Leveraging Multiple Isolated Maps for Visual Inertial Odometry: Framework and Consistency

Zhuqing Zhang, Yanmei Jiao, Shenhan Jia, Rong Xiong, Yue Wang*

***Abstract*— Visual Inertial Odometry (VIO) is widely used in autonomous robots to provide the ego-pose of the robot. With the help of pre-built map information, the drift of the VIO can be constrained. However, constructing a globally consistent map is a tough job, especially for large scenes. In this paper, we propose a framework aiming to leverage multiple isolated maps to improve the performance of VIO. In this framework, the relative transformations between the local VIO reference frame and the multiple map reference frames are regarded as pose features to be online estimated. We call these relative transformations as *augmented variables*. With these *augmented variables*, the map-based information can be tightly coupled into the VIO system to ease the drift of VIO. To fuse these maps consistently, we theoretically analyze the observability properties of our proposed framework. Based on the analysis, the Schmidt extended Kalman filter (EKF) and the first-estimate Jacobian (FEJ) are employed to maintain the consistency of the system. Simulation and real-world experiments are conducted to demonstrate the effectiveness and consistency of our framework.**

I. Introduction

During the recent decade, visual-inertial odometry (VIO) has gained great attention, and a number of excellent VIO systems [1]–[5] have emerged with the efforts of researchers. However, VIO will inevitably suffer drift due to its global unobservable property [6]. To ease the drift, it is required to introduce the measurements beyond self-perception. Pre-built visual maps are commonly employed to relieve the drift of VIO. Especially, it would be favorable if there is more than one visual map available to constrain the drift of VIO and therefore improve positioning accuracy.

There are lots of algorithms improving the performance of VIO by fusing the pre-built map information [7]–[11]. However, the majority of these algorithms only support one global consistent map. Therefore, to assist VIO with multiple maps[1], some tedious and complicated pre-works are required to concatenate and fuse these maps into one global consistent map. Usually, a single visual map can be generated by simultaneous localization and mapping (SLAM) [12] or structure from motion (SFM) [13]. Nevertheless, for a large scenario, either building a big consistent map directly or merging several maps into one big map is quite time-consuming and complicated. Especially, to consistently fuse multiple maps, it is necessary to consider the covariance matrices of the submaps, which makes the problem more complicated.

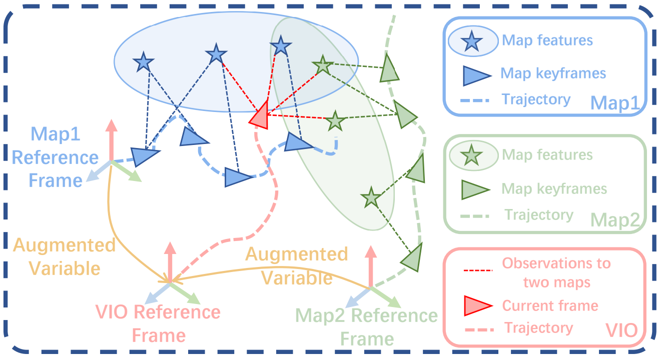


Fig. 1. An intuitive illustration of our proposed Multiple Isolated Maps Based Visual Inertial Odometry (MIMB-VIO). In MIMB, both VIO and the augmented variables are online estimated. When the current frame of the VIO observes the features of the multiple isolated maps, these map-based observations are used to update the state of MIMB-VIO.

An intuitive idea to utilize multiple maps is to regard them as multiple submaps. But how to acquire the relative transformation with covariance matrices between these multiple submaps is a problem. Huang et al. proposed a method to join submaps by performing data association between two adjacent submaps [14]. This method requires the submaps have co-vision areas. Besides, it is an offline map-building algorithm instead of utilizing multiple submaps to assist VIO localization. DuToit et al. proposed a method to compute the relative transformation between two sub-maps by partitioning the original map. After the partition, a relaxation strategy is employed so that these two submaps are independent [9]. However, this method still relies on a pre-built global consistent map, and there is a loss of accuracy in the map because of the relaxation strategy.

Since building a global consistent map from multiple isolated maps is time-consuming and complicated, and partitioning the global map into multiple submaps introduces the approximation, we argue that during the procedure of VIO, it is better to directly utilize multiple isolated (sub)maps and online estimate the relative transformations among them. In order to be lightweight and computationally efficient, a filter-based framework is employed to solve the problem of the multiple isolated maps based VIO (MIMB-VIO). For a filter-based algorithm, how to maintain the consistency of the system is the key. To be specific, on the one hand, because the pre-built maps are not *perfect* (100% accurate), the covariance of these maps needs to be correctly considered. On the other hand, the observability properties of the MIMB-VIO need to be analyzed and correctly maintained. In this

The authors are with the State Key Laboratory of Industrial Control Technology and Institute of Cyber-Systems and Control, Zhejiang University, Zhejiang, China.

* Yue Wang is the corresponding author wangyue@iipc.zju.edu.cn.

[1] In the rest of the paper, when we mention "map(s)", it means the pre-built map(s) instead of the map(s) built online.

paper, based on the above two points, we investigate how to fuse multiple isolated maps into VIO consistently.

As indicated in Fig. 1, in our proposed framework, MIMB-VIO estimates the robot pose in the VIO reference frame as well as the relative transformations between the VIO reference frame and the multiple isolated maps reference frames (the orange curved arrows)[2] simultaneously. In this way, the multiple maps, i.e., Map1 and Map2 in Fig. 1, can be fused into the VIO system without the need to build a global consistent map. We call the MIMB-VIO system that maintains the *augmented variable(s)* as the *augmented system*. By investigating the observability properties of the *augmented system*, we find that the *augmented variables* can be regarded as 6 degree-of-freedom (DoF) features maintained by the state. To maintain the consistency of the *augmented system*, we employ the Schmidt EKF [15] to consider the uncertainties of the maps while keeping the computational complexity at a low level. Besides, the first-estimate Jacobian (FEJ) technique is introduced to preserve the correct observability properties of the system. According to the simulations and experiments, we validated that consistently fusing multiple maps can significantly improve the performance of the VIO whereas the inconsistent *augmented system* will produce even worse results than the VIO. In summary, the contributions of this paper are four-folded:

- Propose a filter-based framework that fuses multiple isolated maps into the VIO system to improve localization accuracy.
- Analyze the observability properties of our proposed MIMB-VIO system.
- Introduce Schmidt EKF and FEJ to make the system consistent.
- Validate the effectiveness and consistency of the proposed method. The source code of our proposed algorithm is released.[3]

II. Related Works

This section gives a brief review of the related works from two perspectives: the map-based localization system and the observability of the VIO system.

A. Map-based Localization Systems

Map-based localization systems can be classified into optimization-based [7], [10], [16] and filter-based [8], [9], [17]. In these works, a trade-off has been made between the computation and the consideration of map uncertainty. To be specific, most of them [7], [8], [10], [16] highly rely on the accuracy of the pre-built map and neglect the uncertainty of the map. [17] considers the map uncertainty by utilizing the covariance intersection (CI) method. The work [9] considers the map uncertainty through the Schmidt filter. However, all these methods perform localization on a global consistent map. The framework of [9] inspires us to design the framework for the MIMB-VIO system. In [9], the relative transformation between the map reference frame and the VIO reference frame (i.e. our defined *augmented variable*) is maintained. To reduce the computational cost, the authors propose a relaxation method to partition the global map into several sub-maps. Nevertheless, as we will demonstrate in Sec. IV, the formulation like [9] will suffer from inconsistency. This problem will be solved in Sec. V. Besides, [9] still needs a global consistent map instead of directly employing several isolated maps.

TABLE I
The definition of involved reference frames/sensors

L	The local inertial reference frame, which is a fixed frame defined by the initial pose of VIO (c.f. L in Fig. 2).
G^i	The ith map reference frame. All the ith-map-related information (including map keyframes and map features) is based on this frame (c.f. G^1 and G^2 in Fig. 2).
I_t	The IMU (body) frame at timestamp t, which is attached to the robot. VIO online estimates the transformation between L and I_t, i.e., ${}^L\mathbf{T}_{I_t}$[4].
C_t	The camera (image) frame at timestamp t. The camera online observes features in C_t.
${}^iKF_j$	The j^{th} image keyframe of the ith pre-built map, whose pose ${}^{G^i}\mathbf{T}_{KF_j}$ is represented in G^i (cf. ${}^2KF_1$ and ${}^1KF_1$ in Fig. 2). After the ith map is built, the keyframes $\{{}^iKF_j\}$ are fixed.

B. Observability Analysis of the Visual Inertial Odometry

The filter-based VIO has the demerit of acquiring spurious information along the unobservable direction due to the ever-changing linearization points [6]. There are several works trying to relieve the inconsistency of the VIO by fixing the linearization point [18], searching for a linearization point in the observable subspace [6], or reformulating the state error on the manifold [19]. Unfortunately, since our proposed *augmented system* introduces additional variables like the *augmented variables* and map information, the observability properties of the pure VIO cannot reflect that of the *augmented system*. The main contribution of this paper is providing a theoretical analysis of the observability properties of the *augmented system*.

III. System Framework

In this section, we will introduce the involved frame of the *augmented system*, the state of the system, the propagation of the state, and the observation functions used for updating the state. The backbone of our framework is a VIO system Open-VINS [2], which is a state-of-the-art implementation of MSCKF [1].

A. Involved Frames

As is shown in Fig. 2 and Table I, for the *augmented system* there are five kinds of frames, i.e., L, I_t, C_t, G^i, and ${}^iKF_j$, where L, I_t, and C_t are VIO-related frames, and G^i and ${}^iKF_j$ are map-related frames.

[2] For simplicity, we call this kind of relative transformation(s) the *augmented variable(s)*.

[3] https://github.com/zhuqingzhang/C-MIMB-VIO

[4] In this paper, we use $\mathbf{T}$ to represent a pose or a transformation. ${}^L\mathbf{T}_{I_t}$ means the pose of the frame I_t in the frame L, or transforming a state from the frame I_t to the frame L.

B. System State

We extend the state of Open-VINS [2] to get the *augmented system* state $\mathbf{x}_t$:

$$\begin{aligned}\mathbf{x}_t &= \begin{bmatrix}\mathbf{x}_{VIO_t}^\top & \mathbf{x}_{SW_t}^\top & \mathbf{x}_{\tau_t}^\top & \mathbf{x}_{KF_t}^\top\end{bmatrix}^\top \\ \mathbf{x}_{VIO_t} &= \begin{bmatrix}{}^{I_t}\mathbf{q}_L^\top & {}^{L}\mathbf{v}_{I_t}^\top & {}^{L}\mathbf{p}_{I_t}^\top & \mathbf{b}_{g_t}^\top & \mathbf{b}_{a_t}^\top\end{bmatrix}^\top \\ \mathbf{x}_{SW_t} &= \begin{bmatrix}{}^{I_t}\mathbf{q}_L^\top & {}^{L}\mathbf{p}_{I_t}^\top & \cdots & {}^{I_{t-B+1}}\mathbf{q}_L^\top & {}^{L}\mathbf{p}_{I_{t-B+1}}^\top\end{bmatrix}^\top \\ \mathbf{x}_{\tau_t} &= \begin{bmatrix}{}^{G_t^1}\mathbf{q}_L^\top & {}^{G_t^1}\mathbf{p}_L^\top & \cdots & {}^{G_t^N}\mathbf{q}_L^\top & {}^{G_t^N}\mathbf{p}_L^\top\end{bmatrix}^\top \\ \mathbf{x}_{KF_t} &= \begin{bmatrix}\cdots {}^{G_t^1}\mathbf{q}_{kf_j}^\top & {}^{G_t^1}\mathbf{p}_{kf_j}^\top \cdots {}^{G_t^N}\mathbf{q}_{kf_k}^\top & {}^{G_t^N}\mathbf{p}_{kf_k}^\top \cdots\end{bmatrix}^\top\end{aligned} \tag{1}$$

where $\mathbf{x}_{VIO_t}$ and $\mathbf{x}_{SW_t}$ form the original state of Open-VINS [2], $\mathbf{x}_{\tau_t}$ and $\mathbf{x}_{KF_t}$ are the augmented parts. To be specific, for $\mathbf{x}_{VIO_t}$, ${}^{I_t}\mathbf{q}_L$ is the unit quaternion that transforms a 3D vector from L to I_t, ${}^{L}\mathbf{p}_{I_t}$ is the position of I_t in L, ${}^{L}\mathbf{v}_{I_t}$ is the velocity of I_t in the frame L, and $\mathbf{b}_{g_t}$ and $\mathbf{b}_{a_t}$ are the bias of gyroscope and accelerometer. VIO outputs the pose ${}^{L}\mathbf{T}_{I_t} \triangleq \{{}^{I_t}\mathbf{q}_L, {}^{L}\mathbf{p}_{I_t}\}$. $\mathbf{x}_{SW_t}$ contains B cloned historical body poses, which is the so-called sliding window. Suppose we have N isolated pre-built maps, then the elements ${}^{G_t^i}\mathbf{q}_L$, ${}^{G_t^i}\mathbf{p}_L$, $i = 1 \cdots N$ represent the relative transformation between the local VIO reference frame L and the ith map reference frame G^i, i.e., the so-called *augmented variable*. $\mathbf{x}_{KF_t}$ contains a set of keyframe poses from N isolated maps. For instance, ${}^{G_t^1}\mathbf{q}_{kf_j}$ and ${}^{G_t^1}\mathbf{p}_{kf_j}$ represent the rotation and the position of the jth keyframe from the 1st map, respectively.

C. State Propagation

For the *augmented system* with the state defined in (1), we have the following kinematics equations:

$$\begin{cases} {}^{I_t}\dot{\mathbf{q}}_L = \dfrac{1}{2}\Omega(\boldsymbol{\omega}_{m_t} - \mathbf{b}_{g_t} - \mathbf{n}_{g_t})\,{}^{I_t}\mathbf{q}_L \\ {}^{L}\dot{\mathbf{v}}_{I_t} = C({}^{I_t}\mathbf{q}_L)^\top(\mathbf{a}_{m_t} - \mathbf{b}_{a_t} - \mathbf{n}_{a_t}) + \mathbf{g} \\ {}^{L}\dot{\mathbf{p}}_{I_t} = {}^{L}\mathbf{v}_{I_t} \\ \dot{\mathbf{b}}_{g_t} = \mathbf{n}_{wg_t} \quad \dot{\mathbf{b}}_{a_t} = \mathbf{n}_{wa_t} \\ {}^{G_t^i}\dot{\mathbf{q}}_L = \mathbf{0} \quad {}^{G_t^i}\dot{\mathbf{p}}_L = \mathbf{0} \\ {}^{G_t^i}\dot{\mathbf{q}}_{kf_j} = \mathbf{0} \quad {}^{G_t^i}\dot{\mathbf{p}}_{kf_j} = \mathbf{0} \end{cases} \tag{2}$$

where we argue that the *augmented variable* and the keyframe pose from maps should be static. $C(\cdot)$ transforms a quaternion to a rotation matrix $\mathbf{R}$, $\boldsymbol{\omega}_{m_t}$ and $\mathbf{a}_{m_t}$ are the measurements from IMU in frame L, $\mathbf{n}_{g_t}$ and $\mathbf{n}_{a_t}$ are the measurement noises, $\mathbf{n}_{wg_t}$ and $\mathbf{n}_{wa_t}$ are the random walk noises, $\mathbf{g} = \begin{bmatrix}0 & 0 & -9.81\end{bmatrix}^\top m/s^2$ is the gravitational acceleration in L. For a vector $\boldsymbol{\omega} \in \mathbb{R}^3$, we have

$$\Omega(\boldsymbol{\omega}) = \begin{bmatrix} -\boldsymbol{\omega}_\times & \boldsymbol{\omega} \\ \boldsymbol{\omega}^\top & 0 \end{bmatrix}.$$

By discretizing (2), we can propagate the state from time step t to time step $t+1$:

$$\mathbf{x}_{t+1|t} = \mathbf{f}(\mathbf{x}_t, \mathbf{a}_{m_t} - \mathbf{n}_{a_t}, \omega_{m_t} - \mathbf{n}_{\omega_t}) \tag{3}$$

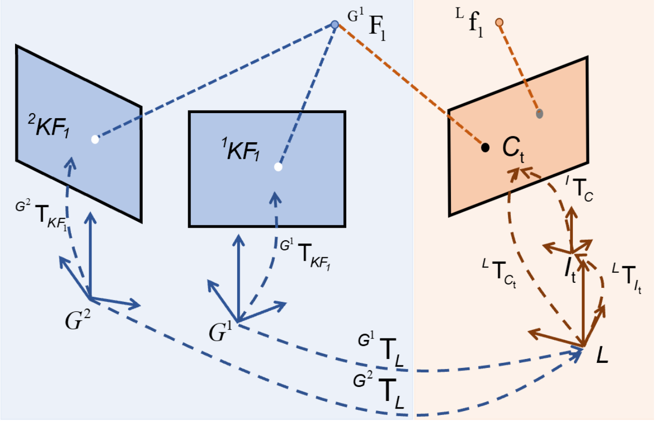


Fig. 2. The observation functions

where $\mathbf{f}(\cdot)$ is the discretized propagation function derived from (2), the subscript $t+1|t$ means the state is propagated from t to $t+1$, with all the measurements up to time step t being processed. Then, the state covariance can be propagated by

$$\mathbf{P}_{t+1|t} = \boldsymbol{\Phi}_{t+1|t}\mathbf{P}_t\boldsymbol{\Phi}_{t+1|t}^\top + \mathbf{G}_{t+1|t}\mathbf{Q}_t\mathbf{G}_{t+1|t}^\top \tag{4}$$

where $\mathbf{P}_t$ and $\mathbf{Q}_t$ are the system state covariance and the noise covariance, respectively. $\boldsymbol{\Phi}_{t+1|t}$ and $\mathbf{G}_{t+1|t}$ are the Jacobians of (3) with respect to the system state and the noise, respectively. $\boldsymbol{\Phi}_{t+1|t}$ is also called state transition matrix. For the detailed derivation, please refer to [1], [2].

D. Observation Functions

As indicated in Fig. 2, there are two kinds of observation functions, the local feature based (the orange shade part) and the map feature based (the blue shade part).

Local Observation: For a local feature ${}^{L}\mathbf{f}_i$, for example, ${}^{L}\mathbf{f}_1$ in Fig. 2, online detected by the camera, we have the following local observation function:

$${}^{loc}\mathbf{y}_t^i = \mathrm{h}({}^{I_t}\mathbf{R}_L({}^{L}\mathbf{f}_{i_t} - {}^{L}\mathbf{p}_{I_t})) + {}^{loc}\boldsymbol{\gamma}_t^i \tag{5}$$

where for simplicity, we assume the extrinsic ${}^{I}\mathbf{T}_L$ between the IMU and the camera is an identity matrix. ${}^{I_t}\mathbf{R}_L$ is the rotation matrix from L to I. $\mathrm{h}(\cdot)$ is the camera projection function that projects a 3D feature in the camera frame into the 2D image plane. ${}^{loc}\boldsymbol{\gamma}_t^i$ is the noise of the local observation.

With (5), the error of the local observation can be computed by:

$$\begin{aligned}{}^{loc}\mathbf{r}_t^i \triangleq\ & \mathrm{h}({}^{I_t}\mathbf{R}_L({}^{L}\mathbf{f}_{i_t} - {}^{L}\mathbf{p}_{I_t})) + {}^{loc}\boldsymbol{\gamma}_t^i - \\ & \mathrm{h}({}^{I_t}\hat{\mathbf{R}}_L({}^{L}\hat{\mathbf{f}}_{i_t} - {}^{L}\hat{\mathbf{p}}_{I_t})) \\ {}^{loc}\mathbf{r}_t^i =\ & {}^{loc}\mathbf{H}_{\mathbf{x}_t}^i\tilde{\mathbf{x}}_t + {}^{loc}\mathbf{H}_{{}^{L}\mathbf{f}_{i_t}}{}^{L}\tilde{\mathbf{f}}_{i_t} + {}^{loc}\boldsymbol{\gamma}_t^i\end{aligned} \tag{6}$$

where $\hat{\cdot}$ represents the estimated value of the variable, $\tilde{\cdot}$ represents the error between the true value and the estimated value of the variable. ${}^{loc}\mathbf{H}_{x_t}^i$ and ${}^{loc}\mathbf{H}_{{}^{L}\mathbf{f}_{i_t}}$ represent the Jacobian matrices of h with respect to $\mathbf{x}_t$ and ${}^{L}\mathbf{f}_{i_t}$, respectively.

Note that in the framework of MSCKF [1] we do not maintain the local features in the state vector, so that the Jacobian matrix of h with respect to the local feature ${}^{L}\mathbf{f}_{i_t}$

should be null-space projected to eliminate. The sliding window $\mathbf{x}_{SW}$ provides us an opportunity to achieve this purpose: a local feature can be observed not only by the current camera frame but also by the historical successive camera frames. This means for each local feature, there is more than one (say $n \geq 2$) local observation function. By stacking these local observation functions, the row number ($2n$) of the Jacobian matrix with respect to the local feature ${}^{L}\mathbf{f}_{i_t}$ is larger than its column number (3, the dimension of the local feature), which guarantees the existence of the left null-space of the Jacobian matrix of h with respect to ${}^{L}\mathbf{f}_{i_t}$. For the details, please refer to [1].

Map-based Observation: For a feature from a pre-built map, it can be observed by the current camera C_t and keyframes ${}^{i}KF_j$ from different maps, as indicated by the blue shade part of Fig. 2. Suppose there is a feature ${}^{G^k}\mathbf{F}_i$ from the kth map, observed by the jth keyframe ${}^{k}KF_j$ in the kth map. Besides, ${}^{G^k}\mathbf{F}_i$ can also be observed by the current camera frame C and the lth keyframe ${}^{s}KF_l$ in the sth ($s \neq k$) map. For example, by setting $k=1, i=1, j=1, s=2, l=1$, the situation is exactly described by Fig. 2. Then, there exist three observation functions:

$$ {}^{cam}\mathbf{y}_t^{ki} = \mathrm{h}\Big[{}^{I_t}\mathbf{R}_L({}^{G_t^k}\mathbf{R}_L^\top({}^{G^k}\mathbf{F}_{i_t} - {}^{G_t^k}\mathbf{p}_L) - {}^{L}\mathbf{p}_{I_t})\Big] + {}^{cam}\boldsymbol{\gamma}_t^{ki} \tag{7} $$

$$ {}^{kKF_j}\mathbf{y}_t^{ki} = \mathrm{h}\left[{}^{G_t^k}\mathbf{R}_{kf_j}^\top({}^{G^k}\mathbf{F}_{i_t} - {}^{G_t^k}\mathbf{p}_{kf_j})\right] + {}^{kKF_j}\boldsymbol{\gamma}_t^{ki} \tag{8} $$

$$ \begin{aligned} {}^{sKF_l}\mathbf{y}_t^{ki} = \mathrm{h}\Big[&{}^{G_t^s}\mathbf{R}_{kf_l}^\top \Big[{}^{G_t^s}\mathbf{R}_L {}^{G_t^k}\mathbf{R}_L^\top({}^{G^k}\mathbf{F}_{i_t} - {}^{G_t^k}\mathbf{p}_L) \\ &+ {}^{G_t^s}\mathbf{p}_L - {}^{G_t^s}\mathbf{p}_{kf_l}\Big]\Big] + {}^{sKF_l}\boldsymbol{\gamma}_t^{ki} \end{aligned} \tag{9} $$

where ${}^{cam}\mathbf{y}_t^{ki}$, ${}^{kKF_j}\mathbf{y}_t^{ki}$ and ${}^{sKF_l}\mathbf{y}_t^{ki}$ represent the observations to the feature ${}^{G^k}\mathbf{F}_i$ by the current camera, the map keyframe ${}^{k}KF_j$ and the map keyframe ${}^{s}KF_l$, respectively. ${}^{cam}\boldsymbol{\gamma}_t^{ki}$, ${}^{kKF_j}\boldsymbol{\gamma}_t^{ki}$ and ${}^{sKF_l}\boldsymbol{\gamma}_t^{ki}$ are the corresponding observation noises. For simplicity, we assume the extrinsic ${}^{I}\mathbf{T}_C$ between the IMU and the camera is an identity matrix.

Similar to (6), we have the following map-based observation errors derived from (7)-(9):

$$ {}^{cam}\mathbf{r}_t^{ki} = {}^{cam}\mathbf{H}_{\mathbf{x}_t}^{ki}\tilde{\mathbf{x}}_t + {}^{cam}\mathbf{H}_{{}^{G^k}\mathbf{F}_{i_t}} {}^{G^k}\tilde{\mathbf{F}}_{i_t} + {}^{cam}\boldsymbol{\gamma}_t^{ki} \tag{10} $$

$$ {}^{kKF_j}\mathbf{r}_t^{ki} = {}^{kKF_j}\mathbf{H}_{\mathbf{x}_t}^{ki}\tilde{\mathbf{x}}_t + {}^{kKF_j}\mathbf{H}_{{}^{G^k}\mathbf{F}_{i_t}} {}^{G^k}\tilde{\mathbf{F}}_{i_t} + {}^{kKF_j}\boldsymbol{\gamma}_t^{ki} \tag{11} $$

$$ {}^{sKF_l}\mathbf{r}_t^{ki} = {}^{sKF_l}\mathbf{H}_{\mathbf{x}_t}^{ki}\tilde{\mathbf{x}}_t + {}^{sKF_l}\mathbf{H}_{{}^{G^k}\mathbf{F}_{i_t}} {}^{G^k}\tilde{\mathbf{F}}_{i_t} + {}^{sKF_l}\boldsymbol{\gamma}_t^{ki} \tag{12} $$

Again, as we do not maintain the map feature ${}^{G^k}\mathbf{F}_{i_t}$ in the state vector, the Jacobian matrices related to ${}^{G^k}\mathbf{F}_{i_t}$ should be null-space projected. Fortunately, as shown in (7)-(9), for each feature ${}^{G^k}\mathbf{F}_{i_t}$, it corresponds to at least two observation functions (if the feature ${}^{G^k}\mathbf{F}_{i_t}$ is observed only by the keyframe in the kth map, (9) vanishes), which means we can find a left null-space matrix to eliminate the ${}^{G^k}\tilde{\mathbf{F}}_{i_t}$ related parts.

IV. Observability Analysis

Given the system mentioned in Sec. III, we need to investigate its observability properties, which can indicate the consistency of the system [6], [18].

The unobservable subspace of the system is the right null space of the observability matrix of the system. The observability matrix of the proposed system can be given by:

$$ \mathbf{M} \triangleq \begin{bmatrix} \mathbf{H}_{L_0} \\ \mathbf{H}_{G_0} \\ \mathbf{H}_{L_1}\boldsymbol{\Phi}_{1|0} \\ \mathbf{H}_{G_1}\boldsymbol{\Phi}_{1|0} \\ \vdots \\ \mathbf{H}_{L_t}\boldsymbol{\Phi}_{t|0} \\ \mathbf{H}_{G_t}\boldsymbol{\Phi}_{t|0} \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{M}_{L_0} \\ \mathbf{M}_{G_0} \\ \mathbf{M}_{L_1} \\ \mathbf{M}_{G_1} \\ \vdots \\ \mathbf{M}_{L_t} \\ \mathbf{M}_{G_t} \end{bmatrix} \tag{13} $$

where $\mathbf{H}_{L_t}$ represents the Jacobian matrix of the local observation function, corresponding to $\left[{}^{loc}\mathbf{H}_{\mathbf{x}_t}^{i} \quad {}^{loc}\mathbf{H}_{{}^{L}\mathbf{f}_{i_t}}\right]$. $\mathbf{H}_{G_t}$ represents the Jacobian matrix of the map-based observation function, which is the stacking matrix along the row direction of $\left[{}^{cam}\mathbf{H}_{\mathbf{x}_t}^{ki} \quad {}^{cam}\mathbf{H}_{{}^{G^k}\mathbf{F}_{i_t}}\right]$, $\left[{}^{kKF_j}\mathbf{H}_{\mathbf{x}_t}^{ki} \quad {}^{kKF_j}\mathbf{H}_{{}^{G^k}\mathbf{F}_{i_t}}\right]$ and $\left[{}^{sKF_l}\mathbf{H}_{\mathbf{x}_t}^{ki} \quad {}^{sKF_l}\mathbf{H}_{{}^{G^k}\mathbf{F}_{i_t}}\right]$. The transition matrix from time step 0 to time step t can be computed by $\boldsymbol{\Phi}_{t|0} \triangleq \boldsymbol{\Phi}_{t|t-1}\ldots\boldsymbol{\Phi}_{2|1}\boldsymbol{\Phi}_{1|0}$.[5]

As the Jacobian matrices of local and map features are considered here, we need to put these features in the state vector. Besides, for simplicity, the IMU bias and cloned state are omitted. Therefore, the new augmented state is given as:

$$ \begin{aligned} \mathbf{x}_{Aug_t} = \Big[&{}^{I_t}\mathbf{q}_L^\top \ {}^{L}\mathbf{v}_{I_t}^\top \ {}^{L}\mathbf{p}_{I_t}^\top \ {}^{L}\mathbf{f}_{i_t}^\top \mid {}^{G_t^k}\mathbf{q}_L^\top \ {}^{G_t^k}\mathbf{p}_L^\top \ {}^{G_t^s}\mathbf{q}_L^\top \\ &{}^{G_t^s}\mathbf{p}_L^\top \mid {}^{G^k}\mathbf{F}_{i_t}^\top \ {}^{G_t^k}\mathbf{q}_{kf_j}^\top \ {}^{G_t^k}\mathbf{p}_{kf_j}^\top \ {}^{G_t^s}\mathbf{q}_{kf_l}^\top \ {}^{G_t^s}\mathbf{p}_{kf_l}^\top \Big]^\top \end{aligned} \tag{14} $$

where the elements before the first $|$ are VIO-related, those after the first $|$ are the *augmented variables*, and those after the second $|$ are map-related. We assume all the features are static, i.e., the derivations of their positions are zero.

To investigate the observability properties of the *augmented system*, we should first consider the ideal case where all the values are given by ground truth. Then, the actual case is considered where all the values are given by estimation. If the observability of the actual case is different from that of the ideal case, the system will suffer from inconsistency.

Theorem 1: (Ideal observability properties of the *augmented system*) For the ideal case, the right null space $\mathcal{N}_i$ of the observability matrix, where the Jacobian matrices are evaluated with the true values, is spanned by ten (four plus

[5] Note that there is a slight abuse of symbol. The dimension of $\boldsymbol{\Phi}$ in this section is different from that in (4). In this section, $\boldsymbol{\Phi}$ is augmented by the local feature related and the map feature related parts.

six) directions:

$$\mathcal{N}_i = \begin{bmatrix} {}^{I_0}\mathbf{R}_L\mathbf{g} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -({}^{L}\mathbf{v}_{I_0})_\times\mathbf{g} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -({}^{L}\mathbf{p}_{I_0})_\times\mathbf{g} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ -({}^{L}\mathbf{f}_i)_\times\mathbf{g} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ {}^{G^k}\mathbf{R}_L\mathbf{g} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -{}^{G^k}\mathbf{R}_L & \mathbf{0} & \mathbf{I} \\ {}^{G^s}\mathbf{R}_L\mathbf{g} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -{}^{G^s}\mathbf{R}_L & \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} & ({}^{G^k}\mathbf{F}_i - {}^{G^k}\mathbf{p}_L)_\times & \mathbf{I} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{15}$$

Proof: See **Appendix** I. ■

Theorem 2: (Actual observability properties of the *augmented system*) For the actual case, the right null space $\mathcal{N}_a$ of the observability matrix, where the Jacobian matrices are evaluated with the estimated values, is spanned by three directions:

$$\mathcal{N}_a = \begin{bmatrix} \mathbf{0}_{3\times15} & \mathbf{I} & \mathbf{0} & \mathbf{I} & \mathbf{I} & \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{I} \end{bmatrix}^\top \tag{16}$$

Proof: See **Appendix** I. ■

From the above two theorems, we can find that compared with the ideal case, seven unobservable directions vanish for the actual case. This means if we use the estimated values to compute Jacobian matrices, the MIMB-VIO system will suffer from inconsistency. Readers can find more detailed analysis in **Appendix** I.

To reveal how the introduced *augmented variables* affect the observability properties of the *augmented system*, we further investigate a special case: when the maps are *perfect*, i.e., the *augmented system* does not need to consider the uncertainty of the maps, then the map-related elements in (14) are removed. In this case, we can derive the following two Theorems:

Theorem 3: (Ideal observability properties of the *augmented system* with *perfect* maps) For the ideal case, when system-used maps are *perfect*, the right null space $\mathcal{N}_{ip}$ of the observability matrix, where the Jacobian matrices are evaluated with the true values, is spanned by four directions:

$$\mathcal{N}_{ip} = \begin{bmatrix} {}^{I_0}\mathbf{R}_L\mathbf{g} & \mathbf{0} \\ -({}^{L}\mathbf{v}_{I_0})_\times\mathbf{g} & \mathbf{0} \\ -({}^{L}\mathbf{p}_{I_0})_\times\mathbf{g} & \mathbf{I} \\ -({}^{L}\mathbf{f}_i)_\times\mathbf{g} & \mathbf{I} \\ {}^{G^k}\mathbf{R}_L\mathbf{g} & \mathbf{0} \\ \mathbf{0} & -{}^{G^k}\mathbf{R}_L \\ {}^{G^s}\mathbf{R}_L\mathbf{g} & \mathbf{0} \\ \mathbf{0} & -{}^{G^s}\mathbf{R}_L \end{bmatrix} \tag{17}$$

Proof: See **Appendix** I. ■

Theorem 4: (Actual observability properties of the *augmented system* with *perfect* maps) For the actual case, when system used maps are *perfect*, the original four unobservable directions vanish:

Proof: See **Appendix** I. ■

We can find when the maps are *perfect*, for the ideal case, the dimension of the unobservable subspace of the *augmented system* is the same as the dimension of the unobservable subspace of the VI-SLAM system — four dimensions. Actually, the *augmented variables* can be regarded as 6 DoF pose features maintained in the state. Every time a map-based observation occurs, it is equivalent to the *augmented system* observing these pose features, similar to that the VI-SLAM system closes a loop. Therefore, consistently estimating these pose features (i.e., *augmented variables*) can efficiently improve the performance of the VI-SLAM system (i.e., *augmented system*). However, in the actual case, the original four unobservable directions vanish because of the introduced *augmented variables*.

For the case that the maps are *imperfect*, we need to consider their uncertainties, then an additional six-dimensional unobservable subspace is introduced to the *augmented system*, as indicated by (15).

For the detailed derivation and analysis, please refers to **Appendix** I.

V. Consistent Estimation

Inconsistency stems from 1) mistakenly considering the uncertainty of the fused information; 2) mistakenly maintaining the observability properties of the system. Aiming to 1), we add the map information (keyframe poses) into the state vector like (1) to consider the uncertainty of the map keyframe poses. However, if we update all the state values like standard EKF, the huge computation would be unaffordable. Aiming to 2), as mentioned in Sec. IV, the ever-changing estimated value of the (3 DoF and 6 DoF) features will break the original observability properties of the system, resulting in inconsistency. Therefore, the Schmidt updating [15] and the first estimate Jacobian (FEJ) [18] techniques are introduced in this section to make the system efficient and consistent.

A. Schmidt Updating

Schmidt updating is only used for map-based observations. When there are measurements from maps, we have (10)-(12). By stacking them along the row and performing left null-space projection to eliminate the map feature related part, we have the following concise expression:

$$\begin{aligned} \mathbf{r}_t^* &= \mathbf{H}_{\mathbf{x}_t}^* \tilde{\mathbf{x}}_t + \boldsymbol{\gamma}_t^* \\ &\triangleq \begin{bmatrix} \mathbf{H}_{A_t}^* & \mathbf{H}_{N_t}^* \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_{A_t} \\ \tilde{\mathbf{x}}_{N_t} \end{bmatrix} + \boldsymbol{\gamma}_t^* \end{aligned} \tag{18}$$

where we partition the state $\mathbf{x}_t$ into the two parts, $\mathbf{x}_{A_t}$ and $\mathbf{x}_{N_t}$. $\mathbf{x}_{A_t}$ is composed of $\mathbf{x}_{VIO_t}$, $\mathbf{x}_{SW_t}$ and $\mathbf{x}_{\mathcal{T}_t}$. $\mathbf{x}_{N_t}$ is map keyframe related part, i.e., $\mathbf{x}_{KF_t}$. $\mathbf{H}_{A_t}^*$ and $\mathbf{H}_{N_t}^*$ are the Jacobian matrices corresponding to $\mathbf{x}_{A_t}$ and $\mathbf{x}_{N_t}$, respectively.

Then, the Schmidt updating can be performed with the following equations:

$$\begin{aligned} \hat{\mathbf{x}}_{A_t} &= \hat{\mathbf{x}}_{A_{t|t-1}} + \mathbf{K}_{A_t}\mathbf{r}_t^* \\ \hat{\mathbf{x}}_{N_t} &= \hat{\mathbf{x}}_{N_{t|t-1}} \end{aligned} \tag{19}$$

$$\begin{bmatrix}\mathbf{K}_{A_t}\\ \mathbf{K}_{N_t}\end{bmatrix} \triangleq \begin{bmatrix}\bar{\mathbf{K}}_{A_t}\\ \bar{\mathbf{K}}_{N_t}\end{bmatrix}\mathbf{S}_t^{-1} = \begin{bmatrix}\mathbf{P}_{AA_{t|t-1}}\mathbf{H}_{A_t}^{*\top} + \mathbf{P}_{AN_{t|t-1}}\mathbf{H}_{N_t}^{*\top}\\ \mathbf{P}_{NA_{t|t-1}}\mathbf{H}_{A_t}^{*\top} + \mathbf{P}_{NN_{t|t-1}}\mathbf{H}_{N_t}^{*\top}\end{bmatrix}\mathbf{S}_t^{-1} \tag{20}$$

$$\mathbf{S}_t = \mathbf{H}_{\mathbf{x}_t}^{*}\mathbf{P}_{t|t-1}\mathbf{H}_{\mathbf{x}_t}^{*\top} + \mathbf{R}_t$$

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \begin{bmatrix}\mathbf{K}_{A_t}\mathbf{S}_t\mathbf{K}_{A_t}^{\top} & \mathbf{K}_{A_t}\mathbf{H}_t^{*}\begin{bmatrix}\mathbf{P}_{AN_{t|t-1}}\\ \mathbf{P}_{NN_{t|t-1}}\end{bmatrix}\\ \begin{bmatrix}\mathbf{P}_{AN_{t|t-1}}\\ \mathbf{P}_{NN_{t|t-1}}\end{bmatrix}^{\top}\mathbf{H}_t^{*\top}\mathbf{K}_{A_t}^{\top} & \mathbf{0}_{\mathfrak{m}\times\mathfrak{m}}\end{bmatrix} \tag{21}$$

where $\mathfrak{m}$ is the dimension of the map keyframe poses.

We can find that, in Schmidt updating, the map keyframe related part in the state is not updated, and neither is the corresponding covariance. This mechanism guarantees low computation. For the standard EKF updating, $\mathbf{0}_{\mathfrak{m}\times\mathfrak{m}}$ in (21) is replaced by $\mathbf{K}_{N_t}\mathbf{S}_t\mathbf{K}_{N_t}^{\top}$, which makes the computational complexity increase from $\mathcal{O}(\mathfrak{m})$ to $\mathcal{O}(\mathfrak{m}^2)$. Besides, the relation between the updated covariance matrices from the Schmidt updating and from the standard EKF updating is that $\mathbf{P}_{Schmidt} = \mathbf{P}_{EKF} + \begin{bmatrix}\mathbf{0}\\ \bar{\mathbf{K}}_N\end{bmatrix}\mathbf{S}^{-1}\begin{bmatrix}\mathbf{0} & \bar{\mathbf{K}}_N^{\top}\end{bmatrix}$, which means compared with the covariance from the standard EKF, the covariance from the Schmidt updating will be relative conservative. If the standard EKF is not overconfident about the estimated state, then neither is the Schmidt updating. Hence, by employing the Schmidt updating, we can keep the low computation while not over-optimistically estimating the state.

B. *Observability Maintainance*

From the analysis of Sec. IV, we conclude that the ever-changing estimated values make the linearization points of nonlinear kinematics and observation functions also change with time, even for those state elements that should be constant. This makes the system think it can obtain information in the original unobservable subspace, leading to inconsistency. To solve this problem, we introduce the FEJ technique to constrain the system only obtaining information in the observable subspace.

To be specific, we carefully extend the traditional first-estimate Jacobian (FEJ) [18] to our *augmented system* with two steps:

- The propagation Jacobian matrix at time step t is evaluated at $\hat{\mathbf{x}}_{t|t-1}$.
- The map-based observation Jacobian matrix at time step t is ever evaluated at the first estimate of ${}^{G^i}\mathbf{p}_L$, ${}^{G^i}\mathbf{R}_L$ and $\hat{\mathbf{x}}_{t|t-1}$.

With these two steps, for the actual case, the transition matrix $\mathbf{\Phi}_{t|0}$ can be given by (29). Besides, the linearization points of ${}^{G^i}\mathbf{R}_L$ and ${}^{G^i}\mathbf{p}_L$ will never change. Moreover, as we do not maintain the map features in the state, their values will also be constant. At this point, the null space of the observability matrix $\mathbf{M}$ in the actual case keeps the same as that in the ideal case.

VI. Experiments

In this section, we will validate the consistency of our proposed algorithm through simulation experiments. Then, real-world experiments are conducted to further illustrate the effectiveness of our proposed algorithm.

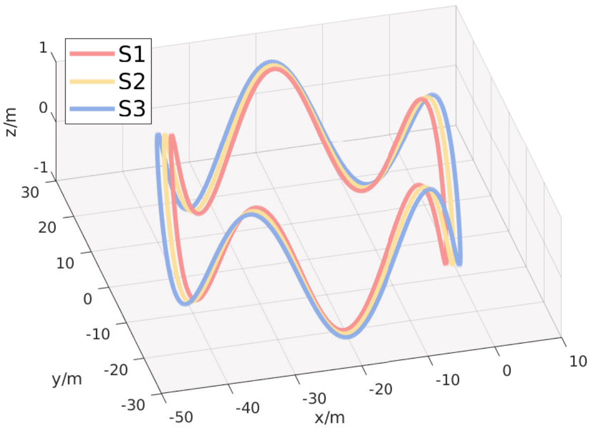


Fig. 3. The trajectories used for simulation.

A. *Simulation Experiments*

In this part, we make some adaptations to the simulator of Open-VINS [2] to evaluate the accuracy and consistency of our proposed algorithm. There are three trajectories used in the simulation, which are plotted in Fig. 3. These three trajectories are denoted as S1 (the red one), S2 (the orange one), and S3 (the blue one), respectively. When we online run one of the three trajectories (for example, S1), the other two trajectories (S2, S3) can be used as the two isolated maps. The map keyframe poses and their covariance matrices are generated by running the VIO algorithm Open-VINS [2]. For the maps S1, S2, and S3, the absolute trajectory errors (ATEs) of the map keyframe poses from Open-VINS are $0.298°/0.114m$, $0.165°/0.116m$ and $0.303°/0.172m$, respectively. The map features are randomly extracted and triangulated with the estimated keyframe poses. The triangulated features are then reprojected into the camera frame of the online running trajectory so that we can get the feature matches between the online running trajectory and the maps.

To test the effectiveness of our proposed algorithm (called Consistent MIMB-VIO, denoted as C-MIMB-VIO), we compare it with the one that fails to maintain the correct observability properties of the system (denoted as MIMB-VIO). Furthermore, Open-VINS [2] is used as a benchmark to show that by correctly fusing the multiple isolated maps, the accuracy of the localization can be improved. The overall results are given in Table II.

In Table II, ATE is used to measure the accuracy of the localization results, and normalized estimation error squared (NEES) [2] is used to measure the consistency of the localization results. The value before and after / are corresponding to the orientation (°) and the position (m), respectively. For ATE, the smaller the value, the better. For NEES, if its value is larger than 1, this means the system is over-optimistic about the estimated results, i.e., the system is inconsistent. For each sequence, there are three settings:

TABLE II

THE ATES $(degree/m)$ AND NEESS OF DIFFERENT ALGORITHMS ON SIMULATION DATA

Sequence \ Algorithm			Open-VINS	MIMB-VIO	C-MIMB-VIO
S1	0map	ATE	0.323/0.191	—	—
		NEES	0.301/0.687	—	—
	1map	ATE	—	1.142/0.545	0.282/0.154
		NEES	—	19.198/9.353	0.213/0.457
	2map	ATE	—	1.131/0.541	0.278/0.149
		NEES	—	19.401/9.445	0.191/0.410
	average	ATE	0.323/0.191	1.137/0.543	**0.280/0.152**
S2	0map	ATE	0.314/0.145	—	—
		NEES	0.357/0.430	—	—
	1map	ATE	—	0.931/0.466	0.275/0.129
		NEES	—	14.373/6.931	0.223/0.346
	2map	ATE	—	0.934/0.464	0.275/0.128
		NEES	—	14.914/7.124	0.208/0.295
	average	ATE	0.314/0.145	0.933/0.465	**0.275/0.129**
S3	0map	ATE	0.404/0.228	—	—
		NEES	0.407/0.772	—	—
	1map	ATE	—	0.693/0.353	0.345/0.198
		NEES	—	6.171/3.382	0.275/0.543
	2map	ATE	—	0.668/0.351	0.350/0.188
		NEES	—	5.941/3.446	0.249/0.426
	average	ATE	0.404/0.228	0.681/0.352	**0.348/0.193**

— means there is no such variable to evaluate.

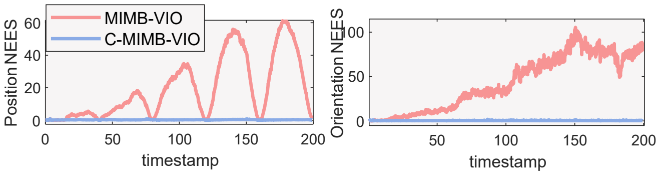


Fig. 4. The NEES of the local VIO poses

0map means there is no map; 1map means there is one map; 2map means there are two isolated maps. From Table II, we can find that the localization accuracy of MIMB-VIO is much worse than C-MIMB-VIO. Even with the map information, ATE of MIMB-VIO is larger than the pure odometry Openv-VINS. Also, the value of NEES is much larger than 1, which means the system is inconsistent. On the contrary, because C-MIMB-VIO can maintain the correct observability properties of the system, the localization accuracy is improved with the help of external map information, and the value of NEES indicates the good consistency of the system. Fig. 4 also gives the trend of NEES derived from MIMB-VIO and C-MIMB-VIO over time. We can find that the estimation of MIMB-VIO is more and more optimistic, i.e., the estimated uncertainty is much less than the true uncertainty, resulting in a bad estimation. Different from MIMB-VIO, C-MIMB-VIO always keeps NEES at a low value. Moreover, the error of the estimated local VIO pose with its 3σ bounds is plotted in Fig. 5. The consistent estimation from C-MIMB-VIO guarantees the estimated errors fall within the 3σ bounds, whereas the estimation error of MIMB-VIO is out of bounds.

B. Real-World Experiments

In this part, we will validate our proposed algorithm in two kinds of real-world datasets (c.f. Fig. 6), which cover the scenarios of the aerial vehicle (EuRoC [22]) and the ground vehicle in urban areas (Kaist [23]).

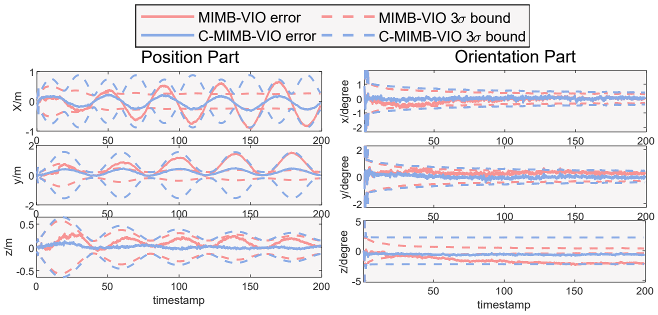
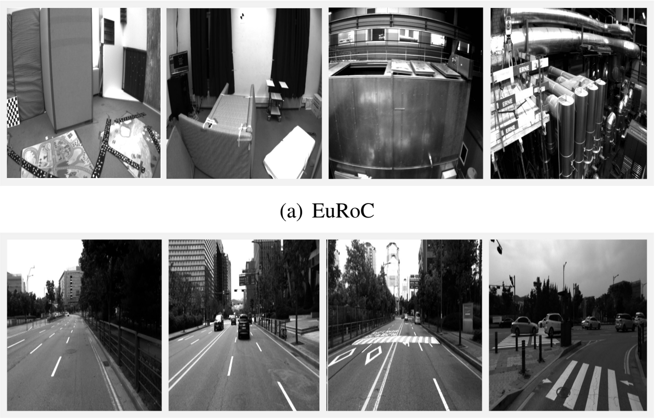


Fig. 5. The errors of the local VIO poses with 3σ bounds.



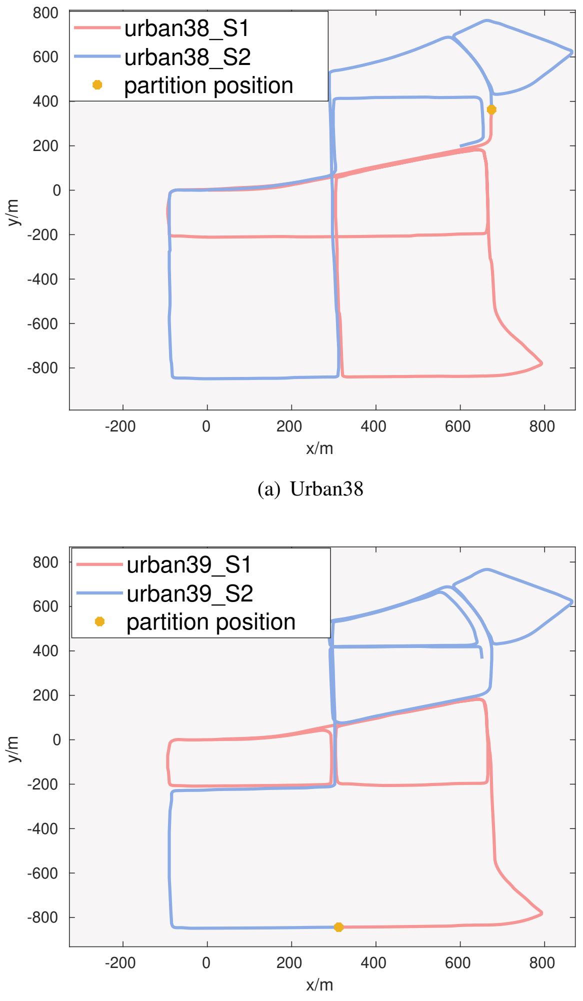
(a) EuRoC

(b) Kaist

Fig. 6. The vehicles and images sampled from the two datasets.

EuRoC: EuRoC [22] dataset contains three scenarios, vicon room 1 (V1), vicon room 2 (V2), and machine hall (MH). For our experiments, we utilize Open-VINS to generate map keyframe poses and their covariance matrices. The 3D map features are triangulated across multiple adjacent images and the corresponding poses. The feature matching procedure is conducted as follows: we first utilize R2D2 [20] to extract new features on the current queried frame and match them with features in map keyframes. The initial matching pairs based on descriptors are then fed into a robust pose estimator [21] to generate accurate feature matching pairs. For each scenario, if one of the sequences is used to query, the other sequences are employed as maps. For instance, if MH01 is the queried sequence, MH02-MH05 are used as maps.

Kaist: Kaist [23] is a dataset collected in urban on-the-road environments with many moving vehicles. In the dataset, there are two sequences, Urban38 and Urban39, having a large overlap. For each sequence, we divide its trajectory into two sessions (denoted as Urban38_S1, Urban38_S2, Urban39_S1, Urban39_S2), so that for each queried sequence, there are two isolated map sessions from the other sequence can be used. The partitioned trajectory of each sequence is plotted in Fig. 7. Because this dataset is challenging, Open-VINS will suffer significant drift and cannot be used to generate maps. Instead, we utilize the ground truth poses provided by this dataset as the keyframe poses of maps. As the ground truth built upon Virtual Reference Station (VRS)–GPS, we set the deviation of map keyframe pose as $0.1m$ for position parts and $2.87°$ $(0.05rad)$ for orientation parts. The procedures for generating 3D map



(a) Urban38

(b) Urban39

Fig. 7. The partitions of the trajectories in Kaist.

features and feature matches are the same as those in EuRoC dataset.

Experimental Results: To validate the accurate estimation of our proposed algorithm, we evaluate the estimated local VIO position (V.P.) as well as the relative transformations between the local VIO reference frame and the different map reference frames (*augmented variables*, A.V.). Besides, the results of the algorithm that regards the map as perfect (denoted as P-MIMB-VIO) are also given to demonstrate the necessity of considering the uncertainty of the map information. All the results are listed in Table III. For each queried sequence, we run experiments three times with the different numbers of maps. For example, for the queried sequence MH01, one (MH02) to four (MH02-MH05) maps will be used to perform localization. Therefore, the ATEs of V.P. for MH01 are the average of all the experimental results. Similarly, the ATEs of A.V. for MH01 are the average of the transformations between the VIO reference frame and reference frames of MH02-MH05 [6]. From Table III, we can find that C-MIMB-VIO has the best performance for almost all cases because it consistently fuses the maps. It is worth noting that even though the map information for Kaist [23] is generated by the ground truth, C-MIMB-VIO also gives better results than P-MIMB-VIO. A trajectory comparison on Urban38 is given by Fig. 8 to provide an intuitive feeling of the performance of different algorithms.

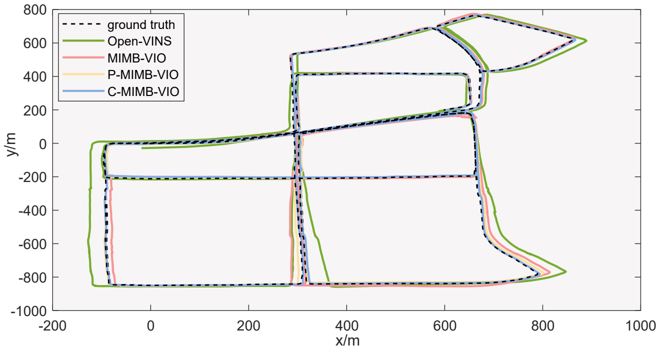


Fig. 8. The trajectory comparison on Kaist Urban38 with different algorithms

Moreover, the average time consumptions of each step for C-MIMB-VIO in four different scenarios are also given in Table IV, where the pure VIO Open-VINS is compared as a baseline to demonstrate the efficiency of our proposed algorithm.

VII. Conclusion

In this paper, we propose a framework that can fuse the information of multiple isolated maps into the VIO system, avoiding the tough task of building a big globally consistent map. By analyzing the observability properties of the system, we introduce the Schmidt updating and FEJ to consider the uncertainty of the maps efficiently and maintain the correct observability properties of the system, so that the system can be consistent and gives good localization results. According to simulation and real-world experiments, we demonstrate the consistency of our proposed algorithm and the necessity of keeping the system consistent.

Appendix I
Observability Analysis of the Augmented System

To analyze the observability properties of *augmented system*, we should compute the observability matrix (13), which contains the transition matrices and observation Jacobian matrices.

With the state (14) and the kinematics equation (2), we can derive the transition matrix $\mathbf{\Phi}_{t+1|t}$ defined by:

$$\mathbf{\Phi}_{t+1|t} = \begin{bmatrix} \mathbf{\Phi}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0}_{3\times 30} \\ \mathbf{\Phi}_2 & \mathbf{I} & \mathbf{0} & \mathbf{0}_{3\times 30} \\ \mathbf{\Phi}_3 & \mathbf{I}\Delta & \mathbf{I} & \mathbf{0}_{3\times 30} \\ \mathbf{0}_{30\times 3} & \mathbf{0}_{30\times 3} & \mathbf{0}_{30\times 3} & \mathbf{I}_{30} \end{bmatrix}, \tag{22}$$

[6] Table VI-B also gives the detailed results for each sequence with the different number of used maps, and Table VI gives the detailed results of the *augmented variables* between the queried sequence and the used pre-built maps. Table III only lists the average data.

TABLE III

ATEs(m) OF LOCAL VIO POSITION (V.P.), ATEs($degree/m$) OF ROTATION PART (R.) AND POSITION PART (P.) OF *augmented variable* (A.V.)

Algorithm		Sequence	MH01	MH02	MH03	MH04	MH05	V101	V102	V103	V201	V202	V203	Urban38	Urban39
V.P.		Open-VINS	0.090	0.101	0.144	0.190	0.246	0.046	0.058	0.047	0.073	0.059	0.117	24.886	7.061
		MIMB-VIO	0.070	0.063	0.152	0.193	0.280	0.039	0.074	0.047	0.068	0.104	0.083	13.773	10.539
		P-MIMB-VIO	0.073	0.074	0.128	0.229	0.218	0.056	0.050	0.081	0.071	0.123	0.105	5.969	5.614
		C-MIMB-VIO	**0.056**	**0.059**	**0.085**	**0.161**	**0.187**	**0.032**	**0.043**	**0.043**	**0.066**	**0.056**	**0.082**	**5.205**	**5.376**
A.V.	R.	MIMB-VIO	1.657	3.126	1.546	3.398	1.747	4.662	5.230	2.169	2.931	9.349	2.494	1.176	1.453
		P-MIMB-VIO	1.567	**2.233**	1.256	1.784	2.031	1.364	**0.736**	1.446	2.220	3.300	**1.851**	0.877	0.760
		C-MIMB-VIO	**1.304**	2.301	**1.149**	**1.237**	**1.388**	**0.972**	0.825	**0.993**	**1.452**	**3.093**	2.038	**0.795**	**0.746**
	P.	MIMB-VIO	0.138	0.162	0.137	0.189	0.391	0.049	0.118	0.076	0.099	0.230	0.124	6.437	2.922
		P-MIMB-VIO	0.202	0.173	0.147	0.228	0.544	0.067	0.112	0.114	0.134	**0.186**	0.133	5.138	1.389
		C-MIMB-VIO	**0.123**	**0.147**	**0.099**	**0.182**	**0.342**	**0.038**	**0.107**	**0.044**	**0.093**	0.209	**0.122**	**4.389**	**1.387**

TABLE IV

THE TIME CONSUMPTION (s) OF EACH STEP ON REAL-WORLD DATASETS

Algorithm / Dataset	MH	V1	V2	urban	average
Open-VINS	0.017	0.016	0.017	0.016	0.017
C-MIMB-VIO	0.024	0.019	0.020	0.024	0.022

where

$$\begin{aligned}
\mathbf{\Phi}_1 &= {}^{I_{t+1}}\hat{\mathbf{R}}_L {}^{I_t}\hat{\mathbf{R}}_L^\top, \\
\mathbf{\Phi}_2 &= -({}^L\hat{\mathbf{v}}_{I_{t+1}} - {}^L\hat{\mathbf{v}}_{I_t} + \mathbf{g}\Delta)_\times {}^{I_t}\hat{\mathbf{R}}_L^\top, \\
\mathbf{\Phi}_3 &= -({}^L\hat{\mathbf{p}}_{I_{t+1}} - {}^L\hat{\mathbf{p}}_{I_t} - {}^L\hat{\mathbf{v}}_{I_t}\Delta + \frac{1}{2}\mathbf{g}\Delta^2)_\times {}^{I_t}\hat{\mathbf{R}}_L^\top
\end{aligned}$$

Δ is one time step from time step t to time step $t+1$. The operation $(\cdot)_\times$ transforms a vector into a skew-symmetric matrix. All the $\mathbf{0}$ and $\mathbf{I}$ without the right subscript are with the dimension of 3.

From (5), we have the local observation Jacobian $\mathbf{H}_{L_t}$ defined by:

$$\mathbf{H}_{L_t} = \mathbf{H}_\pi^i \left[({}^{I_t}\hat{\mathbf{R}}_L({}^L\hat{\mathbf{f}}_{i_t} - {}^L\hat{\mathbf{p}}_{I_t}))_\times \ \mathbf{0} \ -{}^{I_t}\hat{\mathbf{R}}_L \ {}^{I_t}\hat{\mathbf{R}}_L \ \mathbf{0}_{3\times 27}\right] \tag{23}$$

where $\mathbf{H}_\pi^i$ is the Jacobian of the camera projection function, whose dimension is 2×3. The detailed expression is unimportant and omitted here.

For the map-based observation Jaocbian $\mathbf{H}_{G_t}$, it consists of three parts (denoted as ${}^{cam}\mathbf{H}_{G_t}$, ${}^{{}^kKF_j}\mathbf{H}_{G_t}$ and ${}^{{}^sKF_l}\mathbf{H}_{G_t}$) that are derived from (10)-(12):

$${}^{cam}\mathbf{H}_{G_t} = \mathbf{H}_\pi^{ki}\left[H_1 \ \mathbf{0} \ -{}^{I_t}\hat{\mathbf{R}}_L \ \mathbf{0} \ H_2 \ -H_3 \ \mathbf{0}_{3\times 6} \ H_3 \ \mathbf{0}_{3\times 12}\right] \tag{24}$$

$${}^{{}^kKF_j}\mathbf{H}_{G_t} = \mathbf{H}_\pi^{ki'}\left[\mathbf{0}_{3\times 24} \ {}^{G_t^k}\hat{\mathbf{R}}_{kf_j}^\top \ H_4 \ -{}^{G_t^k}\hat{\mathbf{R}}_{kf_j}^\top \ \mathbf{0}_{3\times 6}\right] \tag{25}$$

$$\begin{aligned}{}^{{}^sKF_l}\mathbf{H}_{G_t} = \mathbf{H}_\pi^{ki''}\Big[&\mathbf{0}_{3\times 12} \ H_5 \ -H_6 \ H_7 \ {}^{G_t^s}\hat{\mathbf{R}}_{kf_l}^\top \ H_6 \\ &\mathbf{0}_{3\times 6} \ H_8 \ -{}^{G_t^s}\hat{\mathbf{R}}_{kf_l}^\top\Big]\end{aligned} \tag{26}$$

$$\begin{aligned}
H_1 &= \left[{}^{I_t}\hat{\mathbf{R}}_L({}^{G_t^k}\hat{\mathbf{R}}_L^\top({}^{G^k}\hat{\mathbf{F}}_{i_t} - {}^{G_t^k}\hat{\mathbf{p}}_L) - {}^L\hat{\mathbf{p}}_{I_t})\right]_\times \\
H_2 &= {}^{I_t}\hat{\mathbf{R}}_L {}^{G_t^k}\hat{\mathbf{R}}_L^\top({}^{G^k}\hat{\mathbf{F}}_{i_t} - {}^{G_t^k}\hat{\mathbf{p}}_L)_\times \quad H_3 = {}^{I_t}\hat{\mathbf{R}}_L {}^{G_t^i}\hat{\mathbf{R}}_L^\top \\
H_4 &= -{}^{G_t^k}\hat{\mathbf{R}}_{kf_j}^\top({}^{G^k}\hat{\mathbf{F}}_{i_t} - {}^{G_t^k}\hat{\mathbf{p}}_{kf_j})_\times \\
H_5 &= -{}^{G_t^s}\hat{\mathbf{R}}_{kf_l}^\top {}^{G_t^s}\hat{\mathbf{R}}_L {}^{G_t^k}\hat{\mathbf{R}}_L^\top({}^{G^k}\hat{\mathbf{F}}_{i_t} - {}^{G_t^k}\hat{\mathbf{p}}_L)_\times \\
H_6 &= {}^{G_t^s}\hat{\mathbf{R}}_{kf_l}^\top {}^{G_t^s}\hat{\mathbf{R}}_L {}^{G_t^k}\hat{\mathbf{R}}_L^\top \\
H_7 &= {}^{G_t^s}\hat{\mathbf{R}}_{kf_l}^\top\left[{}^{G_t^s}\hat{\mathbf{R}}_L {}^{G_t^k}\hat{\mathbf{R}}_L^\top({}^{G^k}\hat{\mathbf{F}}_{i_t} - {}^{G_t^k}\hat{\mathbf{p}}_L)\right]_\times \\
H_8 &= -{}^{G_t^s}\hat{\mathbf{R}}_{kf_l}^\top\left[{}^{G_t^s}\hat{\mathbf{R}}_L {}^{G_t^k}\hat{\mathbf{R}}_L^\top({}^{G^k}\hat{\mathbf{F}}_{i_t} - {}^{G_t^k}\hat{\mathbf{p}}_L) + {}^{G_t^s}\hat{\mathbf{p}}_L - {}^{G_t^s}\hat{\mathbf{p}}_{kf_l}\right]_\times
\end{aligned}$$

We first consider the ideal case that all the estimated values are equal to the true values. With this assumption, we can investigate the observability properties that the system should have. Then, on this basis, we will further demonstrate the inconsistency of the system under the actual case.

The Proof of Theorem 1

Proof: Under the ideal case, the following two conditions are satisfied: a) The propagated state $\hat{\mathbf{x}}_{Aug_{t|t-1}}$ is equal to the updated state $\hat{\mathbf{x}}_{Aug_t}$ and is equal to the true state $\mathbf{x}_{Aug_t}$; b) All the features, *augmented variables*, and map keyframe poses do not change with time and are equal to the constant true values. For instance, the map feature ${}^{G^k}\hat{\mathbf{F}}_{i_t} = {}^{G^k}\mathbf{F}_i$.

With condition a) and (22), the transition matrix $\mathbf{\Phi}_{t|0}$ can be given as:

$$\mathbf{\Phi}_{t|0} = \begin{bmatrix} \mathbf{\Phi}(1) & \mathbf{0} & \mathbf{0} & \mathbf{0}_{3\times 30} \\ \mathbf{\Phi}(2) & \mathbf{I} & \mathbf{0} & \mathbf{0}_{3\times 30} \\ \mathbf{\Phi}(3) & \mathbf{I}\Delta_t & \mathbf{I} & \mathbf{0}_{3\times 30} \\ \mathbf{0}_{30\times 3} & \mathbf{0}_{30\times 3} & \mathbf{0}_{30\times 3} & \mathbf{I}_{30} \end{bmatrix}, \tag{27}$$

where Δ_t is t time steps from time setp 0 to time setp t,

$$\begin{aligned}
\mathbf{\Phi}(1) &= {}^{I_t}\mathbf{R}_L {}^{I_0}\mathbf{R}_L^\top, \\
\mathbf{\Phi}(2) &= -({}^L\mathbf{v}_{I_t} - {}^L\mathbf{v}_{I_0} + \mathbf{g}\Delta_t)_\times {}^{I_0}\mathbf{R}_L^\top, \\
\mathbf{\Phi}(3) &= -({}^L\mathbf{p}_{I_t} - {}^L\mathbf{p}_{I_0} - {}^L\mathbf{v}_{I_0}\Delta_t + \frac{1}{2}\mathbf{g}\Delta_t^2)_\times {}^{I_0}\mathbf{R}_L^\top.
\end{aligned}$$

Substituting (23)-(29) into (13), we can get the observability matrix, whose right null space is given by (15).

We can find that for the ideal case, the dimension of the unobservable subspace is ten: four dimensions for the VIO system, and six dimensions for the introduced *imperfect* maps. ■

The Proof of Theorem 2

TABLE V

ATE(M) OF THE LOCAL VIO POSITION ON DIFFERENT DATASETS WITH DIFFERENT ALGORITHMS

Sequence \ Algorithm		Open -VINS	MIMB -VIO	P-MIMB -VIO	C-MIMB -VIO
MH01	0map	0.090	—	—	—
	1map	—	0.067	0.062	0.066
	2map	—	0.086	0.080	0.053
	3map	—	0.062	0.078	0.057
	4map	—	0.065	0.070	0.048
	average	0.090	0.070	0.073	**0.056**
MH02	0map	0.101	—	—	—
	1map	—	0.070	0.074	0.066
	2map	—	0.064	0.070	0.060
	3map	—	0.059	0.073	0.051
	4map	—	0.060	0.077	0.058
	average	0.101	0.063	0.074	**0.059**
MH03	0map	0.144	—	—	—
	1map	—	0.187	0.125	0.083
	2map	—	0.174	0.137	0.113
	3map	—	0.130	0.122	0.072
	4map	—	0.115	0.127	0.073
	average	0.144	0.152	0.128	**0.085**
MH04	0map	0.190	—	—	—
	1map	—	0.199	0.225	0.163
	2map	—	0.158	0.253	0.159
	3map	—	0.168	0.225	0.157
	4map	—	0.247	0.212	0.164
	average	0.190	0.193	0.229	**0.161**
MH05	0map	0.246	—	—	—
	1map	—	0.301	0.224	0.186
	2map	—	0.317	0.239	0.189
	3map	—	0.266	0.264	0.183
	4map	—	0.237	0.145	0.191
	average	0.246	0.280	0.218	**0.187**
V101	0map	0.046	—	—	—
	1map	—	0.040	0.058	0.032
	2map	—	0.037	0.054	0.032
	average	0.046	0.039	0.056	**0.032**
V102	0map	0.058	—	—	—
	1map	—	0.072	0.044	0.044
	2map	—	0.075	0.056	0.041
	average	0.058	0.074	0.050	**0.043**
V103	0map	0.047	—	—	—
	1map	—	0.047	0.058	0.043
	2map	—	0.046	0.104	0.042
	average	0.047	0.047	0.081	**0.043**
V201	0map	0.073	—	—	—
	1map	—	0.070	0.069	0.069
	2map	—	0.066	0.073	0.063
	average	0.073	0.068	0.071	**0.066**
V202	0map	0.059	—	—	—
	1map	—	0.093	0.132	0.049
	2map	—	0.117	0.114	0.062
	average	0.059	0.104	0.123	**0.056**
V203	0map	0.117	—	—	—
	1map	—	0.086	0.106	0.084
	2map	—	0.080	0.103	0.079
	average	0.117	0.083	0.105**0.082**	
Urban38	0map	24.886	—	—	—
	2map	—	13.773	5.969	5.205
	average	24.886	13.773	5.969	**5.205**
Urban39	0map	7.061	—	—	—
	2map	—	10.539	5.614	5.376
	average	7.061	10.539	5.614	**5.376**

TABLE VI

ATE(DEGREE/M) OF THE *augmented variables* ON DIFFERENT DATASETS WITH DIFFERENT ALGORITHMS

queried sequence	map sequence	MIMB -VIO	P-MIMB -VIO	C-MIMB -VIO
MH01	MH02	2.730/0.113	2.652/0.202	**2.147/0.100**
	MH03	1.372/0.141	**0.598**/0.224	0.808/**0.109**
	MH04	1.358/0.129	1.419/**0.120**	**1.334**/0.131
	MH05	1.167/0.169	1.598/0.263	**0.927/0.150**
	average	1.657/0.138	1.567/0.202	**1.304/0.123**
MH02	MH01	3.740/0.203	**2.472/0.153**	2.842/0.160
	MH03	3.134/0.136	2.421/0.118	**2.320/0.106**
	MH04	2.642/**0.141**	**1.795**/0.197	1.916/0.160
	MH05	2.986/0.167	2.242/0.255	**2.127/0.163**
	average	3.126/0.162	**2.233**/0.173	2.301/**0.147**
MH03	MH01	1.357/0.158	0.929/0.166	**0.672/0.096**
	MH02	**2.150**/0.183	2.735/**0.110**	2.502/0.134
	MH04	1.247/0.092	0.790/0.096	**0.724/0.084**
	MH05	1.429/0.116	**0.568**0.214	0.696/**0.080**
	average	1.546/0.137	1.256/0.147	**1.149/0.099**
MH04	MH01	3.479/**0.272**	2.264/0.333	**1.393**/0.274
	MH02	3.801/0.149	2.195/0.222	**1.923/0.137**
	MH03	3.204/0.162	1.553/0.269	**1.005/0.154**
	MH05	3.107/0.172	1.123/**0.089**	**0.628**/0.163
	average	3.398/0.189	1.784/0.228	**1.237/0.182**
MH05	MH01	1.606/0.413	2.416/0.625	**1.483/0.370**
	MH02	2.276/0.430	2.323/0.577	**1.983/0.341**
	MH03	1.911/0.442	2.516/0.595	**1.439/0.433**
	MH04	1.196/0.280	0.869/0.378	**0.648/0.222**
	average	1.747/0.391	2.031/0.544	**1.388/0.342**
V101	V102	4.810/0.047	1.777/0.075	**1.032/0.038**
	V103	4.513/0.051	0.950/0.059	**0.911/0.038**
	average	4.662/0.049	1.364/0.067	**0.972/0.038**
V102	V101	5.302/0.108	**0.709/0.090**	0.968/0.093
	V103	5.157/0.127	0.762/0.133	**0.681/0.120**
	average	5.230/0.118	**0.736**/0.112	0.825/**0.107**
V103	V101	2.312/0.076	1.665/0.114	**1.279/0.055**
	V102	2.026/0.076	1.226/0.113	**0.707/0.033**
	average	2.169/0.076	1.446/0.114	**0.993/0.044**
V201	V202	2.946/0.102	2.025/0.140	**1.426/0.079**
	V203	2.916/**0.096**	2.415/0.128	**1.477**/0.106
	average	2.931/0.099	2.220/0.134	**1.452/0.093**
V202	V201	6.778/**0.233**	3.000/0.233	**2.837**/0.244
	V203	11.920/0.227	3.598/0.138	**3.349/0.173**
	average	9.349/0.230	3.300/**0.186**	**3.093**/0.209
V203	V201	2.831/**0.112**	2.061/0.122	**2.164/0.112**
	V202	2.156/0.135	1.640/0.144	**1.911/0.131**
	average	2.494/0.124	**1.851**/0.133	2.039/**0.122**
Urban38	Urban39_S1	1.556/7.003	1.295/6.161	**1.164/5.861**
	Urban39_S2	0.796/5.871	0.468/4.115	**0.427/3.817**
	average	1.176/6.437	0.877/5.138	**0.795/4.839**
Urban39	Urban38_S1	1.516/3.374	**0.634**/1.400	**0.634/1.310**
	Urban38_S2	1.390/2.470	0.887/**1.375**	0.854/1.463
	average	1.453/2.922	0.760/1.389	**0.746/1.387**

Proof: Under the actual case, conditions a) and b) are not satisfied. Without condition a), $\mathbf{\Phi}$ does not have the elegant form of (29), which leads to the vanishment of the first column block of (15). Without condition b), element values that are not supposed to change over time will become ever-changing estimated values. That is to say, the map feature ${}^{G^k}\hat{\mathbf{F}}_{i_t} \neq {}^{G^k}\hat{\mathbf{F}}_{i_{t-1}} \neq \cdots \neq {}^{G^k}\hat{\mathbf{F}}_{i_0} \neq {}^{G^k}\mathbf{F}_i$, so as ${}^{G^k}\mathbf{R}_L$, ${}^{G^s}\mathbf{R}_L$, and ${}^{G^k}\mathbf{p}_L$, which leads to the vanishment of the second and the third column block of (15). Therefore, for the actual case, the dimension of the unobservable subspace is only three given by (16), not in accordance with the ideal case. This means if we use the estimated values to compute Jacobian matrices, the *augmented system* will suffer from inconsistency. ■

When the used maps are *perfect*, the original state (14) shrinks to

$$\mathbf{x}^{*}_{Aug_t}=\left[{}^{I_t}\mathbf{q}_L^\top \; {}^{L}\mathbf{v}_{I_t}^\top \; {}^{L}\mathbf{p}_{I_t}^\top \;\; {}^{L}\mathbf{f}_{i_t}^\top | {}^{G_t^k}\mathbf{q}_L^\top \;\; {}^{G_t^k}\mathbf{p}_L^\top \;\; {}^{G_t^s}\mathbf{q}_L^\top \; {}^{G_t^s}\mathbf{p}_L^\top\right]^\top \tag{28}$$

where the map-related parts are removed. Then, the dimension of the the state is reduced from 39 to 24.

The Proof of Theorem 3

Proof: With the proof of Theorem 1, the derivation of Theorem 3 is straightforward.

Since the dimension of the state changes, so does that of the transition matrix $\mathbf{\Phi}_{t|0}$. Therefore, (22) becomes to

$$\mathbf{\Phi}^{*}_{t|0} = \begin{bmatrix} \mathbf{\Phi}(1) & \mathbf{0} & \mathbf{0} & \mathbf{0}_{3\times 15} \\ \mathbf{\Phi}(2) & \mathbf{I} & \mathbf{0} & \mathbf{0}_{3\times 15} \\ \mathbf{\Phi}(3) & \mathbf{I}\Delta_t & \mathbf{I} & \mathbf{0}_{3\times 15} \\ \mathbf{0}_{15\times 3} & \mathbf{0}_{15\times 3} & \mathbf{0}_{15\times 3} & \mathbf{I}_{15} \end{bmatrix}, \tag{29}$$

where the key elements, i.e., the upper-left 9×9 blocks of the transition matrix, remain the same.

Besides, for (23)-(26), the map-related Jacobian blocks are also removed.

Based on these analyses and the unobservable subspace getten from Theorem 1, we can easily derive that the unobservable subspace of the *augmented system* with *perfect* maps is given by (17). ■

The Proof of Theorem 4

Proof: As analyzed in the proof of Theorem 2, under the actual case, the first column block of (17) vanishes because the updated state $\hat{\mathbf{x}}^{*}_{Aug_t}$ is not equal to the true state $\mathbf{x}^{*}_{Aug_t}$. Besides, the ever-changing values of the estimated *augmented variables* make the second column block of (17) vanish. ■

References

[1] A. I. Mourikis and S. I. Roumeliotis, "A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation," *Proceedings 2007 IEEE International Conference on Robotics and Automation*, Roma, 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," *2020 IEEE International Conference on Robotics and Automation (ICRA)*, Paris, France, 2020, pp. 4666-4672, doi: 10.1109/ICRA40945.2020.9196524.

[3] M. Bloesch, M. Burri, S. Omari, M. Hutter and R. Siegwart. "Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback". The International Journal of Robotics Research, vol. 36, no. 10, pp. 1053–1072, 2017. https://doi.org/10.1177/0278364917728574

[4] T. Qin, P. Li and S. Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," in *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004-1020, Aug. 2018, doi: 10.1109/TRO.2018.2853729.

[5] C. Forster, Z. Zhang, M. Gassner, M. Werlberger and D. Scaramuzza, "SVO: Semidirect Visual Odometry for Monocular and Multicamera Systems," in IEEE Transactions on Robotics, vol. 33, no. 2, pp. 249-265, April 2017, doi: 10.1109/TRO.2016.2623335.

[6] J. A. Hesch, D. G. Kottas, S. L. Bowman and S. I. Roumeliotis, "Consistency Analysis and Improvement of Vision-aided Inertial Navigation," in *IEEE Transactions on Robotics*, vol. 30, no. 1, pp. 158-176, Feb. 2014, doi: 10.1109/TRO.2013.2277549.

[7] G. Cioffi, and D. Scaramuzza, "Tightly-coupled Fusion of Global Positional Measurements in Optimization-based Visual-Inertial Odometry," arXiv e-prints, 2020, retrieved on March 9th 2020. [Online]. Available: https://arxiv.org/abs/2003.04159

[8] T. Schneider *et al.*, "Maplab: An Open Framework for Research in Visual-Inertial Mapping and Localization," in IEEE Robotics and Automation Letters, vol. 3, no. 3, pp. 1418-1425, July 2018, doi: 10.1109/LRA.2018.2800113.

[9] R. C. DuToit, J. A. Hesch, E. D. Nerurkar and S. I. Roumeliotis, "Consistent map-based 3D localization on mobile devices," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, 2017, pp. 6253-6260, doi: 10.1109/ICRA.2017.7989741.

[10] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM," arXiv e-prints, 2020, retrieved on July 23rd 2020. [Online]. Available: https://arxiv.org/abs/2007.11898v1

[11] M. Burri, H. Oleynikova, M. W. Achtelik and R. Siegwart, "Real-time visual-inertial mapping, re-localization and planning onboard MAVs in unknown environments," 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 1872-1878, doi: 10.1109/IROS.2015.7353622.

[12] M. Labbé, F. Michaud, "RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," Journal of Field Robotics, vol. 35, pp. 416– 446, 2019, doi: 10.1002/rob.21831

[13] A. Fisher, R. Cannizzaro, M. Cochrane, C. Nagahawatte, J. Palmer, " ColMap: A memory-efficient occupancy grid mapping framework," Robotics and Autonomous Systems, vol 142, 2021, doi: 10.1016/j.robot.2021.103755.

[14] S. Huang, Z. Wang and G. Dissanayake, "Sparse Local Submap Joining Filter for Building Large-Scale Maps," in IEEE Transactions on Robotics, vol. 24, no. 5, pp. 1121-1130, Oct. 2008, doi: 10.1109/TRO.2008.2003259.

[15] P. Geneva, K. Eckenhoff and G. Huang, "A Linear-Complexity EKF for Visual-Inertial Navigation with Loop Closures," *2019 International Conference on Robotics and Automation (ICRA)*, Montreal, QC, Canada, 2019, pp. 3535-3541, doi: 10.1109/ICRA.2019.8793836.

[16] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," arXiv e-prints, 2019, retrieved on March 1st, 2020. [Online]. Available: https://arxiv.org/abs/1901.03642v

[17] L. Delobel, R. Aufrère, C. Debain, R. Chapuis and T. Chateau, "A Real-Time Map Refinement Method Using a Multi-Sensor Localization Framework," in IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 5, pp. 1644-1658, May 2019, doi: 10.1109/TITS.2018.2840822

[18] G. P. Huang, A. I. Mourikis and S. I. Roumeliotis, "Analysis and improvement of the consistency of extended Kalman filter based SLAM," 2008 IEEE International Conference on Robotics and Automation, 2008, pp. 473-479, doi: 10.1109/ROBOT.2008.4543252.

[19] K. Wu, T. Zhang, D. Su, S. Huang and G. Dissanayake, "An invariant-EKF VINS algorithm for improving consistency," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 1578-1585, doi: 10.1109/IROS.2017.8205965.

[20] J. Revaud *et al.*"R2d2: repeatable and reliable detector and descriptor," arXiv e-prints, 2019, retrieved on June 17th, 2019. [Online]. Available: https://arxiv.org/abs/1906.06195

[21] Y. Jiao *et al.*, "2-Entity RANSAC for robust visual localization in changing environment," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Macau, China, 2019, pp. 2478-2485, doi: 10.1109/IROS40897.2019.8967671.

[22] B. Siciliano et al., "EuRoC - The Challenge Initiative for European Robotics," *ISR/Robotik 2014; 41st International Symposium on Robotics*, Munich, Germany, 2014, pp. 1-7.

[23] J. Jeong ,Y. Cho, Y.S. Shin, et al. "Complex urban dataset with multi-level sensors from highly diverse urban environments", *The International Journal of Robotics Research*, 2019, 38(6):027836491984399.