have superior accuracy than filtering-based methods because the nonlinear system can be relinearized with the cost of more expensive resources. However, the map uncertainty is ignored in all the above-mentioned methods, which may make the estimation unreliable.

C. Observability Analysis and Consistency Maintaining

The inconsistency issue of point feature-based EKF SLAM is first demonstrated in [28]. Its authors find that when a stationary robot observes a new feature multiple times, the uncertainty along the robot's orientation will decrease, which is against our intuition since the observation of a new feature should not provide information about the ego state of the robot. Further analysis of the inconsistency issue is discussed in [13], where Huang and Dissanayake also consider the case that a robot observes a feature from two positions. Huang et al. [16] first theoretically analyze the observability properties of the EKF-SLAM and argue that due to the linearization toward the nonlinear system, the unobservable dimensions become observable. Based on this analysis, the first-estimated Jacobian is proposed to retain the correct observability properties of the system, which turns out to be effective in improving the system's consistency. Furthermore, in their follow-up work, an OC method is introduced [31]. Different from designing constraints to maintain the correct observability, a series of manifold-based filters (invariant filters) are proposed [4], [20], [21], [22], [23], [24]. For example, Barral and Bonnabel [20] formulate the state on a special Lie group and define nonlinear errors through Lie algebra to perform the invariant EKF SLAM. With this kind of formulation, the proposed invariant EKF is able to automatically maintain the correct observability properties of the system.

The observability analyses and the methods for improving the consistency of the system mentioned above have already been extended and applied to VIO systems [25], [26], [27]. Unfortunately, for the VIL problem we considered, the augmented variable and map measurements are introduced into the system. Therefore, the observability analyses of VIO cannot reflect the observability of the augmented system. In [10], the observability of the VIO fusing GPS is analyzed, which shows some insights into the observability of the augmented system. However, it only points out that the augmented system suffers from inconsistency and proposes a consistent estimation algorithm when the state is estimated in a gravity-aligned map frame, i.e., the map has 4-DoF. To be specific, after initializing the relative transformation between the VIO frame and the 4-DoF map frame, the variables in the state vector are transformed from the 4-DoF VIO frame to the 4-DoF map frame, and the state vector is propagated by IMU measurements in the 4-DoF map frame. Nevertheless, for a more general gravity-unaligned map (a 6-DoF map), the state cannot be correctly propagated by the IMU measurements in the map frame because the gravity in the map frame is unknown. In this article, we propose a general framework that can be used for the case of 6-DoF maps and extend the observability analysis in [10] to the case considering the uncertainty of map information.

III. PRELIMINARY AND PROBLEM STATEMENT

In this section, some preliminaries about Lie group, Lie algebra, and invariant error are introduced first. Then, the problem description for the map-based VIL system is provided, which could give the readers a general idea of the problem addressed in the following sections.

A. Theoretical Background and Preliminaries

Matrix Lie group and Lie algebra [35]: A matrix Lie group \( \mathbb{G} \subset \mathbb{R}^{N \times N} \) is a subset of invertible square matrices with the following properties:

\[
\mathbf{I} \in \mathbb{G}
\]

\[
\forall \mathbf{X} \in \mathbb{G} \text{ s.t. } \mathbf{X}^{-1} \in \mathbb{G}
\]

\[
\forall \mathbf{X}_1, \mathbf{X}_2 \in \mathbb{G} \text{ s.t. } \mathbf{X}_1 \mathbf{X}_2 \in \mathbb{G}
\]

where \( \mathbf{I} \) is the identity matrix, and \( N \) is the order of the square matrices.

Denote \( \mathfrak{g} \cong \mathbb{R}^D \) as the Lie algebra of \( \mathbb{G} \). An exponential map \( \exp : \mathbb{R}^D \rightarrow \mathbb{G} \) is defined by

\[
\exp (\xi) = \exp_{\mathfrak{g}} (\xi)
\]

(2)

where \( \xi \in \mathbb{R}^D, \exp_{\mathfrak{g}} \) is the standard matrix exponential map, and \( \mathfrak{g} \) is the isomorphism of \( \mathbb{R}^D \) and \( \mathfrak{g} \). For simplicity, we denote \( \mathfrak{g} (\cdot) \triangleq (\cdot)^{+} \). Besides, we denote log as the inverse operation of \( \exp \).

The usually used Lie group (and its algebra) includes the special orthogonal group \( SO(3) (\mathfrak{so}(3)) \), which is used to represent rotation matrices; the special Euclidean group \( SE(3) (\mathfrak{se}(3)) \), which is the combination of rotation and translation; and \( SE_{2+K} (3) (\mathfrak{se}_{2+K} (3)) \), which is the extension of the special Euclidean group and consists of one rotation matrix and \( 2 + K \) vectors [20].

Left- and right-invariant error: Suppose \( \mathbf{X}_t, \hat{\mathbf{X}}_t \in \mathbb{G} \) are the true and the estimated states at the time step \( t \), respectively. Then, we have the following definition:

**Definition 1:** The left- and right-invariant errors between the two states are [22] as follows:

\[
\eta_t^L = \mathbf{X}_t^{-1} \hat{\mathbf{X}}_t \quad \text{(Left invariant error)}
\]

(3)

\[
\eta_t^R = \mathbf{X}_t \hat{\mathbf{X}}_t^{-1} \quad \text{(Right invariant error)}
\]

(4)

The reason why (3)/(4) is called left/right-invariant error is that when we left/right multiply an arbitrary element in \( \mathbb{G} \) to both states \( \mathbf{X}_t \) and \( \hat{\mathbf{X}}_t \), the error is invariant. In the rest of this article, the right-invariant error, which is widely used in the localization problem [4], [20], [22], [26], will be used to formulate the problem.

B. Map-Based VIL Problem

Before stating the map-based VIL problem, we define the reference frames and terminologies frequently used in this article in Table I.

As shown in Fig. 1 and Table I, there are five reference frames in the map-based VIL problem, i.e., \( L, I_t, C_t, G \), and \( K F_t \). The

\[1\] Throughout this article, the symbol with/without * denotes the estimated/true value.
prebuilt map could be built by visual SLAM or SFM algorithms and could involve hundreds of map keyframes \( \{KF_i\} \) with known poses and thousands of map features \( \{F_j\} \) with known positions in \( G \).

Compared with the VIO solution, such as [2], the solution for VIL needs to additionally fuse the prebuilt map information so that it can produce a more accurate local pose \( L_TI_t \) than the article presented in [2], as well as an accurate map-based pose \( G_TI_t \).

Thereby, \( G_TI_t \) is derived, and the map-based VIL problem is solved.

However, such a problem requires the augmented system to consistently and efficiently estimate \( L_TI_t \) and \( G_TI_t \), given a sequence of IMU measurements in frame \( I_t \) and images in frame \( C_t \), and the prebuilt map information consisting of keyframes \( \{KF_i\} \) and features \( \{F_j\} \) in frame \( G \). In the rest of this article, we elaborate on how to design a consistent and efficient filter to solve this problem.

### C. Flow of the Whole Algorithm

Before going into details, we give a flowchart in Fig. 2 to overview the whole procedure of our proposed framework. The framework consists of three main blocks: local VIO, feature matching, and map-based localization.

1) **Local VIO:** This block is a VIO, which processes IMU and camera data to online estimate the local pose \( L_TI_t \) (cf. Sections IV-C and VII-B).

2) **Feature matching:** This block outputs the matching information between the prebuilt map and the VIO. The matching information consists of the poses and covariance matrices of matched keyframes, the positions of the matched map features in the frame \( G \), the two-dimensional (2-D) pixel observations of the matched map features in the matched keyframes, and the 2-D pixel observations of the matched map features in the current image \( C_t \) (cf. Feature matching of Section VIII-B).
3) Map-based localization: This block consists of two main parts, i.e., initializing the augmented variable (cf. Section VII-C) and consistently updating the state with map-based measurements (cf. Section VII-A).

The system will output the local pose and the augmented variable at the same frequency as the camera frame rate. By multiplying the local pose and the augmented variable, we can get the pose in the map (global) frame.

IV. INVARINT EKF FOR PERFECT MAP-BASED VII

In this section, we assume the map is perfect without uncertainty. With the perfect map assumption, we propose an invariant EKF for the perfect augmented system. First, a new Lie group with its algebra is introduced to construct the kinematics of the augmented system. Then, the state of the system, propagating procedure, and updating procedure are illustrated step-by-step.

A. Novel Lie Group and Algebra

Although the invariant EKF based on $SE_{2+K}(3)$ is successfully applied to the SLAM problem to achieve consistent estimators [20], [36], [37], $SE_{2+K}(3)$ is not adaptable to the augmented system. This is because $SE_{2+K}(3)$ contains only one rotation matrix, whereas another rotation matrix in the augmented variable also needs to be considered.

Therefore, we introduce a novel Lie group $SE_{2+K}(3)$ with its algebra $se_{2+K}(3)$.

Definition 2: With $SE_{2+K}(3)$ and SO(3), a new Lie group denoted as $SE_{2+K}(3)$ is defined as follows:

$$SE_{2+K}(3) \triangleq \{ T = \begin{bmatrix} T_1 & 0_{(5+K+M) \times (3M)} \\ 0_{(5+K+M) \times (3M)} & T_2 \end{bmatrix} \}

T_1 \in SE_{2+K+M}(3), T_2 \triangleq \text{diag}(R_1, \ldots, R_M)

R_i \in SO(3), i = 1, \ldots, M\}$$

(5)

where diag(...) represents a diagonal block matrix.

The corresponding Lie algebra $se_{2+K}(3)$ is defined by

$$se_{2+K}(3) \triangleq \{ \mathcal{L}_{se_{2+K}}(\phi), \phi = \begin{bmatrix} \theta_0^T \phi_1 \cdots \phi_{2+K+M} \\
\theta_1^T \cdots \theta_M^T \end{bmatrix} \in \mathbb{R}^{9+3K+3M} \}

\mathcal{L}_{se_{2+K}}(\phi) = \phi^A

\triangleq \begin{bmatrix}
(\theta_0)_x \\
0_{(2+K+M) \times (3M) × 3} \\
\phi_1 \cdots \phi_{2+K+M} \\
0_{3M \times (2+K+M)} \\
0_{(2+K+M) \times (3M)} \\
0_{(3M \times (2+K+M)} \\
0_{3M \times (2+K+M)} \\
0_{3M \times (2+K+M)} \\
0_{3M \times (2+K+M)}
\end{bmatrix}\Theta

\triangleq \text{diag}(\theta_1)_x, \ldots, (\theta_M)_x.$$

(6)

In the above definition, both $\theta_i, i = 0, \ldots, M$ and $\phi_i, i = 1, \ldots, 2+K+M$ are 3-D vectors. Especially, $\theta_i$ is related with $R_i \in SO(3)$ and is derived by $\theta_i = \log(R_i)$. It is clear that $SE_{2+K}(3)$ satisfies the properties in (1). We can regard $SE_{2+K}(3)$ as the combination of one $SE_{2+K+M}(3)$ and $M$ SO(3) $\times \mathbb{R}^3$, and we insert the $M \mathbb{R}^3$ into $SE_{2+K+M}(3)$ to get $SE_{2+K+M+3}(3)$ while the $M$ SO(3) are arranged diagonally. As presented in the following part, $SE_{2+K}(3)$ is used for the augmented system.

B. State for the Perfect Augmented System

We define the state of the augmented system at time step $t$ as $X_t \in SE_{2+K}(3) \times \mathbb{R}^6$

$$X_t = (\begin{bmatrix} \phi_{L_t} \phi_{I_{L_t}} \phi_{p_{L_t}} \cdots \phi_{p_{f_{L_t}}} \phi_{G_t} \end{bmatrix}^T \theta_{L_{I_t}} \theta_{L_{I_t}} \theta_{L_{p_{L_t}}} \theta_{L_{p_{L_t}}} \theta_{L_{G_t}} \theta_{L_{G_t}})^T 

(7)

$$

where $\phi_{L_t}$ represents the rotation of the body frame $I$ in $L$, $\phi_{I_{L_t}}$ and $\phi_{p_{L_t}}$ are the velocity and the position of the body in $L$, respectively, $\phi_{p_{f_{L_t}}} i = 1, \ldots, K$ is the position of the $i$th (local) feature in $L$, $\phi_{G_t}$ and $\phi_{G_t}$ are the translation and the rotation parts of the relative transformation between $G$ and $L$ (augmented variable), and $b_9$ and $b_{10}$ represent the gyroscope and accelerometer bias. For simplicity, in the rest of this article, only one local feature $\phi_{p_{f_{L}}}$ is employed to formulate the state. In this case, $X_t$ will belong to $SE_{2+K+1}(3)$.

Error definition: With the state defined above, we formulate the error state as follows:

$$e_t = (\hat{X}_t \bar{X}_t^{-1} (\hat{B}_t - B_t)) \triangleq (\eta_{A_t}, \xi_{B_t}).$$

(8)

Note that $\eta_{A_t} \in SE_{2+K+1}(3)$. Applying (2) and (6) to $\eta_{A_t}$, the following approximation holds up to the first order:

$$\eta_{A_t} = \text{exp}_{\hat{L}}(\mathcal{L}_{se_{2+K}}(\xi_{A_t})) \approx I_{10} + \mathcal{L}_{se_{2+K}}(\xi_{A_t})$$

(9)

where

$$\xi_{A_t} \triangleq \begin{bmatrix} \xi_{\theta_{L_{I_t}}} & \xi_{\phi_{L_{I_t}}} & \xi_{\phi_{L_{p_{L_t}}}} & \xi_{\phi_{L_{G_t}}} & \xi_{\theta_{L_{G_t}}} \end{bmatrix}^T$$

(10)

and

$$\xi_{\phi_{L_{I_t}}} = \hat{\phi}_{L_{I_t}}$$

$$\xi_{\phi_{L_{p_{L_t}}} = \hat{\phi}_{L_{p_{L_t}}} - (I_3 + (\hat{\phi}_{L_{L_{I_t}}})_x) \hat{\phi}_{L_{p_{L_t}}}$$

$$\xi_{\phi_{L_{G_t}}} = \hat{\phi}_{L_{G_t}} - (I_3 + (\hat{\phi}_{L_{L_{I_t}}})_x) \hat{\phi}_{L_{G_t}}$$

$$\xi_{\phi_{L_{G_t}}} = \hat{\phi}_{L_{G_t}}$$

$$\xi_{\phi_{L_{G_t}}} = \hat{\phi}_{L_{G_t}}$$

(11)

In the above equations, $\hat{\phi}$ represents the error of the rotation matrix, which is defined as $\hat{\phi} \triangleq \log(RR^{-1})$, and $\eta_{R_t} \triangleq L_{R_t} \hat{R}_{L_{I_t}} = I_3 + (\hat{\phi}_{L_{L_{I_t}}})_x$ up to the first order. $e_t \triangleq [\xi_{A_t} \xi_{B_t}]^T$ defines the right-invariant error, which is used to formulate the error propagation and update equations.
C. Propagation for the Perfect Augmented System

With the state defined by (7), the system kinematics is given as follows:
\[
\begin{align*}
\dot{L}\mathbf{R}_{t} &= L\mathbf{R}_{t}(\omega_{t} - b_{gt} - w_{gt})^\times, \\
\dot{L}\mathbf{v}_{t} &= L\mathbf{v}_{t}(a_{t} - b_{at} - w_{at}) + g, \\
\dot{L}\mathbf{p}_{t} &= L\mathbf{p}_{t} + L\mathbf{R}_{Gt} = 0_{3 \times 1}, \\
\dot{b}_{gt} &= w_{bg}, \quad b_{at} = w_{bat},
\end{align*}
\]
where $\omega$ and $a$ are the measurements of angular velocity and linear acceleration from the IMU, $w_{g}$ and $w_{a}$ are the IMU measurement noises, $w_{bg}$ and $w_{bat}$ are the Gaussian random walk noises, and $g = [0 \ 0 \ -9.8]^T$ is the gravitational acceleration in the local inertial frame $L$.

Different from the standard EKF error propagation function, we need to formulate the error propagation function with the right-invariant error defined by (11), which leads to
\[
\frac{d}{dt} e_t = \mathbf{A}_t e_t + \mathbf{W}_t w_t
\]
where
\[
\mathbf{A}_t = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{L}\mathbf{R}_{t} & 0 \\
0 & 0 & 0 & 0 & \mathbf{L}v_{t} & \mathbf{L}\mathbf{R}_{t} & 0 & \mathbf{L}\mathbf{R}_{t} & 0 \\
0 & 0 & 0 & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{p}_{t} & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{R}_{t} & 0 \\
0 & 0 & \mathbf{L}\mathbf{p}_{t} & \mathbf{L}\mathbf{R}_{Gt} & \mathbf{L}\mathbf{R}_{t} & 0 & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{R}_{Gt} \\
0 & \mathbf{L}\mathbf{p}_{t} & \mathbf{L}\mathbf{p}_{t} & \mathbf{L}\mathbf{R}_{Gt} & \mathbf{L}\mathbf{R}_{t} & 0 & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{R}_{Gt} \\
0 & 0 & 0 & \mathbf{L}\mathbf{R}_{Gt} & \mathbf{L}\mathbf{R}_{t} & 0 & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{R}_{Gt} \\
0 & 0 & 0 & 0 & \mathbf{L}\mathbf{R}_{Gt} & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{R}_{t} & \mathbf{L}\mathbf{R}_{Gt}
\end{bmatrix}
\]
and $w_{t} \triangleq \begin{bmatrix} w_{bg}^T, w_{at}^T, 0_{12 \times 1}, w_{bg}, w_{bat}^T \end{bmatrix}^T$. All the 0 in the above representations are 3 $\times$ 3 zero matrices. The detailed derivation is given in Appendix A of the supplementary material [50].

As indicated by (13) and (14), $\mathbf{A}_{X_{A_t}}$ only consists of constants. This is because the right-invariant error defined by (11) is propagated independently of the system state, according to Theorem 4 in [22].

Combining $X_{A_t}$ with IMU bias parameters $B_t$, we can design an “imperfect-invariant EKF” as mentioned in [36] and [38]. The linearized error kinematics of (12) is given by (13). The covariance of the right-invariant error is computed by solving the Riccati equation
\[
\frac{d}{dt} \mathbf{P}_{t} = \mathbf{A}_t \mathbf{P}_{t} + \mathbf{P}_{t} \mathbf{A}_t^T + \mathbf{Q}_{t}
\]
where $\mathbf{Q}_{t} \triangleq \mathbf{W}_t \text{Cov}(w_t) \mathbf{W}_t^T$, and Cov$(w_t)$ is the covariance matrix of the noise vector $w_t$.

D. Updating for the Perfect Augmented System

The updating procedure is closely related to observation functions. As indicated by Barrau and Bonnaibel [20], for a general observation function, $y_t = h(X_{A_t}, B_t) + \gamma_t$, where $y_t$ is the observation of the visual sensor and $\gamma_t$ is the zero-mean Gaussian noise with the covariance $\mathbf{V}_t$. The observation error (i.e., the innovation term) is written as
\[
r_t = y_t - h(X_{A_t}, B_t)
\]
\[
= h(X_{A_t}, B_t) - h(X_{A_t}, B_t) + \gamma_t
\]
\[
= h(\exp(-\xi_{A_t} g \mathbf{X}_{A_t}, \mathbf{B}_t - \xi_{B_t}) - h(X_{A_t}, B_t) + \gamma_t.
\]
Then, we perform the first-order Taylor expansion to the above function. However, unlike the standard Taylor expansion represented with the standard vector error, we need to formulate the Taylor expansion with the right-invariant error $e_t \triangleq \begin{bmatrix} \xi_{A_t}^T, \xi_{B_t}^T \end{bmatrix}^T$ as follows:
\[
r_t = y_t - h(X_{A_t}, B_t) = -H_t e_t + \gamma_t + O(||e_t||^2). \tag{18}
\]
The observation functions in this article can be divided into the local feature-based observation function and the map feature-based observation function.

Local feature-based observation function: Suppose the robot observes a feature $L\mathbf{p}_f$ in the local odometry frame $L$ at the current time step $t$. For simplicity, we assume that the extrinsic between the camera and the IMU is an identity matrix so that the observation function can be given as follows:
\[
y_{L_t} = h(L\mathbf{R}_{t}^T (L\mathbf{p}_f - L\mathbf{p}_{t}) + \gamma_{L_t} \tag{19}
\]
where $h(\cdot)$ is the projection function of the camera. The subscript $L_t$ of $y_{L_t}$ and $\gamma_{L_t}$ indicates the local observation. The Jacobian matrix of (19) is given by
\[
\mathbf{H}_{L_t} = -\nabla h|_q L\mathbf{R}_{t}^T [0_3, 0_3, I_3, -I_3, 0_3, 0_3, 0_3]
\tag{20}
\]
where $q \triangleq L\mathbf{R}_{t}^T (L\mathbf{p}_f - L\mathbf{p}_{t})$ and $\nabla h \triangleq \frac{\partial h}{\partial q}$. The detailed derivation is in Appendix B of our supplementary material [50].

Map feature-based observation function: Suppose we have a feature $G\mathbf{p}_F$ provided by the map, which is observed by the robot again at the current time step $t$ (cf. Fig. 1). Then, with the augmented variable $T_G$ as a bridge, the observation function is formulated as follows:
\[
y_{G_t} = h(T_G^T (L\mathbf{R}_{Gt}^T G\mathbf{p}_F + L\mathbf{p}_{Gt}) - L\mathbf{p}_F) + \gamma_{G_t}. \tag{21}
\]
The subscript $G_t$ of $y_{G_t}$ and $\gamma_{G_t}$ indicates the map-based (global) observation. The Jacobian matrix of (21) is given by
\[
\mathbf{H}_{G_t} = -\nabla h|_q T_G^T L\mathbf{R}_{Gt}^T [0_3, 0_3, I_3, -I_3, 0_3, 0_3]
\tag{22}
\]
where $q \triangleq L\mathbf{R}_{Gt}^T (L\mathbf{R}_{Gt}^T G\mathbf{p}_F + L\mathbf{p}_{Gt} - L\mathbf{p}_F)$ and $\nabla h \triangleq \frac{\partial h}{\partial q}$. The detailed derivation is in Appendix B of the article presented in [50].

With (13), (19), and (21), the state of the augmented system can be propagated and updated with the procedure, such as EKF. The whole process of invariant EKF for the perfect augmented system is summarized in Algorithm 1.
Algorithm 1: Invariant EKF for the Perfect Augmented System.

**Input:** the posterior of the state at time step $t-1$, $\mathbf{X}_{t-1}$, the covariance of the state $\mathbf{P}_{t-1}$, the IMU measurements $\mathbf{a}_t, \mathbf{\omega}_t$.

**Output:** $\mathbf{X}_t, \mathbf{P}_t$.

1. Propagate the state with (12) to get the prediction $\mathbf{X}_{t|t-1}$, and propagate the covariance with (13) and (16) to get $\mathbf{P}_{t|t-1}$.
2. if there are tracked local features then
3. compute the innovation term $r_L$, and $H_L$, with (18)–(20).
4. end if
5. if there are matched features from the map then
6. compute the innovation term $r_{G_i}$ and $H_{G_i}$ with (18), (21), and (22).
7. end if
8. Stack $r_L$, and $r_{G_i}$ as $\mathbf{r}_t$ and $H_L$, and $H_{G_i}$ as $\mathbf{H}_t$.
9. Update the state and the covariance: $\mathbf{S}_t = \mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^\top + \mathbf{V}_t$

$$
\mathbf{K}_t \triangleq \begin{bmatrix} \mathbf{K}_{A_t} \\ \mathbf{K}_{B_t} \end{bmatrix} = \mathbf{P}_{t|t-1} \mathbf{H}_t^\top \mathbf{S}_t^{-1} 
$$

$$
\mathbf{X}_t = \mathbf{X}_{t|t-1} + \exp(\mathbf{K}_{A_t} \mathbf{r}_t) \mathbf{X}_{A_t|t-1} 
$$

$$
\mathbf{P}_t = \mathbf{P}_{t|t-1} - \mathbf{K}_t \mathbf{H}_t \mathbf{P}_{t|t-1} 
$$

10. return $\mathbf{X}_t, \mathbf{P}_t$.

V. INVARIANT EKF FOR IMPERFECT MAP-BASED VIL

In Section IV, we design an invariant EKF for the perfect augmented system by employing a novel Lie group and its algebra. However, in practice, the prebuilt map is actually imperfect and has uncertainty. Neglecting the map uncertainty will lead to overconfident estimation, which encourages us to consider the uncertainty of map information. This kind of imperfect map-based VIL system is called the imperfect augmented system.

A. State for the Imperfect Augmented System

In order to consider the uncertainty of map information, we need to add the map-related variables into the system state so that the state of the imperfect augmented system at time step $t$ is defined as follows:

$$
\mathbf{X}^r_t = (\mathbf{X}_A_t, \mathbf{B}_t, \mathbf{X}_{M_t}) = (\mathbf{X}_t, \mathbf{X}_{M_t})
$$

where $\mathbf{X}_{M_t}$ contains $m$ map keyframe poses $\{\mathbf{X}_{KF_i}, t, \ldots, \mathbf{X}_{KF_{m}, t}\}$ and $n$ map feature positions $\{G_{p_{F_j}, t}, \ldots, G_{p_{F_{n}, t}}\}$. Each map keyframe pose $\{\mathbf{X}_{KF_i}, t\}, i = 1, \ldots, m$, consists of $\{G_{R_{KF_i}, t}, G_{p_{F, i}}\}$.

Following the definition of (11), the right-invariant error of the map keyframe pose, $\epsilon_{KF_i, t} \triangleq \begin{bmatrix} \mathbf{\xi}_{KF_i, t}^T \\ \mathbf{\xi}_{p_{KF_i}, t}^T \end{bmatrix}$, is given as

$$
\epsilon_{KF_i, t} = G_{\tilde{\theta}_{KF_i, t}} - \mathbf{I}_3 + \tilde{\theta} \mathbf{I}_3
$$

while for the map feature position, we define the error in the Euclidean vector space instead of the Lie group space to facilitate the derivation of the Jacobian matrix

$$
\mathbf{\epsilon}_{F_i, t} = \begin{bmatrix} \mathbf{\xi}_{\mathbf{p}_{F_i}, t}^T \\ \mathbf{\xi}_{p_{F_i}, t}^T \end{bmatrix}
$$

We define the nonlinear error of $\mathbf{X}_t$ as

$$
\mathbf{\epsilon}_t \triangleq \mathbf{\xi}_{t}^T \mathbf{\xi}_{t} \quad \mathbf{\epsilon}_{KF_i, t} \cdots \epsilon_{KF_i, t} \cdots \mathbf{\epsilon}_{KF_i, t}
$$

B. Propagation for the Imperfect Augmented System

The kinematics of $\mathbf{X}_t$ is identical to (12). For $\mathbf{X}_{M_t}$, we assume that the map-related variables should remain unchanged

$$
\begin{bmatrix} G_{R_{KF_i, t}} = \mathbf{0}_3 \\ G_{p_{F, i}} = \mathbf{0}_3 \times 1 \\ G_{p_{F, i}} = \mathbf{0}_3 \times 1 \end{bmatrix}
$$

C. Updating for the Imperfect Augmented System

Since the local feature-based observation function is not related to the map information, in this section, we only introduce the map feature-based observation function. As shown in Fig. 1, supposing a map feature $G_{p_{F, i}}$ is observed by the current frame $C_i$ and the map keyframe $KF_i$, we have the following observation functions:

$$
y_{C_i} = h(L_R I_t (I_t L_R G_{p_{F, i}} + L \mathbf{p}_{G, t} - L \mathbf{p}_{I_i})) + \gamma_{C_i}
$$

$$
y_{G_{KF, i}} = h(G_{R_{KF_i, t}} (G_{p_{F, i}} - G_{p_{G, t}})) + \gamma_{G_{KF, i}}
$$

where the superscript $C_i$ means the observation of $C_i$ toward the map feature $F_j$, the superscript $KF_i$ means the observation of the map keyframe $KF_i$ toward the map feature $F_j$, and $\gamma_{C_i}$ and $\gamma_{G_{KF, i}}$ represent the observation noises of $C_i$ and $KF_i$, respectively.

Note that (27) is almost identical to (21), except that the feature $G_{p_{F, i}}$ is variable, whereas the feature $G_{p_{F, i}}$ in (21) is constant.

The Jacobian matrix of (27) is given by

$$
\mathbf{H}_{G_{KF, i}} = \nabla h|_{q_{C_i}} (L_R G_{KF_i} (G_{p_{F, i}} - G_{p_{G, t}}) \times \mathbf{0}_3) \mathbf{I}_3 \mathbf{0}_3 \mathbf{0}_3 - \mathbf{I}_3
$$

$$
(L_R G_{KF_i} G_{p_{F, i}} \times) \mathbf{0}_3 \mathbf{0}_3 \cdots \mathbf{0}_3 \mathbf{0}_3 \cdots - L_R G_{KF_i}, \cdots)
$$

where the elements before $|$ are identical to the elements of (22), and the elements after $|$ are the Jacobians for the map-related variables. Specifically, the Jacobian of $y_{C_i}$ with respect to $K_{F_i}$ is $[0_3 \mathbf{0}_3]$, and the Jacobian of $y_{G_{KF, i}}$ with respect to $F_j$ is $-L_R \mathbf{g}_{G, t}$. The other omitted Jacobians are zeros. The detailed derivation is in Appendix B of the supplementary material [50].

The Jacobian matrix of (28) is given by

$$
\mathbf{H}_{G_{KF, i}} = \nabla h|_{q_{C_i}} (L_R G_{KF_i} (G_{p_{F, i}} - G_{p_{G, t}}) \times \mathbf{I}_3 \cdots \mathbf{I}_3)
$$

where $q_{C_i} \triangleq \nabla h|_{q_{C_i}} (L_R G_{KF_i} (G_{p_{F, i}} - G_{p_{G, t}}))$ and $\nabla h|_{q_{C_i}}$. Specifically, the Jacobian of $y_{G_{KF, i}}$ with respect to $K_{F_i}$ is $[-(G_{p_{F, i}} \times) I_3]$, and the Jacobian of $y_{G_{KF, i}}$ with respect to $F_j$ is $-I_3$. The other omitted Jacobians are zeros. The detailed derivation is in Appendix B of the supplementary material [50].

VI. OBSERVABILITY ANALYSIS OF MAP-BASED VIL SYSTEM

In this section, we theoretically analyze the observability properties of the perfect augmented system and the imperfect
augmented system from the perspectives of standard EKF and invariant EKF. Besides, intuitive explanations are also given.

To analyze the unobservable subspace of the augmented system, we need to compute the observability matrix of the augmented system, and the unobservable subspace is spanned by the right null space of the observability matrix.

For simplicity, the analysis below neglects the IMU bias, assumes the extrinsic between the camera and the IMU is an identity matrix, and sets the number of the local features, the map features, and the map keyframes to be one. The conclusions can be extended to general cases.

A. Observability of the Perfect Augmented System With Standard EKF

For standard EKF, the state is represented in the Euclidean vector space as

$$\mathbf{x}_{st} = \begin{bmatrix} L^T \mathbf{q}_{t_1} \mathbf{v}_{t_1}^T \mathbf{p}_{t_1}^T \mathbf{q}^T_{G1} \mathbf{p}^T_{G1} \end{bmatrix}^T$$ (31)

where $L^T \mathbf{q}_{t_1}$ is the quaternion of $L \mathbf{R}_{t_1}$, and $L^T \mathbf{q}_{G1}$ is the quaternion of $L \mathbf{R}_{G1}$.

Denoting the state transition matrix of the perfect augmented system from $t - 1$ to $t$ as $\Phi_{st|t-1}$, we have

$$\Phi_{st|t-1} \triangleq \Phi_{st|t-2} \cdots \Phi_{st|1|0}$$ (32)

where the left superscript $st$ represents that the matrix is derived from the standard EKF. Accordingly, denoting the Jacobian matrix of the local feature-based observation function at time step $t$ as $\mathbf{H}_{st|t}$, and that of the map-based observation function as $\mathbf{H}_{st|0}$, we have the observability matrix following [10] as

$$\mathbf{M}_{st} \triangleq \begin{bmatrix} \mathbf{M}_{st|0} \\ \mathbf{M}_{st|0} \\ \vdots \\ \mathbf{M}_{st|0} \\ \mathbf{M}_{st|1} \\ \mathbf{M}_{st|1} \\ \vdots \\ \mathbf{M}_{st|1} \\ \mathbf{M}_{st|1} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{st|0} \\ \mathbf{H}_{st|0} \\ \vdots \\ \mathbf{H}_{st|0} \\ \mathbf{H}_{st|1} \\ \mathbf{H}_{st|1} \\ \vdots \\ \mathbf{H}_{st|1} \\ \mathbf{H}_{st|1} \end{bmatrix} \mathbf{M}_{st_{10}}$$ (33)

We first assume that the Jacobian matrices are evaluated at the ground truth, which is ideal but can demonstrate the theoretical implication.

**Lemma 3**: (Ideal observability of perfect augmented system with standard EKF) The right null space $\mathbf{M}_{st_{10}}$ of the observability matrix $\mathbf{M}_{st}$, where the Jacobian matrices are evaluated at the ground truth, is spanned by four directions as

$$\mathbf{M}_{st_{10}} = \text{span} \begin{bmatrix} L^T \mathbf{R}_{t_0} \mathbf{g} \mathbf{0}_3 \\ -L^T \mathbf{v}_{t_0} \times \mathbf{g} \mathbf{0}_3 \\ -L^T \mathbf{p}_{t_0} \times \mathbf{g} \mathbf{I}_3 \\ -L^T \mathbf{D}_{t_0} \times \mathbf{g} \mathbf{I}_3 \\ L^T G \mathbf{R} \mathbf{g} \mathbf{0}_3 \\ -L^T G \mathbf{v} \times \mathbf{g} \mathbf{I}_3 \\ -L^T G \mathbf{p} \times \mathbf{g} \mathbf{I}_3 \\ -L^T G \mathbf{D} \times \mathbf{g} \mathbf{I}_3 \end{bmatrix}$$ (34)

**Proof**: See Appendix C of the supplementary material [50].

However, in practice, we have no access to the ground truth, so the Jacobian matrices can only be evaluated at some estimated points. This causes spurious information and breaks the original observability of the augmented system with standard EKF.

**Theorem 4**: (Real observability of perfect augmented system with standard EKF) The right null space $\mathbf{N}_{st_{2}}$ of the observability matrix $\mathbf{M}_{st}$, where the Jacobian matrices are evaluated at changing estimated values, is spanned by three directions as

$$\mathbf{N}_{st_{2}} = \text{span} \begin{bmatrix} 0_3 & 0_3 & \mathbf{I}_3 & \mathbf{I}_3 & 0_3 & 0_3 \end{bmatrix}^T$$ (35)

**Proof**: See Appendix C of the supplementary material [50].

B. Observability of the Perfect Augmented System With Invariant EKF

For invariant EKF, the state is represented as $\mathbf{X}_{st} = [\mathbf{q}^T, 0_1 \times 3, \mathbf{0}_1 \times 3, \mathbf{0}_1 \times 3, \mathbf{0}_1 \times 3]$.

**Theorem 5**: (Observability of perfect augmented system with invariant EKF) The right null space $\mathbf{N}_{in}$ of the observability matrix $\mathbf{M}_{in}$ is independent of the state values and spanned by four directions as

$$\mathbf{N}_{in} = \text{span} \begin{bmatrix} \mathbf{g}^T & 0_1 \times 3 & \mathbf{0}_1 \times 3 & \mathbf{0}_1 \times 3 & \mathbf{0}_1 \times 3 & \mathbf{g}^T \end{bmatrix}^T$$ (36)

**Proof**: See Appendix D of the supplementary material [50].

As $\mathbf{N}_{in}$ only consists of constant values, the unobservable subspace of the perfect augmented system will not be broken even if the state values are ever changing. Therefore, the correct observability of the perfect augmented system is preserved naturally with our proposed invariant EKF (cf. Section IV).

C. Observability of the Imperfect Augmented System With Standard EKF

As mentioned before, the map components always have more or less uncertainty. It is necessary to consider the uncertainty of the map components so that the system can be of good consistency.

For standard EKF, the state of the imperfect augmented system is defined as

$$\mathbf{x}_{st} = \begin{bmatrix} L^T \mathbf{q}_{t_1} \mathbf{v}_{t_1}^T \mathbf{p}_{t_1}^T \mathbf{q}^T_{G1} \mathbf{p}^T_{G1} \end{bmatrix} G^T \mathbf{q}_{KF} \mathbf{p}^T_{KF} \mathbf{p}^T_{KF}$$ (37)

where the elements after $|$ are map related. $G^T \mathbf{q}_{KF}$ and $G^T \mathbf{p}_{KF}$ form a map keyframe pose, and $\mathbf{p}^T_{KF}$ is a map feature position.

The state transition matrix, Jacobian matrix of the observation function, and the observability matrix derived from the standard EKF are denoted similarly to those in (32) and (33), except that the left superscript $st$ is replaced by in.

**Theorem 6**: (Ideal observability of imperfect augmented system with standard EKF) The right null space $\mathbf{N}_{st_{10}}$ of the observability matrix $\mathbf{M}_{st}$, where the Jacobian matrices are evaluated at the ground truth, is spanned by ten (four plus six)
perfect augmented system has the same unobservable dimensions as and map feature positions in $C$ or $I$ or with invariant EKF [cf. (33)]. Therefore, this 6-D unobservable subspace will be broken for the real case:

$$\mathcal{N}_1^* = \text{span} \begin{bmatrix} \mathbf{e}_3 & \mathbf{0}_{3 \times 3} \end{bmatrix}$$

$$\mathbf{1} = -(G \mathbf{p}_{KF}) \times (G \mathbf{R}_{KF}).$$

Proof: See Appendix E of the supplementary material [50].

The right null space $\mathcal{N}_1^*$ consists of two parts: the unobservable subspace of the original perfect augmented system (the first four columns) and the unobservable subspace introduced by the map-related variables (the last six columns). Furthermore, as the elements of the last six columns of $\mathcal{N}_1^*$ include ever-changing variables, the six dimensions of the unobservable subspace will be broken for the real case, which leads to the following theorem.

**Theorem 7 (Real observability of imperfect augmented system with standard EKF):** The right null space $\mathcal{N}_2^*$ of the observability matrix $\mathbf{M}_1^*$, where the Jacobian matrices are evaluated at changing estimated values, is spanned by ten (four plus six) directions as

$$\mathcal{N}_2^* = \text{span} \begin{bmatrix} \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{I}_3 & \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{0}_3 \end{bmatrix}^\top.$$

With the prior knowledge of Section VI-A, the proof is trivial and is omitted.

### D. Observability of the Imperfect Augmented System With Invariant EKF

The state of the imperfect augmented system with invariant EKF is given by (23), and the state error is also defined by (24) and (25). The state transition matrix, Jacobian matrix of the observation function, and the observability matrix derived from the invariant EKF are denoted similarly to those in (32) and (33).

**Theorem 8 (Ideal observability of imperfect augmented system with invariant EKF):** The right null space $\mathcal{N}_1^*$ of the observability matrix $\mathbf{M}_1^*$, where the Jacobian matrices are evaluated at the ground truth, is spanned by ten (four plus six) directions as

$$\mathcal{N}_1^* = \text{span} \begin{bmatrix} \mathbf{g} & \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_{3 \times 1} \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_{3 \times 1} \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_{3 \times 1} \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_{3 \times 1} \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_{3 \times 1} \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_{3 \times 1} \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_{3 \times 1} \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_{3 \times 1} \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_{3 \times 1} \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \end{bmatrix}.$$

Proof: See Appendix F of the supplementary material [50].

Unlike the unobservable subspace of the perfect augmented system with invariant EKF [cf. (36)], the elements of the last six columns of $\mathcal{N}_2^*$ include ever-changing variables, such as the standard EKF system in Section VI-C. Therefore, this 6-D unobservable subspace will be broken for the real case:

**Theorem 9 (Real observability of imperfect augmented system with invariant EKF):** The right null space $\mathcal{N}_2^*$ of the observability matrix $\mathbf{M}_2^*$, where the Jacobian matrices are evaluated at changing estimated values, is spanned by four directions as

$$\mathcal{N}_2^* = \text{span} \begin{bmatrix} \mathbf{g} & \mathbf{0}_{1 \times 3} & \mathbf{0}_{1 \times 3} & \mathbf{0}_{1 \times 3} & \mathbf{0}_{1 \times 3} & \mathbf{0}_{1 \times 3} \\ \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{I}_3 & \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{I}_3 & \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{I}_3 & \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{I}_3 & \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{I}_3 & \mathbf{I}_3 & \mathbf{0}_3 & \mathbf{0}_3 \end{bmatrix}^\top.$$

The proof of Theorem 9 is straightforward and is omitted.

### E. Intuitive Explanations Toward Observability

In this section, intuitive explanations toward the observability of the (imperfect augmented system) are given in order to give readers a better understanding of the unobservable subspace of the augmented systems.

As shown in Fig. 3, there are four kinds of frames, i.e., $C$, $L$, $G$, and $KF$. The orange dots represent the map features. The dotted-curved arrows represent the relative poses or positions between different frames (for example, the dotted-curved arrow between a map feature and frame $L$ means the position of the map feature in frame $G$). The dotted-straight lines represent the observation of $KF$ or $C$ toward map features. The black arrows/lines mean that they are fixed variables and should be independent of the transformation (the purple curved arrow) applied to the system, while the gray arrows represent the variables to be estimated and will be changed when applying a transformation to the system.

**Perfect augmented system:** For the ideal case, the perfect augmented system has the same unobservable dimensions as the standard visual-inertial system. This result is not surprising.
We can understand the proposed result by regarding the augmented variable as another 6-DoF pose feature in the odometry frame and never being marginalized. In this way, the perfect augmented system is equivalent to a visual-inertial SLAM system, demonstrating similar properties. Fig. 3(a) indicates the four unobservable dimensions of the perfect augmented system. When a 4-DoF transformation (one for the rotation along the gravity axis and three for the translation) is applied to frame $L$, the pose of $C$ and $G$ in frame $L$ can be adjusted [see the changed gray dotted-curved arrows in Fig. 3(a)] so that the observation of $C$ (the black dotted-straight lines) and the map feature positions (the black dotted-curved arrows) does not change. As the system is gravity aligned, the dimension of the applied transformation is four (three dimensions for translation, and one dimension for rotation along the gravity axis), which is the dimension of the unobservable subspace. A similar conclusion is found in [10], where Lee et al. argue that even with GPS-based global measurements, the unobservable subspace of the system (i.e., the perfect augmented system in this article) has four dimensions.

Following this idea, for the real case, when standard EKF is employed, the observability deficiency of the perfect augmented system would be the same as that of the visual-inertial SLAM system, i.e., along one direction (rotation around gravity direction) [25], which is consistent with Theorem 5. However, when invariant EKF is employed, the correct observability can be naturally maintained as indicated by Theorem 6, which is similar to other invariant EKF-based VIO systems [26, 27].

Imperfect augmented system: For the imperfect augmented system, the map-related parts can be treated as another visual SLAM system where keyframe poses and features are considered, and the dimension of the unobservable subspace of the visual SLAM system is six. From this point of view, it is easy to understand why the imperfect augmented system has another six unobservable directions compared with the perfect augmented system. Fig. 3(b) and (c) provides good insights toward this conclusion. In Fig. 3(b), we apply a 4-DoF transformation (one for the rotation along the gravity axis and three for the translation) to frame $L$ while keeping the observations of $K F$ and $C$ unchanged. In Fig. 3(c), a 6-DoF transformation is applied to frame $G$. By adjusting the pose of $K F$ and the position of map features in frame $G$, the observations of $K F$ and $C$ are unchanged.

VII. SCHMIDT-INARIANT KALMAN FILTER WITH MULTISTATE AND OBSERVABILITY CONSTRAINTS

From Sections IV and VI-B, we have developed an invariant EKF algorithm that is naturally consistent for the perfect augmented system. However, it is not enough for practical applications. On the one hand, as the map information is not processed in real-time, the robot pose at the current time step is insufficient because, in this way, informative historical data would fail to constrain the estimation of the current state—a sliding-window-based (multistate-constrained) technique is preferred to get more accurate estimates. From these two aspects, we propose an MSOC-S-IKF.

A. OC Schmidt-Invariant Filter for Imperfect Map

In this part, to integrate map information consistently and efficiently, a Schmidt filter is introduced to consider the uncertainty of map information efficiently, while an OC technique is proposed to maintain the correct observability of the imperfect augmented system.

Schmidt-invariant filter: To consider the uncertainty of map information, we need to add map-related variables into the state. If we put all the map features, the number of which could be tens of thousands, into the state of the augmented system, such as (23), it requires significant computation to update the state and large storage to record covariance. Therefore, we choose to keep the map keyframe poses and their observations toward features instead of saving computation and storage. Besides, by performing the null space projection to map features with (66), the uncertainty of map features can be taken into consideration.

Based on the above idea, the state of the augmented system $x_t$ in (7) will be augmented as follows:

$$X_{SCH} = (x_{An}, b_1, X_{KF})$$

where $X_{KF}$ contains the matched map keyframe poses \{${X_{KF_i}, t, \ldots, X_{KF_m}, t}$\} defined in (23).

For brevity, we divide (42) into two parts: the active part $X_{an} \triangleq (X_{An}, b_1)$ and the nuisance part $X_{na} \triangleq X_{KF}$. This partition is very useful for Schmidt EKF updating, as illustrated in the following.

When the current frame matches a map keyframe, we have the observation functions (27) and (28). Linearizing these two functions, we have

$$R_{Gi} = y_{Gi} - \hat{y}_{Gi} = -H_{ai} e_{ai} - H_{Fi} e_{Fi} + \gamma_{SCH}$$

where $e_{ai}$ and $e_{Fi}$ are the errors of the active part and the nuisance part, respectively, $e_{Fi}$ is the error of the map feature $F_i$, and $\gamma_{SCH}$ are the observation noises. The Jacobian matrices $H_{ai}$ and $H_{Fi}$ form the Jacobian matrix defined by (29). $H_{KF_i}$ and $H_{KF_i}$ form the Jacobian matrix defined by (30). To keep the correct observability of the imperfect augmented system, we employ an OC technique to modify the values of these observation Jacobian matrices. The details of the OC technique will be given later.

By stacking (43) and (44), and performing null space projection toward $F_j$, we have

$$r_{Gi} = -H_{ai} e_{ai} - H_{Fi} e_{Fi} + \gamma_{SCH}$$

$$N_{Fi} r_{Gi} = -N_{Fi} H_{ai} e_{ai} - N_{Fi} H_{Fi} e_{Fi} + N_{Fi} \gamma_{SCH}$$

$$r_{Gi}^{T} = -H_{ai} e_{ai} - H_{Fi} e_{Fi} + \gamma_{SCH}$$

$$= [H_{ai}^{T} H_{Fi}^{T}] e_{ai} + \gamma_{SCH}$$

$$\triangleq -H_{X_{SCH}} e_{X_{SCH}} + \gamma_{SCH}$$

(45)
where \( r^t_{Gi} \) is the stacking vector of \( r^C_{Gi} \) and \( r^{KF}_{Gi} \), \( \mathbf{H}_{a^t}, \mathbf{H}_{n^t}, \mathbf{H}_{F^t}, \), and \( \gamma'_{SCH} \) are the similar formulations, and \( \mathbf{N}_F \) is the left null space of \( \mathbf{H}_{F^t} \). Note that (45) only considers one map feature. In practice, at time step \( t \), there would be more than one map feature that can be observed. By stacking these observation functions together, we have

\[
\mathbf{r}^t_{Gi} = -\mathbf{H}'_{SCH} e'_{SCH} + \gamma'_{SCH}
\]

(46)

which is used to perform the Schmidt updating.

Since the state \( \mathbf{X}_{SCH} \) can be divided into \( \mathbf{X}_{SCH} = (\mathbf{X}_{a^t}, \mathbf{X}_{n^t}) \), its covariance can also be partitioned as

\[
\mathbf{P}_t = \begin{bmatrix} \mathbf{P}_{aa^t} & \mathbf{P}_{an^t} \\ \mathbf{P}_{na^t} & \mathbf{P}_{nn^t} \end{bmatrix}.
\]

(47)

Then, the following equations are employed to update the state, which is so-called the Schmidt updating:

\[
\begin{align*}
\mathbf{S}_t &= \mathbf{H}'_{SCH} \mathbf{P}_t \mathbf{H}'_{SCH} + \mathbf{V}'_{SCH} \\
\mathbf{K}_t &= \begin{bmatrix} \mathbf{K}_{X_{a^t}} \\ \mathbf{K}_{X_{n^t}} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{X_{a^t}} \\ \mathbf{K}_{X_{n^t}} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{P}_{aa^t} - \mathbf{H}'_{SCH} \mathbf{H}^{\dagger}_{SCH} \mathbf{P}_{aa^t} + \mathbf{P}_{an^t} - \mathbf{H}'_{SCH} \mathbf{H}^{\dagger}_{SCH} \mathbf{P}_{an^t} \\ \mathbf{P}_{na^t} - \mathbf{H}'_{SCH} \mathbf{H}^{\dagger}_{SCH} \mathbf{P}_{na^t} + \mathbf{P}_{nn^t} - \mathbf{H}'_{SCH} \mathbf{H}^{\dagger}_{SCH} \mathbf{P}_{nn^t} \end{bmatrix} \mathbf{S}_t^{-1} \\
\mathbf{P}_t &= \mathbf{P}_{t|-1} \\
\mathbf{X}_{a^t} &= \mathbf{X}_{a^t} - \begin{bmatrix} \mathbf{K}_{X_{a^t}} \mathbf{S}_t \mathbf{K}_{X_{a^t}}^{\dagger} \\ \mathbf{P}_{na^t} - \mathbf{H}'_{SCH} \mathbf{H}^{\dagger}_{SCH} \mathbf{K}_{X_{a^t}} \end{bmatrix} \\
\mathbf{X}_{n^t} &= \mathbf{X}_{n^t} - \begin{bmatrix} \mathbf{P}_{an^t} - \mathbf{H}'_{SCH} \mathbf{H}^{\dagger}_{SCH} \mathbf{P}_{na^t} \\ \mathbf{P}_{nn^t} - \mathbf{H}'_{SCH} \mathbf{H}^{\dagger}_{SCH} \mathbf{P}_{nn^t} \end{bmatrix} \mathbf{S}_t^{-1} \mathbf{K}_{X_{a^t}} \mathbf{S}_t^{-1}
\end{align*}
\]

(48)

(49)

(50)

(51)

(52)

where \( \mathbf{V}'_{SCH} \) is the covariance of the noise \( \gamma'_{SCH} \) in (46).

In summary, for the Schmidt updating, the active part is updated in the form of invariant EKF given in Algorithm 1, while the nuisance part (map keyframes) remains unchanged.

**Observability maintenance:** Reviewing Theorem 8 and Theorem 9, we find that due to the ever-changing estimated values of \( \mathbf{L}_G \) and \( \mathbf{G}_P \), the last six columns of (40) vanish for the real case. In the framework of the Schmidt updating, the value of \( \mathbf{G}_P \mathbf{F}_G \) stays the same, whereas the value of \( \mathbf{L}_G \mathbf{F}_G \) keeps changing. A simple and intuitive way to preserve the proper unobservable subspace of the system is defining the null space (40) based on the first-estimated value of \( \mathbf{L}_G \mathbf{F}_G \), i.e., \( \mathbf{L}_G \mathbf{F}_G \), and compute the Jacobian matrix (29) with \( \mathbf{L}_G \mathbf{F}_G \). However, as indicated by Kottas et al. [31], the performance of this solution relies heavily on \( \mathbf{L}_G \mathbf{F}_G \), because \( \mathbf{L}_G \mathbf{F}_G \) is used at all time steps to compute Jacobians. If \( \mathbf{L}_G \mathbf{F}_G \) is far from the true value, the system’s performance would be degraded. Therefore, we design an OC technique following the idea of Huang et al. [31] to search for the optimal values of Jacobians while satisfying the observability requirements.

The key idea of the OC technique is to select a suitable null space \( \mathbf{inH}_{3}^{C} \) of the observability matrix \( \mathbf{inM}^{C} \) and suitable Jacobians \( \mathbf{inH}_{L}^{C}, \mathbf{inH}_{G}^{C}, \mathbf{inF}_{j}^{C}, \) and \( \mathbf{inP}_{i}^{C} \) [cf. (33)] to fulfill the following conditions [32]:

\[
\begin{align*}
\mathbf{inH}_{L}^{C} \mathbf{inN}_{3}^{C} &= \mathbf{0} \\
\mathbf{inH}_{G}^{C} \mathbf{inN}_{3}^{C} &= \mathbf{0} \\
\mathbf{inN}_{3}^{C} &= \mathbf{inP}_{i}^{C} \mathbf{inN}_{3}^{C} \
\end{align*}
\]

(53)

(54)

(55)

A natural and realizable choice of \( \mathbf{inN}_{3}^{C} \) [cf. (40)] is given by

\[
\mathbf{inN}_{3}^{C} = \text{span} \begin{bmatrix} \mathbf{g} & \mathbf{0}_{3 \times 1} & \mathbf{0}_{3 \times 1} \\
\mathbf{0}_{3 \times 1} & \mathbf{I}_{3} & \mathbf{0}_{3 \times 1} \\
\mathbf{0}_{3 \times 1} & \mathbf{0}_{3 \times 1} & \mathbf{I}_{3} \end{bmatrix} \mathbf{G}_{P} \mathbf{F}_{G} \mathbf{R}_{G}^{\dagger}
\]

(56)

It is worth mentioning that the OC technique is a general and useful way to make the estimator consistent. Even for the standard EKF, as long as the conditions (53)–(55) are satisfied, the whole system can be consistent. However, we argue that a good estimator should introduce as few artificial-designed constraints as possible. Therefore, we still formulate the state error in the nonlinear form as (11) instead of the standard error form. In this way, the special OC design for the transformation matrix \( \mathbf{inP}_{i}^{C} \) and the local observation Jacobian matrix \( \mathbf{inH}_{L}^{C} \) can be avoided. Readers can verify that, with (20) and \( \mathbf{inP}_{i}^{C} \) (the detailed expression is given in Appendix F of the supplementary material[50]), (53) and (55) are naturally satisfied. Therefore, we only need to design an OC technique satisfying (54). Furthermore, since \( \mathbf{inH}_{G}^{C} \) consists of (29) and (30), and (30) automatically meets the condition (54), our problem is simplified to modifying (29) such that the linearization error of (27) is minimum while the observability of the imperfect augmented system is correctly maintained

\[
\min_{\mathbf{H}^{C}_{G}} \| \mathbf{H}^{C}_{G} - \mathbf{H}_{G}^{C} \|_{F}^{2}, \text{s.t. } \mathbf{H}^{C}_{G} \mathbf{inN}_{3}^{C} = \mathbf{0}
\]

(57)

where \( \cdot \) \( \cdot \) denotes the Frobenius matrix norm. \( \mathbf{H}^{C}_{G} \) is calculated by (29) with only one map feature and one map keyframe being considered. The optimal \( \mathbf{H}^{C}_{G} \) that fulfills (57) is given by

\[
\mathbf{H}^{C}_{G} = \mathbf{H}^{C}_{G} - \mathbf{H}^{C}_{G} \mathbf{inN}_{3}^{C} \mathbf{inN}_{3}^{C} \mathbf{inN}_{3}^{C} - \mathbf{inN}_{3}^{C}
\]

(58)

Replacing (29) with (58) while computing Jacobians, the proper unobservable subspace \( \mathbf{inN}_{3}^{C} \) of \( \mathbf{inM}^{C} \) is preserved.

**Complexity analysis:** Suppose a map has \( m \) keyframes and \( n \) features (\( m \ll n \)). In terms of storage, if all the map features are maintained in the state, we need to store the positions and the covariance matrices of these features, then the complexity of the storage will be \( \mathcal{O}(n + m^2) = \mathcal{O}(n^2) \). Instead, if the map keyframe poses are maintained as mentioned before, we need to store the poses and the covariance matrices of the map keyframes and the positions of the map features,\(^3\) then the complexity of the storage will be \( \mathcal{O}(m + m^2 + n) = \mathcal{O}(m^2 + n) \), which is much more storage-saving.

\(^3\)Note that in this case, we do not need to store the covariance matrices of the map features because the map features can be marginalized from the state by (45).
TABLE II

| Complexity of EKF/IKF Under Different Settings |
|-----------------------------------------------|
| **Update scheme to consider map uncertainty** |
| **Maintained map components in the state vector** | **Complexity** |
| Standard Schmidt features keysframes | Storage | Computation |
| ✔ | ✔ | ✔ | ∇ (n²) | O(n³) |
| ✔ | ✔ | ✔ | ∇ (n² + n) | O(n³) |
| ✔ | ✔ | ✔ | ∇ (n³) | O(n³) |
| ✔ | ✔ | ✔ | ∇ (n³ + n) | O(n³) |

- n/m: the number of map features/keyframes; "✔": be adopted.
- "-": do not consider the map uncertainty or maintain the map.
- "constant": the complexity is constant compared with O(m).

In terms of computation, the main difference between the standard (invariant) EKF and the Schmidt (invariant) EKF is the covariance update (50). For the standard (invariant) EKF, 0 in (50) is replaced by X_{n_{s}}S_{i}X_{n_{s}}^{-1}, whose computational complexity is quadratic of the size of the nuance part (O(n²)). This limits the real-time performance of the augmented system. In contrast, the computational complexity of (50) is linear (O(m)).

The complexity differences among the different settings are summarized in Table II for better readability. The blue shade highlights the setting we used in our proposed algorithm.

B. Multistate-Constrained-Invariant Filter

The multistate-constrained filter is designed to improve the performance of VIO [1]. In our system, it is applied to the local feature-based observation function [cf. (19)].

**State with multistate constraint:** Suppose we keep a sliding window with the size of s, where the historical IMU poses are saved. Then, we augment the state X_t defined by (7) with the cloned poses as

\[
X_{MSC_t} = (X_{A_t}, \mathbf{B}_t, \mathbf{T}_{C_t}, X_{Clone_t})
\]

where \( \mathbf{T}_{C_t} = [\mathbf{t}^T \mathbf{R}_{C_t}, \mathbf{p}_{C_t}] \) is the extrinsic between the camera and the IMU. We estimate this variable online for better localization results. X_{Clone_t} contains s IMU poses \{X_{CP_1}, \ldots, X_{CP_{s+1}} \}. For each cloned IMU pose \{X_{CP_i} \}, it consists of \( \{\mathbf{t}^T \mathbf{R}_{C_t}, \mathbf{p}_{C_t}\} \). Following the definition of (11), the *right-invariant error* of \( \mathbf{T}_{C_t} \) and \( X_{Clone_t} \) is given as follows:

\[
\begin{align*}
\xi_{p_{ci}} &= \mathbf{t}^T \mathbf{R}_{C_t} - (I_3 + (\mathbf{t}^T \mathbf{R}_{C_t})_x)^T \mathbf{p}_{C_t}, \\
\xi_{p_{Li}} &= \mathbf{t}^T \mathbf{R}_{C_t} - (I_3 + (\mathbf{t}^T \mathbf{R}_{C_t})_x)^T \mathbf{p}_{Li_t},
\end{align*}
\]

Observation function with multistate constraint: Suppose there is a local feature f observed by multiple cloned poses. For each cloned pose, we formulate the observation function as

\[
y_{MSC} = h(\mathbf{t}^T \mathbf{R}_{C_t}, \mathbf{p}_{C_t}) = \mathbf{t}^T \mathbf{R}_{C_t} - (\mathbf{t}^T \mathbf{R}_{C_t})_x (\mathbf{p}_{C_t} + \mathbf{L}_t \mathbf{p}_{Li_t}),
\]

where \( y_{MSC} \) is the measurement of feature \( f \) in the image captured by \( X_{CP_i} \), and \( \gamma_{MSC} \) is the observation noise. Linearizing the above function, such as (18), we get

\[
r_{MSC} = y_{MSC} - \mathbf{H}_{MSC} f + \gamma_{MSC}
\]

where \( r_{MSC} \) is the reprojection error of \( f \) under \( X_{CP_i} \), \( \mathbf{H}_{MSC} \) is the Jacobian of the observation function with respect to all related variables in \( X_{MSC} \). \( \mathbf{H}_f \) is the Jacobian matrix of the observation function with respect to the observed feature \( f \). \( \epsilon_{MSC} \) and \( \epsilon_f \) are the *right-invariant error* of the related state and the feature, respectively.

Noting that in (62), although the error of state, \( \epsilon_{MSC} \), is related to different cloned poses, the error of the same feature \( f \), \( \epsilon_f \) should be identical. Recall (11) that the feature error is defined as \( \epsilon_f = \mathbf{t}^T \mathbf{R}_{f} - R^T_{f} \mathbf{p}_{f} \), where \( R \) is a rotation matrix. This means that we need to bind a rotation matrix to each feature’s error. In this article, we assume that this rotation matrix is the rotation part of the current IMU cloned pose \( X_{CP_i} \), \( \mathbf{t}^T \mathbf{R}_{C_t}, \mathbf{p}_{C_t} \), and we call this current IMU cloned pose as the *anchored frame*. Then, there are two cases to be considered.

1) The feature \( f \) is observed by the anchored frame \( X_{CP_i} \),

\[
r_{MSC} = y_{MSC} - \mathbf{H}_{MSC} f + \gamma_{MSC}
\]

\[
\begin{align*}
\mathbf{r}_{MSC} &= \mathbf{y}_{MSC} - \hat{\mathbf{y}}_{MSC} \\
&= \nabla h(\mathbf{t}^T \mathbf{R}_{C_t}, \mathbf{p}_{C_t})^T \left[ -\xi_f + \mathbf{t}^T \mathbf{R}_{C_t} \xi_{p_{Ci}} + \xi_{p_{Li}},
\right. \\
&\left. + (\mathbf{t}^T \mathbf{R}_{C_t} - \mathbf{t}^T \mathbf{R}_{f}) \times (\mathbf{t}^T \mathbf{R}_{C_t} \xi_{p_{Ci}}) \right] + \gamma_{MSC}.
\end{align*}
\]

2) The feature \( f \) is observed by the other cloned frame \( X_{CP_i} \), where \( i = t, \ldots, s + 1 \).

\[
r_{MSC} = y_{MSC} - \mathbf{H}_{MSC} f + \gamma_{MSC}
\]

\[
\begin{align*}
\mathbf{r}_{MSC} &= \mathbf{y}_{MSC} - \mathbf{H}_{MSC} f + \gamma_{MSC} \\
&= \nabla h(\mathbf{t}^T \mathbf{R}_{C_t}, \mathbf{p}_{C_t}, \mathbf{t}^T \mathbf{R}_{f}, \mathbf{p}_{f})^T \left[ -\xi_f + \mathbf{t}^T \mathbf{R}_{C_t} \xi_{p_{Ci}} + \xi_{p_{Li}} 
\right. \\
&\left. + (\mathbf{t}^T \mathbf{R}_{C_t} - \mathbf{t}^T \mathbf{R}_{f}) \times (\mathbf{t}^T \mathbf{R}_{C_t} \xi_{p_{Ci}}) \right] + \gamma_{MSC}.
\end{align*}
\]

Stacking (63) and (64), we have the following expression:

\[
\begin{align*}
\mathbf{r}_{MSC} &= \mathbf{y}_{MSC} - \mathbf{H}_{MSC} f + \gamma_{MSC} \\
&= \nabla h(\mathbf{t}^T \mathbf{R}_{C_t}, \mathbf{p}_{C_t}, \mathbf{t}^T \mathbf{R}_{f}, \mathbf{p}_{f})^T \left[ -\xi_f + \mathbf{t}^T \mathbf{R}_{C_t} \xi_{p_{Ci}} + \xi_{p_{Li}} 
\right. \\
&\left. + (\mathbf{t}^T \mathbf{R}_{C_t} - \mathbf{t}^T \mathbf{R}_{f}) \times (\mathbf{t}^T \mathbf{R}_{C_t} \xi_{p_{Ci}}) \right] + \gamma_{MSC}.
\end{align*}
\]

Following the method in [1], we do not maintain the feature \( f \) in the state vector. Instead, we perform null space projection to merge the uncertainty of the feature into the uncertainty of
the maintained state as
\[ r_{\text{MSC}} = -H_{\text{MSC}} \epsilon_{\text{MSC}} - H_f \epsilon_f + \gamma_{\text{MSC}} \]
\[ N_f r_{\text{MSC},i} = -N_f H_{\text{MSC}} \epsilon_{\text{MSC}} - N_f H_f \epsilon_f + N_f \gamma_{\text{MSC}} \]
\[ r_{i,\text{MSC}}' = -H_{\text{MSC}} \epsilon_{\text{MSC}} + \gamma_{i,\text{MSC}}' \]
where \( N_f \) is the left null space of \( H_f \). With (66), we perform the invariant EKF update.

We stress that although cloned poses and the extrinsic are introduced into the state, the observability of the augmented system remains unchanged, and the dimension of the unobservable subspace is still four. Readers can verify it following the definition of the observability matrix (33).

C. Initialization of the Augmented Variable

As is shown in Fig. 1, when there are feature matches between the map keyframe \( KF \) and the current frame \( C \), we utilize the matched features’ 3-D positions in \( G \) combined with the 2-D observations in \( C \) to perform the Perspective-N-Points (PnP [40]) algorithm. Then, an estimate of the transformation between \( G \) and \( C \) is acquired. PnP is essentially an optimization problem. By solving this problem, we can get the optimal estimated \( GT_C \), and its information (covariance) matrix. Finally, the initial value of \( LT_G \) is given as \( LT_G = L T_{G,i} L T_{C,i} G T_{C,i}^{-1} \). Since all the variables and their covariance matrices on the right-hand side of this function are known, we can analytically derive the covariance of \( LT_G \), with the method mentioned in [49].

Moreover, according to our tests given in Appendix H of the supplementary material [50], assigning a large value to the initial covariance of \( LT_G \) is also a simple and safe strategy in practice.

VIII. EXPERIMENTAL RESULTS

In this section, we perform extensive experiments with simulated and real-world data. Before introducing the experiments in detail, the evaluation metrics are given first.

We utilize the evaluation tools of Open-VINS [2] to evaluate the performance of different algorithms. The accuracy of the localization can be measured by root-mean-squared error (RMSE) or absolute trajectory error (ATE), the drift of the localization can be reflected by relative pose error (RPE) [2], [41], and the consistency of the system can be evaluated by normalized estimation error squared (NEES). For a given dataset with \( N \) runs of the same algorithm and \( K \) time steps for each run, we define the estimated state value at time step \( k \) of the \( i \)th run as \( \hat{x}_{k,i} \), its corresponding true value as \( x_{k,i} \), and its corresponding covariance matrix as \( P_{k,i} \). Following the article presented in [2], the detailed definitions of these four metrics are as follows:

RMSE:
\[ \epsilon_{\text{rmse},k} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |x_{k,i} \sqcap \hat{x}_{k,i}|^2} \]
\[ \epsilon_{\text{rmse}} = \frac{1}{K} \sum_{k=1}^{K} \epsilon_{\text{rmse},k} \]  

ATE:
\[ \epsilon_{\text{ate}} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\frac{1}{K} \sum_{k=1}^{K} |x_{k,i} \sqcap \hat{x}_{k,i}|^2} \]  

RPE:
\[ \epsilon_{\text{rpe},i} = \frac{1}{N \times S} \sum_{i=1}^{N} \sum_{s=1}^{S} |\hat{x}_{s,i} - x_{s,i}^*| \]
\[ \epsilon_{\text{rpe}} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\frac{1}{K} \sum_{k=1}^{K} |\hat{x}_{k,i} - x_{k,i}^*|^2} \]

where we assume that the trajectory is split into \( S \) segments, and each segment has the length of \( l \). For each segment, it corresponds to a pair of state \( \{x_{s,i}^*, \hat{x}_{s,i}\} \) for \( s = 1, \ldots, S \), where \( x_{s,i}^* \) and \( \hat{x}_{s,i} \) mean the beginning and the end states of this segment, respectively.

NEES:
\[ \epsilon_{\text{nees},k} = \frac{1}{N \times d} \sum_{i=1}^{N} (x_{k,i} \sqcap \hat{x}_{k,i})^T P_{k,i}^{-1} (x_{k,i} \sqcap \hat{x}_{k,i}) \]
\[ \epsilon_{\text{nees}} = \frac{1}{K} \sum_{k=1}^{K} \epsilon_{\text{nees},k} \]

This metric indicates the consistency of the system. If the system is consistent, NEES should be approximately 1 for large \( N \). If NEES is much larger than 1, then the system underestimates the covariance of the state. It is worth noting that, to compute NEES, the estimated covariance derived from invariant-related algorithms corresponds to the nonlinear error defined in (11). However, for a fair comparison of all algorithms, RMSE, ATE, and RPE are computed by the standard vector error for the translation part and the nonlinear error defined in (11) for the rotation part.

A. Simulations

In this part, we make some adaptations to the simulator of Open-VINS [2] to evaluate the accuracy and consistency of different algorithms. We feed the simulator with a ground truth trajectory shaped like a “saddle” (cf. the orange trajectory of Fig. 4). The map keyframe poses come from another trajectory, which has a similar shape as the previous trajectory (cf. the blue trajectory of Fig. 4). These two trajectories are denoted as S1 (the blue one) and S2 (the orange one), respectively. S1 is utilized to generate the map-related information, while S2 is used for the robot to track. Note that the reference frame of S1 (map reference frame) is different from that of S2 (local VIO reference frame). The relative transformation between these two frames (i.e., the augmented variable \( LT_G \)) is one of the variables to be estimated. In the simulation, the robot will repeat the orange trajectory ten times for each run. So, the total running distance is around 10 \times 128 m = 1280 m. The locations where feature matches between the map and the running VIO occur are also marked in Fig. 4 by black asterisks (20 locations for each loop). During the ten loops of each run, the feature matches only occur in the second to fourth loop and the ninth to tenth loop (100 locations) such that there is a long-term absence of the map information from the fifth to the eighth loop.

To demonstrate the necessity of considering the uncertainty of the map, we conduct simulation experiments on a perfect map and an imperfect map. For the perfect map, we use the
ground truth of map keyframe poses. For the imperfect map, we artificially add noises to the ground truth of S1 to simulate the imperfect but real map keyframes. The position of the ground truth is perturbed by the Gaussian white noise \( \gamma_{tp} \sim N(0, \sigma_p^2 I_{3\times3}) \), \( \sigma_p = 0.1 \) m, and the orientation is perturbed by \( \gamma_o \sim N(0, \sigma_o^2 I_{3\times3}) \), \( \sigma_o = 0.9^\circ \). After the perturbation, the RMSE of the map trajectory S1 is 0.172 m/1.566°.

For the map-matching information, we first randomly generate 3-D features, then utilize the map-based observation functions (27) and (28) to reproject 3-D features into the frames obtained from the running trajectory (S2) and the perfect map keyframes so that the 2D–2D matching features are obtained. After that, we add Gaussian white noises to the 2-D features. Finally, for each 3-D map feature, we utilize the noisy 2-D features and the noisy map keyframe poses to triangulate and optimize its 3-D position as the estimated 3-D feature. Note that for the perfect map, we directly use the ground truth of the map keyframe poses to reproject 3-D features into the frames.

To test the consistency of our proposed algorithm (MSOC-S-IKF), we compare it with the Schmidt EKF version (MSC-S-EKF) and the ones without considering the uncertainty of the map (MSC-IKF and MSC-EKF). We regard Open-VINS [2] as a pure VIO baseline method to show that the consistent map-based algorithm is capable of effectively alleviating the drift of odometry. The overall results of the five algorithms with a perfect map and an imperfect map are given in Table III and Table IV, respectively. As MSC-S-EKF and MSOC-S-IKF require uncertainties of the map keyframe poses, for the perfect map, we assign a small uncertainty for each map keyframe pose: the standard deviation of the position and orientation parts are 0.1 mm and 0.01°, respectively. In the tables, three types of results are given: the pose of the robot in the local reference system (local-pose), the relative transformation between the local reference system and the map reference system (relative-trans), and the pose of the robot in the map reference system (map-pose). For each type, its RMSE and NEES of orientation (°)/position (m) are given. All the results are the average of results from 20 Monte Carlo simulations. The seed used to generate the Gaussian noises is different for each run. It should be noted that as Open-VINS is a VIO, it does not estimate the relative-trans; therefore, only the results of the local-pose are given. Besides, the other four algorithms estimate the local-pose and the relative-trans instead of the map-pose. The results of map-pose are derived from the estimated local-pose and relative-trans. Therefore, we only give the NEESs of the local-pose and the relative-trans to demonstrate the consistency of the algorithms.

### Trajectory accuracy
Table III reveals that for the perfect map, both MSC-IKF and MSOC-S-IKF have good performance in local-pose and relative-trans. This is because these two algorithms correctly maintain the observability of the perfect augmented system (cf. Theorem 5). However, for the imperfect map, as shown in Table IV, the situation is different, demonstrating the necessity of considering the uncertainty of map information. Moreover, readers may find that the inconsistent algorithms (MSC-EKF and MSC-S-EKF) are able to produce competitive results for map-pose. This is due to the fact that the map-based observation function (21) provides the constraint of relative transformation between frame I and frame G, i.e., the map-pose. Therefore, the results of map-pose \((G_T I_L)\) can be good even, although the results of local-pose \((L_T I_L)\) and relative-trans \((L_T G)\) are inaccurate and inconsistent. Nevertheless, although the map-based observation function constrains the map-pose, if the algorithm does not consider the uncertainty of map information, the accuracy of the map-pose will be poor (cf. the map-pose of MSC-EKF in Tables III and IV).

From Table IV, we can find that MSOC-S-IKF has the minimum RMSE for local-pose, relative-trans, and map-pose. It is worth noting that for the augmented system, MSOC-S-IKF online estimates local-pose and relative-trans instead of map-pose. Therefore, the performance in estimating local-pose and relative-trans is what we are concerned about. And listing the

### Table III
**RMSEs (°/m) and NEESs of Different Algorithms on Simulation Data With a Perfect Map**

| Algorithms          | Local-pose | Relative-trans | Map-pose |
|---------------------|------------|----------------|----------|
| Open-VINS          | RMSE: 0.247/0.132 | —              | —        |
| MSC-EKF            | RMSE: 1.152/0.338 | 1.204/0.190   | 0.101/0.042 |
| MSC-S-EKF          | RMSE: 1.106/0.510 | 1.156/0.197   | 0.099/0.049 |
| MSC-IKF            | RMSE: 0.172/0.106 | 0.153/0.108   | 0.100/0.043 |
| MSOC-S-IKF         | RMSE: 0.165/0.106 | 0.139/0.099   | 0.097/0.048 |

— means there is no such variable to evaluate.

The bolded elements indicate the best performance.

### Table IV
**RMSEs (°/m) and NEESs of Different Algorithms on Simulation Data With an Imperfect Map**

| Algorithms          | Local-pose | Relative-trans | Map-pose |
|---------------------|------------|----------------|----------|
| Open-VINS          | RMSE: 0.233/0.144 | —              | —        |
| MSC-EKF            | RMSE: 1.163/0.531 | 1.799/0.854   | 0.708/0.429 |
| MSC-S-EKF          | RMSE: 1.197/0.545 | 1.304/0.233   | 0.176/0.089 |
| MSC-IKF            | RMSE: 0.230/0.160 | 0.702/0.705   | 0.766/0.430 |
| MSOC-S-IKF         | RMSE: 0.176/0.126 | 0.185/0.132   | 0.175/0.088 |

— means there is no such variable to evaluate.

The bolded elements indicate the best performance.
results of map-pose is just for illustrating that with a good estimation of local-pose and relative-trans, we can get a competitive or even the best result of map-pose. In contrast, MSC-IKF has poor performance, even though it correctly preserves the observability of the system. This is because MSC-IKF does not consider the uncertainty of the map. Similarly, MSC-EKF, which neither considers the uncertainty of the map nor correctly keeps the observability of the system, and MSC-S-EKF, which does not correctly preserve the observability of the system, have (much) worse performance than MSOC-S-IKF, which not only considers the uncertainty of the map but keep the correct observability of the system. Moreover, Open-VINS, a consistent solution for VIO, produces the accurate local-pose because it consistently fuses the IMU measurements and the local feature observations. But contrast, MSOC-S-IKF, a consistent solution for VIL, is capable of producing more accurate local-pose because it additionally fuses the map consistently. Fig. 5 gives an intuitive comparison among MSC-EKF, MSC-S-EKF, MSC-IKF, and MSOC-S-IKF for the local trajectory on the $x$–$y$ plane with the imperfect map. The zoom-in area of Fig. 5 shows that the trajectory derived from MSOC-S-IKF (the green one) is the closest to the ground truth.

Consistency: To show the consistency properties of the different algorithms, the NEESs for local-pose and relative-trans are also given in Tables III and IV. When the map is perfect, the NEES values of both MSC-IKF and MSOC-S-IKF are around 1, which indicates the good consistency of these two algorithms. However, when the map is imperfect, MSC-EKF has large values of NEES for relative-trans because it neglects the uncertainty of the map, whereas MSOC-S-IKF has good consistency property (the NEES values are around 1). For both situations, MSC-EKF and MSC-S-EKF have large NEES values, which indicates that the systems are overconfident in their uncertainty estimates.

Fig. 6 shows the local-pose error with $3\sigma$ bounds from MSC-S-EKF, MSC-IKF, and MSOC-S-IKF. Fig. 7 gives the NEES of local-pose and relative-trans over timestamps. Both figures are under the situation that the map is imperfect. For MSC-S-EKF, since the unobservable subspace is collapsed (cf. Theorems 4 and 7), its estimation shows a trend of divergence. For MSC-IKF, although it correctly maintains the observability of the system (cf. Theorem 5), the uncertainty of map information is not considered, which makes the estimator over-reliant on the imperfect map information and leads to inaccurate and inconsistent results.

In this part, we will validate our proposed algorithm in four kinds of real-world datasets, which cover the scenarios of an aerial vehicle (EuRoC [44]), ground vehicles in urban areas (Kaist [45], 4Seasons [46]), and on campus (YQ [48]). All the experiments are run on a device with an Intel i7-10700@2.9 GHZ CPU and 16 Gb RAM.

EuRoC: EuRoC [44] is a visual-inertial dataset collected onboard a microaerial vehicle. This dataset contains three scenarios: vicon room 1 (V1), vicon room 2 (V2), and machine hall (MH). For our experiments, the sequences V101, V201, and MH01 are used to build the map, while the sequences V102-V103, V202-V203, and MH02-MH05 are used for localization. The map contains keyframes and 3-D features. As shown in dataset documents, the ground truth poses have measurement errors in millimeters, so we regard the ground truth poses as the noisy map keyframes with the standard deviations of 1 cm and $1\degree$.

B. Real-World Experiments

The length of the trajectory is $10.67$ km. The map keyframes and features are calculated following that in EuRoC part. As the ground truth in this dataset is built upon virtual reference station–GPS, we set the standard deviation of map keyframe poses as $0.1$ m for the position part and $2.87\degree$ ($0.05\ rad$) for the orientation part.
**4Seasons:** 4Seasons [46] is a dataset collected in different scenarios and under various weather conditions and illuminations. This dataset is for the research on VIO, global place recognition, and map-based relocalization, suitable to test our proposed algorithm. In this dataset, we select the first two sequences (2020-03-24_17-36-22, denoted as Office-Loop1, 2020-03-24_17-45-31, denoted as Office-Loop2) from the collection called “Office Loop,” where the vehicle loops around an industrial area of the city [cf. Fig. 8(e)]. We pick these two sequences because the vehicle at the beginning of the sequences is static, which is necessary for the IMU initialization of our system. Besides, due to lots of overlaps of their trajectories and similar illuminations, these two sequences are easier to detect matching images. We employ Office-Loop-1 to build the map and Office-Loop-2 for localization testing. The map keyframes and features are generated following the procedure mentioned in EuRoC part. As the ground truth in this dataset is built upon the real-time kinematic global navigation satellite system that provides global positioning with up-to-centimeter accuracy [46], and the standard deviation of the map keyframe pose is set the same as that in Kaist.

**YQ:** YQ [47], [48] is a dataset collected by ourselves in Yuquan Campus, Zhejiang University, China, by a four-wheel ground vehicle. This dataset aims at testing in off-the-road scenarios and under various weather conditions and illuminations. It contains four sequences, YQ1–YQ4, where YQ1–YQ3 were recorded on three separate days with different weathers in summer, and YQ4 was collected in winter after snowing [cf. Fig. 8(d)]. The changing weather and season severely degenerate the feature matching, causing the long absence of map-based measurements. For this dataset, we use YQ1 to build the map and YQ2–YQ4 to test the performance of the algorithms. The standard deviation of the map keyframe pose is set the same as that in Kaist.

**Feature matching:** For all real-world datasets, the matching procedure is conducted as follows: we first utilize R2D2 [42] to extract new features on the current query frame and match them with features in map keyframes. The initial matching pairs based on the feature descriptors are then fed into a robust pose estimator in [43] to generate accurate feature-matching pairs.

**Benchmark with comparative methods:** In this part, we make comparisons between our proposed methods with the benchmark from Open-VINS [2] and VINS-Fusion [6], [9] on EuRoC, Kaist, 4Seasons, and YQ. We regard Open-VINS as a pure odometry baseline with drift and validate the correction of map-based measurements. For VINS-Fusion, its localization mode is used. We keep its map settings the same as the ones in our method. The only difference is that VINS-Fusion ignores the uncertainty of the map, so we do not set the map covariance for VINS-Fusion. We record the localization result of VINS-Fusion by concatenating the odometry $\hat{T}_k$ and the estimated $\hat{f}_k$ instead of the optimized trajectory due to the real-time causality.

**Trajectory accuracy:** Table V lists the ATE results of position (m) in the local frame (local position) and in the map frame (map position), which are derived from different algorithms on different dataset sequences. The ATE results of the local position are derived by aligning the local trajectories with the ground truths, whereas for the map position, we directly compare the computed trajectories in the map frame with the ground truths without alignments. All the results are the average of three runs. As indicated by Table V, our consistent algorithm MSOC-S-IKF has great (or the best for most cases) performance for both local and map (global) positions, whereas VINS-Fusion, MSC-EKF, and MSC-IKF fail to survive in the long-time and large-scale scene (YQ, Kaist, and 4seasons). VINS-Fusion diverges due to the significant drifts and the overconfident belief in the noisy map, while MSC-EKF and MSC-IKF diverge mainly due to ignoring the uncertainty of map information, so the estimators use incorrect covariance matrices leading to wrong state updating. It is worth mentioning that although MSC-S-EKF, which does not maintain the correct unobservable subspace, is able to successfully run through all of these dataset sequences and sometimes has competitive results for some evaluated variables (e.g., map position of YQ), the results of the variable (local-position) that is actually estimated by the estimator turn out to be poor. For some sequences, the accuracy of MSC-S-EKF of local-pose is worse than the pure odometry (Open-VINS), such as V203, MH04, Urban39, and Offie-Loop2. Besides, the local trajectories of MSC-S-EKF and MSOC-S-IKF are plotted in Fig. 9, where the two trajectories are aligned with the ground truth by the first pose such that we can see how the local trajectories change over time. Fig. 9 shows that the trajectory of MSC-S-EKF wildly deviates from the ground truth after a short-term running. This phenomenon stems from inconsistent estimation, which introduces spurious information to the system and degrades the performance of the estimator even though sometimes the results of map position (derived though the estimated local pose and the augmented variable) are acceptable.

In particular, there is an interesting phenomenon from Table V for EuRoC datasets, the values of the map keyframe poses are set to be the same as the ground truth, and the averaged reprojection error of triangulated 3-D features is around 0.45 pixels, which means that the prebuild maps have pretty good accuracy. MSC-IKF, which is able to naturally maintain
the correct observability of the augmented system, is still inferior to MSOC-S-IKF in the vast majority of cases. This demonstrates the necessity to consider the uncertainty of map information.

To show the drift of the localization and the smoothness of the trajectory, in Fig. 10, we give the box plots of RPEs of Open-VINS, MSC-S-EKF, and MSOC-S-IKF for the position in both local and map reference frames. As defined in (70), we select the segment length $l$ as 100 m, 200 m, and 500 m. In Fig. 10, with the segment length increasing, both local and global position errors increase, indicating the existence of the drift. However, as MSOC-S-IKF consistently fuses the map information, the drift is alleviated, which is illustrated by comparing the local position errors of Open-VINS and MSOC-S-IKF. On the contrary, MSC-S-EKF has apparent drift, and the local RPE for each segment length is also large. This is because MSC-S-EKF is inconsistent, and spurious information is introduced to the system. Moreover,
by comparing MSOC-S-IKF and MSC-S-EKF, even though the global RPE of MSC-S-EKF is sometimes competitive, the local RPE of MSC-S-EKF is much larger than that of MSOC-S-IKF, indicating that the smoothness of the trajectory of MSOC-S-IKF in the local reference system is much better than that of MSC-S-EKF. This conclusion can also be illustrated in Fig. 9, where the trajectory of MSC-S-EKF is severely distorted.

**Real-time efficiency:** The computing efficiency of our proposed algorithm is also validated. The average time consumptions per step of VINS-Fusion and MSC-S-IKF on EuRoC sequences are given in Fig. 11, with MSC-EKF and Open-VINS as baselines. Note that what we validate is the efficiency of the proposed back-end localization algorithm. Therefore, the time consumption of the front-end, including feature extraction and feature matching, is not counted for the compared algorithms. The results indicate that our proposed filter-based solution (MSOC-S-IKF) is more time-saving than VINS-Fusion, and, thus, is more appropriate for the on-board deployment.

**C. Discussion**

Through the simulations and the real-world experiments, we validate that MSC-IKF is consistent for the perfect augmented system but inconsistent for the imperfect augmented system due to ignoring the uncertainty of the imperfect map. On the contrary, because MSOC-S-IKF not only correctly maintains the observability of the imperfect augmented system but also considers the uncertainty of the map, the trajectory accuracy in both the local odometry frame and the map frame is the best in the vast majority of situations.

From the results of simulation and experiments, there are some interesting phenomena.

1) Consistently fusing the map information can improve the local localization accuracy of VIO, whereas inconsistently fusing the map information can even deteriorate the local localization accuracy of VIO.

2) According to Table V, in the small-scale scenes (V101–V203), the algorithms that do not consider the uncertainty of map information (MSC-EKF, MSC-IKF, and VINS-Fusion) can have moderate or even good results. However, as the experiment scene gets larger (MH02-Urban39), the results show the necessity of considering the map uncertainty—the algorithms that regard the map information as perfect produce poor results or even can hardly survive during the experiments. On the contrary, the algorithms that consider the uncertainty of map information (MSC-S-EKF and MSOC-S-IKF) can complete each dataset sequence.

3) According to Tables IV, V, and Fig. 10, MSC-S-EKF sometimes produces competitive results for “pose in map frame,” even though this algorithm is in fact inconsistent. This can be explained by the fact that the map-based observation function (21) provides the constraint of relative transformation between the robot frame $I_i$ and the map frame $G$. Therefore, the results of the robot pose in map frame $(G T_{I_i} = L T_{G} T_{I_i})$ can sometimes be good, even though the results of “local-pose” $(L T_{I_i})$ and “relative-trans” $(L T_{C})$ are inaccurate and inconsistent. However, the inaccurate “local-pose” will make this algorithm impractical. As shown in Fig. 9, MSC-S-EKF has poor local trajectory accuracy and smoothness, which will make subsequent tasks, such as trajectory planning, fail. Therefore, it is necessary for an algorithm to produce good localization accuracy in both local and global frames, and our proposed consistent algorithm, MSOC-S-IKF, can give promising results.

**IX. CONCLUSION**

In this article, a general situation of map-based VIL was considered, where the prebuilt map was imperfect and gravity unaligned. The map-based VIL problem was formulated as estimating both the local VIO and the augmented variable simultaneously. Based on the proposed $SE(3)$ Lie group, the Schmidt filter, and the multistate OC technique, we derived a consistent and efficient map-based VIL algorithm, MSOC-S-IKF. Through observability-based analysis and computational complexity analysis, we theoretically demonstrated the consistency and efficiency of our proposed algorithm, respectively. The simulations and real-data experiments further validated the correctness of the theory and the effectiveness of our proposed algorithm. Future works include extending MSOC-S-IKF to a distributed system for multirobot localization. Besides, how to add loop closure into this algorithm consistently and efficiently will also be our future research direction.

**REFERENCES**

[1] A. I. Mourikis and S. I. Roumeliotis, “A multi-state constraint Kalman filter for vision-aided inertial navigation,” in Proc. IEEE Int. Conf. Robot. Autom., 2007, pp. 3565–3572, doi: 10.1109/ROBOT.2007.364024.

[2] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, “OpenVINS: A research platform for visual-inertial estimation,” in Proc. IEEE Int. Conf. Robot. Autom., 2020, pp. 4666–4672, doi: 10.1109/ICRA40945.2020.9196524.

[3] K. Sun et al., “Robust stereo visual inertial odometry for fast autonomous flight,” IEEE Robot. Autom. Lett., vol. 3, no. 2, pp. 965–972, Apr. 2018, doi: 10.1109/LRA.2018.2793349.

[4] M. Brossard, S. Bonnabel, and A. Barrau, “Uncalmented Kalman filter on lie groups for visual inertial odometry,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., 2018, pp. 649–655, doi: 10.1109/IROS.2018.8593627.

[5] M. Blöesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart, “Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback,” Int. J. Robot. Res., vol. 36, no. 10, pp. 1053–1072, 2017. [Online]. Available: https://doi.org/10.1177/0278364917728574.

[6] T. Qin, P. Li, and S. Shen, “VINS-Mono: A robust and versatile monocular visual-inertial state estimator,” IEEE Trans. Robot., vol. 34, no. 4, pp. 1004–1020, Aug. 2018, doi: 10.1109/TRO.2018.2853729.

[7] C. Förster, Z. Zhang, M. Gassner, M. Wörberger, and D. Scaramuzza, “SVO: Semidirect visual-inertial odometry for monocular and multicamera systems,” IEEE Trans. Robot., vol. 33, no. 2, pp. 249–265, Apr. 2017, doi: 10.1109/TRO.2016.2623335.

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

Fig. 11. Average time consumptions (ms) per step of different methods on EuRoC.
[9] T. Qin, S. Cao, J. Pan, and S. Shen, “A general optimization-based framework for global pose estimation with multiple sensors,” Accessed on: Mar. 1, 2020. [Online]. Available: https://arxiv.org/abs/1901.03642v1

[10] W. Lee, K. Eckenhoff, P. Geneva, and G. Huang, “Intermittent GPS-aided VIO: online initialization and calibration,” in Proc. IEEE Int. Conf. Robot. Autom., 2020, pp. 5724–5731. doi: 10.1109/ICRA4954.2020.9197029

[11] T. Schneider et al., “Maplab: An open framework for research in visual-inertial mapping and localization,” IEEE Robot. Autom. Lett., vol. 3, no. 3, pp. 1418–1425, Jul. 2018, doi: 10.1109/LRA.2018.2800113

[12] R. C. DuToit, J. A. Hesch, E. D. Nerurkar, and S. I. Roumeliotis, “Consistent-extended EKF with loop closures,” in Proc. IEEE Int. Conf. Robot. Autom., 2017, pp. 6253–6260, doi: 10.1109/ICRA.2017.7909741

[13] S. Huang and G. Dissanayake, “Convergence and consistency analysis for extended Kalman filter based SLAM,” IEEE Trans. Robot., vol. 23, no. 5, pp. 1036–1049, Oct. 2007, doi: 10.1109/TRO.2007.903811.

[14] P. Geneva, K. Eckenhoff, and G. Huang, “A linear-complexity EKF for visual-inertial navigation with loop closures,” in Proc. Int. Conf. Robot. Autom., Montreal, QC, Canada, 2019, pp. 3535–3541, doi: 10.1109/ICRA.2019.8793836.

[15] S. Lynen, M. W. Achtelik, S. Weiss, M. Chli, and R. Siegwart, “A robust and modular multi-sensor fusion approach applied to MAV navigation,” in Proc. IEEE/RSJ Int. Conf. Intel. Robots Syst., 2013, pp. 3923–3929, doi: 10.1109/IROS.2013.6690917.

[16] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, “Analysis and improvement of the consistency of extended Kalman filter based SLAM,” in Proc. IEEE Int. Conf. Robot. Autom., 2008, pp. 473–479, doi: 10.1109/ROBOT.2008.4543522.

[17] J. Kelly, S. Saripalli, and G. S. Sukhatme, “Combined visual and inertial navigation for an unmanned aerial vehicle,” in Field and Service Robotics. Berlin, Germany: Springer, 2008, pp. 255–264.

[18] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, “On-manifold preintegration for real-time visual-inertial odometry,” IEEE Trans. Robot., vol. 33, no. 1, pp. 1–21, Feb. 2017, doi: 10.1109/TRO.2016.2597321.

[19] X. Ding, F. Han, T. Yang, W. Wang, and R. Xiong, “Degeneration-aware localization with arbitrary global-local sensor fusion,” Sensors, vol. 21, no. 12, 2021, Art. no. 4042.

[20] A. Barrau and S. Bonnabel, “An EKF-SLAM algorithm with consistency properties,” 2015. [Online]. Available: http://arxiv.org/abs/1510.06263.

[21] S. Bonnabel, “Symmetries in observer design: Review of some recent results and applications to EKF-based SLAM,” in Proc. Conf. Robot Motion Control, 2011, pp. 3–15.

[22] A. Barrau and S. Bonnabel, “The invariant extended Kalman filter as a stable observer,” IEEE Trans. Autom. Control, vol. 62, no. 4, pp. 1797–1812, Apr. 2017.

[23] M. Brossard, A. Barrau, and S. Bonnabel, “Exploiting symmetries to design EKFs with consistency properties for navigation and SLAM,” IEEE Trans. Robot., vol. 31, no. 4, pp. 1572–1579, Aug. 2015.

[24] T. Bailey, J. Nieto, J. Guivant, M. Stevens, and E. Nebot, “Consistency of observations represented by implicit functions,” in Proc. IEEE/RSJ Int. Conf. Intel. Robots Syst., 2020, pp. 3662–3669, doi: 10.1109/IROS.2018.8593603.

[25] T. D. Barfoot, State Estimation for Robotics. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[26] S. Heo and C. G. Park, “Consistent EKF-based visual-inertial odometry with multi-sensor fusion,” in Proc. IEEE/RSJ Int. Conf. Intel. Robots Syst., 2018, pp. 7244–7251, doi: 10.1109/IROS.2018.8593941.

[27] J. Zhao, L. Zhao, S. Huang, and Y. Wang, “2D laser SLAM with general model for global pose estimation with multiple sensors. [Online]. Available: https://arxiv.org/abs/1906.06195

[28] Y. Jiao et al., “2-entity RANSAC for robust visual localization in changing environment,” in Proc. IEEE/RSJ Int. Conf. Intel. Robots Syst., 2019, pp. 2478–2485, doi: 10.1109/IROS40897.2019.8967671.

[29] B. Siciliano et al., “EuRoC - The challenge initiative for European robotics,” in Proc. 41st Int. Symp. Robot., 2014, pp. 1–7.

[30] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, “Complex urban dataset with multi-level sensors from highly diverse urban environments,” Int. J. Robot. Res., vol. 38, no. 6, pp. 642–657, 2019.

[31] P. Wenzel et al., “4Seasons: A cross-season dataset for multi-weather SLAM in autonomous driving,” in Proc. 42nd DAGM German Conf. Pattern Recognit., 2020, pp. 404–417.

[32] X. Ding et al., “Persistent stereo visual localization on cross-modal invariant map,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 11, pp. 4646–4658, Nov. 2020, doi: 10.1109/TITS.2019.2942760.

[33] Y. Jiao, Y. Wang, X. Ding, B. Fu, S. Huang, and R. Xiong, “2-entity random sample consensus for robust visual localization: Framework, methods, and verifications,” IEEE Trans. Int. Comput., vol. 68, no. 5, pp. 4519–4528, May 2021, doi: 10.1109/TIComp.2020.2984970.

[34] J. Zhao, L. Zhao, S. Huang, and Y. Wang, “2D laser SLAM with general model for global pose estimation,” in Proc. 42nd DAGM German Conf. Pattern Recognit., 2020, pp. 404–417.

[35] Z. Zhang, “Supplementary material of toward consistent and efficient state estimation,” 2023. [Online]. Available: https://drive.google.com/file/d/1TID9CVy3xAso9vs05qDj3s5gU1TJQBwh/view?usp=share_link

Zhuqing Zhang received the B.S. degree from the School of Astronautic, Northwestern Polytechnical University, Xi’an, China, in 2017, and the M.S. degree in aeronautical and astronomical science and technology from the School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai, China, in 2020. He is currently working toward the Ph.D. degree in the information fusion department with the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China.

His current research interests include multisensor fusion localization and SLAM.
Yang Song (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in mathematics from the School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China, in 2017 and 2020, respectively. He is currently working toward the Ph.D. degree in engineering with Robotics Institute, University of Technology Sydney, Ultimo, NSW, Australia.

His current research interests include EKF, SLAM, and optimization.

Shoudong Huang (Senior Member, IEEE) received the bachelor’s and master’s degrees in mathematics, and the Ph.D. degree in automatic control from North-eastern University, Shenyang, China, in 1987, 1990, and 1998, respectively.

He is currently a Professor with Robotics Institute, University of Technology Sydney, Ultimo, NSW, Australia. His research interests include nonlinear control systems and mobile robots simultaneous localization and mapping, exploration, and navigation.

Rong Xiong (Member, IEEE) received the Ph.D. degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2009.

She is currently a Professor with the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China. Her latest research interests include motion planning and SLAM.

Yue Wang (Member, IEEE) received the Ph.D. degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2016.

He is currently an Associate Professor with the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China. His latest research interests include mobile robotics and robot perception.