Online Self-Calibration for Visual-Inertial Navigation: Models, Analysis and Degeneracy

Yulin Yang[1], Patrick Geneva[1], Xingxing Zuo[2], and Guoquan Huang[1], *Senior Member, IEEE,*

***Abstract*—As sensor calibration plays an important role in visual-inertial sensor fusion, this paper performs an in-depth investigation of online self-calibration for robust and accurate visual-inertial state estimation. To this end, we first conduct complete observability analysis for visual-inertial navigation systems (VINS) with full calibration of sensing parameters, including IMU/camera intrinsics and IMU-camera spatial-temporal extrinsic calibration, along with readout time of rolling shutter (RS) cameras (if used). We study different inertial model variants containing intrinsic parameters that encompass most commonly used models for low-cost inertial sensors. With these models, the observability analysis of linearized VINS with full sensor calibration is performed. Our analysis theoretically proves the intuition commonly assumed in the literature – that is, VINS with full sensor calibration has four unobservable directions, corresponding to the system's global yaw and position, while all sensor calibration parameters are observable given fully-excited motions. Moreover, we, for the first time, identify degenerate motion primitives for IMU and camera intrinsic calibration, which, when combined, may produce complex degenerate motions. We compare the proposed *online* self-calibration on commonly-used IMUs against the state-of-art *offline* calibration toolbox Kalibr, showing that the proposed system achieves better consistency and repeatability. Based on our analysis and experimental evaluations, we also offer practical guidelines to effectively perform online IMU-camera self-calibration in practice.**

***Index Terms*—Sensor self-calibration, visual inertial systems, state estimation, observability analysis, degenerate motions**

I. Introduction

Due to the decreasing cost of integrated inertial/visual sensor rigs, visual-inertial navigation system (VINS) – which fuses high-rate inertial readings from an IMU and images of the surrounding environment from a camera – has gained great popularity in 6 degree-of-freedom (6-DoF) motion tracking for mobile devices and autonomous vehicles [1], such as micro aerial vehicles (MAV) [2], self-driving cars [3], unmanned ground vehicles (UGV) [4], [5] and smart phones [6], [7]. Many efficient and robust VINS algorithms have been developed in recent years, either based on filtering [8]–[11] or batch least-squares optimization [12]–[14].

There are many factors which attribute to VINS performance, such as visual feature tracking, velocity initialization and sensor calibration. Among them, robust and accurate sensor calibration – including the rigid transformation between sensors (spatial calibration), time offset between IMU-camera (temporal calibration), image line readout time for rolling shutter (RS) cameras, and IMU/camera intrinsics – is crucial, especially when plug-and-play visual-inertial sensor rigs with widely available off-the-shelf low-cost IMUs and RS cameras are used. In addition, sensor configuration itself can change slowly due to extended usage, sensor replacement, non-rigid sensor mounting, mechanical vibrations, environmental effects such as varying temperature, humidity, and among others. For example, IMU biases and intrinsics suffer from temperature and humidity changes [15], and rigid transformation between IMU and camera can vary if the sensor is replaced, reassembled or subjected to vibration. As such, online sensor self-calibration in VINS has attracted significant research efforts in recent years [6], [7], [11], [16]–[18], due to its potential to handle poor prior calibration or calibration changes, which can degrade estimation accuracy in the case where these calibrations are blindly treated to be true.

System observability analysis for VINS with online IMU-camera [19]–[21] or IMU/camera intrinsic [17], [22] calibration has also been carried out to show that these calibration parameters can be identified given fully excited motions. System observability can also be leveraged to improve motion planning for robust self-calibration [23]. Recent research efforts [16], [17] have investigated degenerate motions (e.g., planar motion or one-axis rotation) that might cause certain calibration parameters unobservable. However, comprehensive degeneracy analysis for VINS with full calibration parameters – including IMU/camera intrinsics, IMU-camera rigid transformation, temporal time offset, and camera RS readout time – is still missing in the existing literature. In this paper, we seek to bridge this significant gap.

Blindly performing online calibration is risky, as in most cases domain knowledge on specific motions and prior distribution choices are needed to ensure calibration can converge consistently [24]. In the meantime, existing research efforts [16], [17], [25] have also identified degenerate motions that cause online self-calibration to fail. Most approaches on VINS sensor self-calibration are limited to either handheld or trajectory segments involving rich motion information [15], [24]. This paper deeply focuses on degenerate motions that impact the deployment of VINS on mobile robots which typically have constrained motions, when jointly estimating IMU/camera intrinsics, IMU-camera spatial-temporal calibration, and RS readout time.

Our recent work [17] performed observability analysis for monocular VINS with only IMU intrinsic calibration (including scale and axis-misalignment for gyroscope and accelerom-

Manuscript received on October 5, 2022; revised on February 2, 2023; accepted on April 17, 2023. Date of publication May 17, 2023. This article was recommended for publication by Editor F. Chaumette upon evaluation of the reviews comments. This work was partially supported by the University of Delaware (UD) College of Engineering, the NSF (IIS-1924897, MRI-2018905, SCH-2014264). Yang was partially supported by the University Doctoral Fellowship, while Geneva was partially supported by the Delaware Space Grant Fellowship.

[1]The authors are with the Robot Perception and Navigation Group, University of Delaware, Newark, DE 19716, USA. E-mail:{yuyang, pgeneva, ghuang}@udel.edu

[2]The author is with the Department of Informatics, Technical University of Munich, Munich, Germany. E-mail:{xingxing.zuo}@tum.de

eter, but without g-sensitivity) and identified their degenerate motions. In this work, building upon these results, we develop an accurate and robust monocular VINS estimator with full self-calibration, while extending that observability analysis to visual-inertial self-calibration systems and performing degenerate motion analysis of *all* calibration parameters.

Specifically, to highlight the difference from our prior conference publication [17], in this paper, we have incorporated full calibration parameters including g-sensitivity in IMU models, camera intrinsics, and readout time of RS cameras, all of which are missing in [17]. We have performed extensive numerical studies of IMU model variants with g-sensitivity on four typical trajectories in simulations. We also thoroughly evaluate the proposed method using both the public benchmarking datasets [26] and our own datasets, capturing both fully-excited and degenerate motions for online calibration. Additionally, we perform a fair comparison to Kalibr [27] for the first time, in order to further validate the accuracy and convergence of the proposed online self-calibration.

In particular, the main contributions of this work include:

- An efficient filter-based visual-inertial estimator capable of performing self-calibration for all spatial-temporal extrinsic and intrinsic calibration parameters.
- We perform a comprehensive observability and degeneracy analysis for the proposed visual-inertial models and, *for the first time*, identify the degenerate motions that cause IMU and camera intrinsic parameters to be unobservable.
- Extensive simulations and real-world experiments are performed to verify the parameter convergence of the estimator with online self-calibration under fully-excited 6-DoF motion and a series of identified degenerate motions of practical significance. Additionally, we show that degenerate motions can and do have a significant negative impact on the performance of the estimator, leading to a series of recommendation guidelines.

II. Related Work

A. IMU Intrinsic Calibration

Generally, the gyroscope and acceleration biases are needed for accurate inertial modeling. It is a common practice to estimate biases online in VINS such as [8], [20], [28], [29]. Besides these biases, the IMU intrinsic parameters – including the scale correction and axis misalignment for gyroscope and accelerometer, the rotation from gyroscope or accelerometer frame to IMU frame, and the g-sensitivity – also need to be calibrated offline or online, especially for low-cost inertial sensors. Xiao et al. [30] improved the IMU pre-integration [12] to incorporate the IMU intrinsic parameters in a keyframe based VINS algorithm for online self-calibration. Jung et al. [31] studied IMU intrinsic calibration within multi-state constrained Kalman filter (MSCKF [8]) by using a stereo camera and an IMU sensor, where they also examined the inertial calibration results under planar and random motions.

Building upon our prior work [17], in which we have investigated online IMU intrinsic calibration with the minimal sensor configuration of a single IMU and a monocular camera and compared the performance of four different IMU intrinsic model variants in VINS, in this work, we perform online self-calibration and study 18 different IMU intrinsic model variants which can encompass or be equivalent to most published IMU models for inertial navigation. Comprehensive analysis of degenerate motions, which can cause online self-calibration to fail, is also provided.

B. Joint IMU-Camera Self-Calibration

Extensive works have studied joint sensor calibration in VINS. For instance, Mirzaei et al. [19] proposed to use an extended Kalman filter (EKF) for the spatial calibration (i.e., the rigid transformation between the camera and IMU) of VINS and performed an observability analysis. They showed that the rigid transformation is not fully observable under one-axis rotation. However, the camera intrinsics or IMU-camera time offset are not calibrated and chessboards are needed for calibration. Furgale et al. [27] developed the well-known calibration toolbox: Kalibr, a continuous-time spline-based batch estimator, for IMU-camera extrinsics, time offset and camera intrinsics calibration. Rehder et al. [32] extended Kalibr to incorporate IMU intrinsics (including scaling parameters, axis misalignments, and g-sensitivity). The above mentioned works are all offline methods and need calibration targets. In addition, they do not support full-parameter joint optimization of camera intrinsics with other calibration parameters. Schneider et al. [24] reduced optimization complexity for IMU-camera calibration by selecting the most informative trajectory segments for calibration.

Many recent filter based VINS algorithms perform online IMU-camera joint calibration. Guo et al. [7] proposed to use linear pose interpolation to model RS effects and calibrate readout time. Eckenhoff et al. [11] proposed a generalized polynomial based pose interpolation for readout time calibration of RS cameras. However, the IMU intrinsics were not considered in the above systems. The closest works to ours are by Li et al. [15] and Huai et al. [33] which included IMU-camera extrinsics, time offset, rolling-shutter readout time, camera and IMU intrinsics into the state vector of VINS. The former is built with MSCKF [8] based visual-inertial odometry while the latter uses a key-frame sliding-window filter based VINS. Both systems can calibrate all these parameters. However, no system observability was present in [15] and degenerate motion analysis was still missing from [33]. Instead, system observability and degenerate motion analysis are the focus of our work along with more extensive multi-run statistical validations of the calibration results. In addition, we also evaluate different IMU model variants which have appeared in literature.

C. Observability, Degeneracy and Noise

Observability analysis plays an important role in state estimation [34], especially when the system incorporates biases and calibration parameters [16], [17], [25], [35], [36]. Hernandez et al. [36] studied VINS observability with biases (not noise) as unknown input to examine the bounds for a set of the indistinguishable trajectories. The observability analysis in this paper is performed based on the corresponding deterministic, noise-free systems (e.g., [37]), in order to understand whether the states are estimable given measurements. We wish to understand whether these calibration parameters can be calibrated with visual-inertial measurements, and also identify degenerate motions, which might cause calibration to fail. In addition, observability properties can be leveraged for consistent estimator design [34], [37]. Kelly et al. [20] studied the IMU-camera self-calibration and performed nonlinear observability analysis using Lie derivative to show that the rigid transformation between IMU-camera is observable given random motions. Guo et al. [21] simplified the proof and analytically showed that the spatial calibration between the IMU and RGBD camera is observable. Li et al. [25] analyzed the identifiability for IMU-camera temporal calibration given the measurements of a monocular visual-inertial system and identified a set of degenerate motions that can cause the IMU-camera time offset to become unobservable. Tsao et al. [22] built the observability matrix for linearized VINS and showed that the camera intrinsics (only including focal length and principal points) is observable. However, none of the above mentioned works ever performed and verified the observability analysis with full-parameter calibration for VINS.

In our previous work [16], we built the observability matrix for VINS using the linearized system with IMU-camera spatial-temporal extrinsic calibration and and identified four degenerate motions that can cause these parameters to become unobservable. In our recent work [17], we performed observability analysis for monocular VINS with only IMU intrinsic calibration. Building upon these, we perform full-parameter calibration – IMU intrinsics (including g-sensitivity), camera intrinsics and the IMU-camera spatial-temporal calibration (including RS readout time) – for VINS with a single IMU and a monocular RS camera. Comprehensive observability analysis and degenerate motion identification are performed for these calibration parameters. Both simulations and real world experiments are also leveraged to verify our analysis.

It is difficult, if not impossible, to theoretically quantify the effects of measurement noise on system observability. Noise is often treated as random, uninformative input for VINS (and other robotic systems). In general, since VINS is a nonlinear estimation problem and often uses a linearized estimator (e.g., EKF or window BA), large noise can cause large linearization errors (but not observability), thus degrading its estimation performance, even overturning its convergence. Specifically, in VINS, the image and IMU noises affect its performance. In our recent work [38], the impact of image noise on the VINS accuracy was investigated, showing that the smaller the image noise is, the higher the estimation accuracy is. In this paper, we study the IMU noise impact on calibration, by using different quality IMUs of small and large noises. The calibration results confirm our observability analysis that the IMU calibration can converge to reference value if given fully excited motions. It is also evident from these results that the IMU noises affect the calibration; the lower the IMU noise is, the better calibration convergence can be achieved.

III. Sensing Models

A. IMU Intrinsic Model

We define an IMU as containing two separate frames of reference (see Fig. 1): gyroscope frame $\{w\}$, accelerometer frame $\{a\}$. The base "inertial" frame $\{I\}$ should be determined to coincide with either $\{w\}$ or $\{a\}$. Different from the model in [24], we define the raw angular velocity reading ${}^w\boldsymbol{\omega}_m$ from the gyroscope and linear acceleration readings ${}^a\mathbf{a}_m$ from the accelerometer as follows[1]:

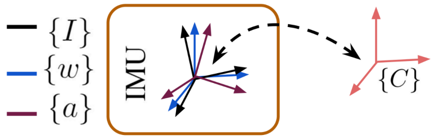


Fig. 1: An IMU sensor composed of accelerometer and gyroscope. The base "inertial" frame $\{I\}$ can be determined to coincide with either accelerometer frame $\{a\}$ or gyroscope frame $\{w\}$. $\{C\}$ represents the camera frame.

$$ {}^w\boldsymbol{\omega}_m = \mathbf{T}_w {}^w_I\mathbf{R}\, {}^I\boldsymbol{\omega} + \mathbf{T}_g {}^I\mathbf{a} + \mathbf{b}_g + \mathbf{n}_g \tag{1} $$

$$ {}^a\mathbf{a}_m = \mathbf{T}_a {}^a_I\mathbf{R}\, {}^I\mathbf{a} + \mathbf{b}_a + \mathbf{n}_a \tag{2} $$

where $\mathbf{T}_w$ and $\mathbf{T}_a$ are invertible 3×3 matrices which account for the scale imperfection and axis misalignment for $\{w\}$ and $\{a\}$, respectively. ${}^w_I\mathbf{R}$ and ${}^a_I\mathbf{R}$ denote the rotation from the gyroscope frame and acceleration frame to base "inertial" frame $\{I\}$, respectively. Note that, if we choose $\{I\}$ coincides with $\{w\}$, then ${}^w_I\mathbf{R} = \mathbf{I}_3$. Otherwise, ${}^a_I\mathbf{R} = \mathbf{I}_3$. $\mathbf{b}_g$ and $\mathbf{b}_a$ are the gyroscope and accelerometer biases, which are modeled as random walks. $\mathbf{n}_g$ and $\mathbf{n}_a$ are the zero-mean Gaussian noises contaminating the measurements. $\mathbf{T}_g$ denotes the g-sensitivity to account for the effects of acceleration to the gyroscope readings. Note that as in [15] and [24], we do not consider the translation between the gyroscope and accelerometer, as it is often negligible or safely excluded from the state vector by assuming $\{I\}$ coincides with $\{a\}$ frame (because any point in the IMU as a rigid body shares the same angular velocity). We can write the true (or corrected) angular velocity ${}^I\boldsymbol{\omega}$ and linear acceleration ${}^I\mathbf{a}$ as:

$$ {}^I\boldsymbol{\omega} = {}^I_w\mathbf{R}\mathbf{D}_w \left({}^w\boldsymbol{\omega}_m - \mathbf{T}_g {}^I\mathbf{a} - \mathbf{b}_g - \mathbf{n}_g \right) \tag{3} $$

$$ {}^I\mathbf{a} = {}^I_a\mathbf{R}\mathbf{D}_a \left({}^a\mathbf{a}_m - \mathbf{b}_a - \mathbf{n}_a \right) \tag{4} $$

where $\mathbf{D}_w = \mathbf{T}_w^{-1}$ and $\mathbf{D}_a = \mathbf{T}_a^{-1}$. In practice we calibrate $\mathbf{D}_a$, $\mathbf{D}_w$, ${}^I_a\mathbf{R}$ (or ${}^I_w\mathbf{R}$) and $\mathbf{T}_g$ for convenience. We only calibrate either ${}^I_w\mathbf{R}$ or ${}^I_a\mathbf{R}$ in Eq. (3)-(4) since the base "inertial" frame coincides with one of these sensor frames. If both ${}^I_w\mathbf{R}$ and ${}^I_a\mathbf{R}$ were calibrated, it would make the rotation between the IMU and camera unobservable due to over parameterization (validated in Section VIII-D).

1) Intrinsic model variants: Given the general model [see Eq. (3) and (4)], different parameters can be chosen to be estimated and, this results in different IMU intrinsic models (see [15], [24], [30]–[32], [41]). In the following, we will present and evaluate these variants.

- `imu1`: includes the rotation ${}^I_w\mathbf{R}$, 6 parameters for $\mathbf{D}_w$ (and thus denoted by $\mathbf{D}_{w6}$) and 6 parameters for $\mathbf{D}_a$ (denoted by $\mathbf{D}_{a6}$), as they assume the upper-triangular structure:

$$ \mathbf{D}_{*6} = \begin{bmatrix} d_{*1} & d_{*2} & d_{*4} \\ 0 & d_{*3} & d_{*5} \\ 0 & 0 & d_{*6} \end{bmatrix} \tag{5} $$

- `imu2`: includes the rotation ${}^I_a\mathbf{R}$ instead, $\mathbf{D}_{a6}$ and $\mathbf{D}_{w6}$, which is the model used by [24].

[1] Note that the IMU measurement model is based on the "flat earth" assumption, instead of the "rotating earth" model [8], [39], [40], where the Coriolis effect is considered.

TABLE I: IMU model variants and parameters.

Model	Dim.	$\mathbf{D}_w$	$\mathbf{D}_a$	${}^I_w\mathbf{R}$	${}^I_a\mathbf{R}$	$\mathbf{T}_g$
imu0	0	-	-	-	-	-
imu1	15	$\mathbf{D}_{w6}$	$\mathbf{D}_{a6}$	${}^I_w\mathbf{R}$	-	-
imu2	15	$\mathbf{D}_{w6}$	$\mathbf{D}_{a6}$	-	${}^I_a\mathbf{R}$	-
imu3	15	$\mathbf{D}_{w9}$	$\mathbf{D}_{a6}$	-	-	-
imu4	15	$\mathbf{D}_{w6}$	$\mathbf{D}_{a9}$	-	-	-
imu5	18	$\mathbf{D}_{w6}$	$\mathbf{D}_{a6}$	${}^I_w\mathbf{R}$	${}^I_a\mathbf{R}$	-
imu6	24	$\mathbf{D}'_{w6}$	$\mathbf{D}'_{a6}$	${}^I_w\mathbf{R}$	-	$\mathbf{T}_{g9}$
imu11	21	$\mathbf{D}_{w6}$	$\mathbf{D}_{a6}$	${}^I_w\mathbf{R}$	-	$\mathbf{T}_{g6}$
imu12	21	$\mathbf{D}_{w6}$	$\mathbf{D}_{a6}$	-	${}^I_a\mathbf{R}$	$\mathbf{T}_{g6}$
imu13	21	$\mathbf{D}_{w9}$	$\mathbf{D}_{a6}$	-	-	$\mathbf{T}_{g6}$
imu14	21	$\mathbf{D}_{w6}$	$\mathbf{D}_{a9}$	-	-	$\mathbf{T}_{g6}$
imu21	24	$\mathbf{D}_{w6}$	$\mathbf{D}_{a6}$	${}^I_w\mathbf{R}$	-	$\mathbf{T}_{g9}$
imu22	24	$\mathbf{D}_{w6}$	$\mathbf{D}_{a6}$	-	${}^I_a\mathbf{R}$	$\mathbf{T}_{g9}$
imu23	24	$\mathbf{D}_{w9}$	$\mathbf{D}_{a6}$	-	-	$\mathbf{T}_{g9}$
imu24	24	$\mathbf{D}_{w6}$	$\mathbf{D}_{a9}$	-	-	$\mathbf{T}_{g9}$
imu31	9	-	$\mathbf{D}_{a9}$	-	-	-
imu32	9	$\mathbf{D}_{w9}$	-	-	-	-
imu33	6	-	-	-	-	$\mathbf{T}_{g6}$
imu34	9	-	-	-	-	$\mathbf{T}_{g9}$

- imu3: combines imu1's $\mathbf{D}_{w6}$ and ${}^I_w\mathbf{R}$ into a general 3×3 matrix containing 9 parameters in total. Thus, we estimate the upper-triangle $\mathbf{D}_{a6}$ and a full matrix $\mathbf{D}_{w9}$ as:

$$\mathbf{D}_{*9} = \begin{bmatrix} d_{*1} & d_{*4} & d_{*7} \\ d_{*2} & d_{*5} & d_{*8} \\ d_{*3} & d_{*6} & d_{*9} \end{bmatrix} \tag{6}$$

- imu4: is an extension of imu2 with a combination of the $\mathbf{D}_{a6}$ and ${}^I_a\mathbf{R}$. Similarly, in this variant we estimate the upper-triangle $\mathbf{D}_{w6}$ and a full matrix $\mathbf{D}_{a9}$.
- imu1A ($A = 1, \cdots, 4$): combines imuA with a 6-parameter g-sensitivity $\mathbf{T}_{g6}$ as:

$$\mathbf{T}_{g6} = \begin{bmatrix} t_{g1} & t_{g2} & t_{g4} \\ 0 & t_{g3} & t_{g5} \\ 0 & 0 & t_{g6} \end{bmatrix} \tag{7}$$

- imu2A ($A = 1, \cdots, 4$): combines imuA with a the 9-parameter g-sensitivity $\mathbf{T}_{g9}$ as:

$$\mathbf{T}_{g9} = \begin{bmatrix} t_{g1} & t_{g4} & t_{g7} \\ t_{g2} & t_{g5} & t_{g8} \\ t_{g3} & t_{g6} & t_{g9} \end{bmatrix} \tag{8}$$

- imu5: contains $\mathbf{D}_{w6}$, $\mathbf{D}_{a6}$, ${}^I_w\mathbf{R}$ and ${}^I_a\mathbf{R}$. This is a redundant over-parameterized model which will be used to verify that ${}^I_w\mathbf{R}$ and ${}^I_a\mathbf{R}$ should not be calibrated simultaneously.
- imu6: contains $\mathbf{D}'_{w6}$, $\mathbf{D}'_{a6}$, ${}^I_w\mathbf{R}$ and $\mathbf{T}_{g9}$. This is equivalent to the *scale-misalignment* IMU model [32] used in Kalibr [27]. $\mathbf{D}'_{*6}$ assumes the lower triangular structure:

$$\mathbf{D}'_{*6} = \begin{bmatrix} d_{*1} & 0 & 0 \\ d_{*2} & d_{*4} & 0 \\ d_{*3} & d_{*5} & d_{*6} \end{bmatrix} \tag{9}$$

- imu3A ($A = 1, \cdots, 4$): models a subset of the parameters of the general model while assuming the others known; that is, only calibrates $\mathbf{D}_{a9}$ in imu31, $\mathbf{D}_{w9}$ in imu32, $\mathbf{T}_{g6}$ in imu33, and $\mathbf{T}_{g9}$ in imu34.

These different models are summarized in Table I. For presentation clarity, imu22$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a6}$, ${}^I_a\mathbf{R}$, $\mathbf{T}_{g9}\}$ is used in the ensuing system derivations and analysis.

B. Camera Model

If a 3D point feature is captured by a camera, the visual measurement function is:

$$\begin{aligned} \mathbf{z}_C &= \begin{bmatrix} u & v \end{bmatrix}^\top + \mathbf{n}_C & (10) \\ &= \mathbf{h}_d\left(\mathbf{z}_n, \mathbf{x}_{Cin}\right) + \mathbf{n}_C & (11) \end{aligned}$$

where $\mathbf{n}_C$ denotes the measurement noise; u and v are the distorted image pixel coordinates, $\mathbf{z}_n = [u_n \ v_n]^\top$ represents the normalized image pixel and $\mathbf{h}_d(\cdot)$ maps the normalized image pixel onto the image plane based on the camera intrinsic parameters $\mathbf{x}_{Cin}$ and camera model. While a pinhole model with radial-tangential (*radtan*) or equivalent-distant (*equidist*) distortion can be used, the *radtan* model is used in the ensuing derivations (see [27]). Specifically, the *radtan* $\mathbf{x}_{Cin}$ and $\mathbf{h}_d(\cdot)$ are given by:

$$\mathbf{x}_{Cin} = \begin{bmatrix} f_u & f_v & c_u & c_v & k_1 & k_2 & p_1 & p_2 \end{bmatrix}^\top \tag{12}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f_u & 0 \\ 0 & f_v \end{bmatrix} \begin{bmatrix} u_d \\ v_d \end{bmatrix} + \begin{bmatrix} c_u \\ c_v \end{bmatrix} \tag{13}$$

$$\begin{bmatrix} u_d \\ v_d \end{bmatrix} = \begin{bmatrix} du_n + 2p_1u_nv_n + p_2(r^2 + 2u_n^2) \\ dv_n + p_1(r^2 + 2v_n^2) + 2p_2u_nv_n \end{bmatrix} \tag{14}$$

where $r^2 = u_n^2 + v_n^2$; $d = 1 + k_1r^2 + k_2r^4$; f_u and f_v are the camera focal length; $\{c_u, c_v\}$ denotes the image principal point; k_1 and k_2 represent the radial distortion coefficients while p_1 and p_2 are tangential distortion coefficients. Normalized image pixel u_n and v_n is obtained by projecting 3D feature ${}^C\mathbf{p}_f = [{}^Cx_f \ {}^Cy_f \ {}^Cz_f]^\top$ into 2D image plane:

$$\mathbf{z}_n = \mathbf{h}_p({}^C\mathbf{p}_f) \triangleq \frac{1}{{}^Cz_f} \begin{bmatrix} {}^Cx_f \\ {}^Cy_f \end{bmatrix} \tag{15}$$

$$\begin{aligned} {}^C\mathbf{p}_f &= \mathbf{h}_t({}^I_G\mathbf{R}, {}^G\mathbf{p}_I, {}^C_I\mathbf{R}, {}^C\mathbf{p}_I, {}^G\mathbf{p}_f) & (16) \\ &\triangleq {}^C_I\mathbf{R}{}^I_G\mathbf{R}({}^G\mathbf{p}_f - {}^G\mathbf{p}_I) + {}^C\mathbf{p}_I \end{aligned}$$

where $\{{}^C_I\mathbf{R}, {}^C\mathbf{p}_I\}$ represents the rigid transformation between the IMU and camera frames.

In addition, global shutter (GS) and rolling shutter (RS) are two common variants of camera sensing modes. GS cameras expose all pixels at a single time instance, while, typically lower-cost, RS cameras expose each row sequentially. As shown by [7], it may lead to large estimation errors if this RS effect is not taken into account when using RS cameras for VINS. Additionally, the camera and IMU measurement timestamps can be incorrect due to processing or communication delays, or different clock references. To address these, we model both the time offset and camera readout time to ensure all measurements are processed in a common clock frame of reference and at the correct corresponding poses. Specifically, t_d denotes the time offset between IMU and camera timeline, and t_r denotes the RS readout time for the whole image. If t denotes the time when the pixel is captured, the measurement function for pixels captured in the m-th row (out of total M rows) is:

$$\begin{aligned} {}^C\mathbf{p}_f &= \mathbf{h}_t({}^{I(t)}_G\mathbf{R}, {}^G\mathbf{p}_{I(t)}, {}^C_I\mathbf{R}, {}^C\mathbf{p}_I, {}^G\mathbf{p}_f) & (17) \\ &\triangleq {}^C_I\mathbf{R}{}^{I(t)}_G\mathbf{R}\left({}^G\mathbf{p}_f - {}^G\mathbf{p}_{I(t)}\right) + {}^C\mathbf{p}_I \end{aligned}$$

$$t_I = t_C + t_d \tag{18}$$

$$t = t_I + \frac{m}{M}t_r \tag{19}$$

where t_I is the IMU state time corresponding to the captured image time t_C when the first row of the image is collected. If the readout time $t_r = 0$, then the camera is actually a

GS camera and all rows are a function of the same pose. $\{{}^{G}_{I(t)}\mathbf{R}, {}^{G}\mathbf{p}_{I(t)}\}$ is the IMU global pose corresponding to the camera measurement time t.

IV. VINS Models with Self-Calibration

A. State Vector

The state vector $\mathbf{x}$ of the proposed visual-inertial system includes the inertial navigation state $\mathbf{x}_I$, IMU intrinsic parameter $\mathbf{x}_{in}$, IMU-camera spatial-temporal extrinsic calibration $\mathbf{x}_{IC}$, camera intrinsic calibration $\mathbf{x}_{Cin}$ and feature positions $\mathbf{x}_f$.

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_I^\top & \mathbf{x}_{IC}^\top & \mathbf{x}_{Cin}^\top & \mathbf{x}_f^\top \end{bmatrix}^\top \tag{20}$$

$$\begin{aligned} \mathbf{x}_I &= \begin{bmatrix} \mathbf{x}_n^\top & | & \mathbf{x}_b^\top & | & \mathbf{x}_{in}^\top \end{bmatrix}^\top \qquad (21) \\ &= \begin{bmatrix} {}^{I}_{G}\bar{q}^\top & {}^{G}\mathbf{p}_I^\top & {}^{G}\mathbf{v}_I^\top & | & \mathbf{b}_g^\top & \mathbf{b}_a^\top & | & \mathbf{x}_{in}^\top \end{bmatrix}^\top \end{aligned}$$

$$\mathbf{x}_{in} = \begin{bmatrix} \mathbf{x}_{Dw}^\top & \mathbf{x}_{Da}^\top & {}^{I}_{a}\bar{q}^\top & \mathbf{x}_{Tg}^\top \end{bmatrix}^\top \tag{22}$$

$$\mathbf{x}_{IC} = \begin{bmatrix} {}^{C}_{I}\bar{q}^\top & {}^{C}\mathbf{p}_I^\top & t_d & t_r \end{bmatrix}^\top \tag{23}$$

where ${}^{I}_{G}\bar{q}$ denotes quaternion with JPL convention [42] and corresponds to the rotation matrix ${}^{I}_{G}\mathbf{R}$, which represents the rotation from $\{G\}$ to $\{I\}$. ${}^{G}\mathbf{p}_I$ and ${}^{G}\mathbf{v}_I$ denote the IMU position and velocity in $\{G\}$. $\mathbf{x}_n$ denotes the IMU navigation states containing the ${}^{I}_{G}\bar{q}$, ${}^{G}\mathbf{p}_I$ and ${}^{G}\mathbf{v}_I$. $\mathbf{x}_b$ denotes the IMU bias states containing $\mathbf{b}_g$ and $\mathbf{b}_a$. $\{{}^{C}_{I}\bar{q}, {}^{C}\mathbf{p}_I\}$ denotes the rigid transformation between $\{C\}$ and $\{I\}$. t_d and t_r represent the IMU-camera time offset and camera readout time. IMU intrinsics, $\mathbf{x}_{in}$, contains $\mathbf{x}_{Dw}$, $\mathbf{x}_{Da}$, $\mathbf{x}_{Tg}$ and ${}^{I}_{a}\bar{q}$, where $\mathbf{x}_{Dw}$, $\mathbf{x}_{Da}$ and $\mathbf{x}_{Tg}$ are non-zero elements stored column-wise in $\mathbf{D}_w$, $\mathbf{D}_a$ and $\mathbf{T}_g$. We have:

$$\mathbf{x}_{D*} = \begin{bmatrix} d_{*1} & d_{*2} & d_{*3} & d_{*4} & d_{*5} & d_{*6} \end{bmatrix}^\top \tag{24}$$

$$\mathbf{x}_{Tg} = \begin{bmatrix} t_{g1} & t_{g2} & t_{g3} & t_{g4} & t_{g5} & t_{g6} & t_{g7} & t_{g8} & t_{g9} \end{bmatrix}^\top \tag{25}$$

for $\mathtt{imu22}\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}, \mathbf{T}_{g9}\}$. In this paper, $\hat{\mathbf{x}}$ denotes the estimated value for state $\mathbf{x}$ and $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$ is the error state. We use the quaternion left multiplicative error defined by $\bar{q} \approx [\frac{1}{2}\delta\boldsymbol{\theta}^\top \; 1]^\top \otimes \hat{\bar{q}}$, where $\otimes$ denotes quaternion multiplication [42]. This rotation error state is equivalent to the $SO(3)$ error ${}^{I}_{G}\mathbf{R} \approx (\mathbf{I}_3 - \lfloor \delta\boldsymbol{\theta} \rfloor){}^{I}_{G}\hat{\mathbf{R}}$. Note that $\lfloor \mathbf{v} \rfloor$ denotes the skew-symmetric matrix [42] of the vector $\mathbf{v}$.

B. Analytic Inertial Integration

The dynamics of inertial navigation state $\mathbf{x}_I$ is (see [39]):

$$\begin{aligned} {}^{I}_{G}\dot{\bar{q}} &= \frac{1}{2}\boldsymbol{\Omega}({}^{I}\boldsymbol{\omega}){}^{I}_{G}\bar{q} \;, \quad {}^{G}\dot{\mathbf{p}}_I = {}^{G}\mathbf{v}_I \qquad (26) \\ {}^{G}\dot{\mathbf{v}}_I &= {}^{I}_{G}\mathbf{R}^\top {}^{I}\mathbf{a} - {}^{G}\mathbf{g} \;, \quad \dot{\mathbf{b}}_g = \mathbf{n}_{wg} \;, \quad \dot{\mathbf{b}}_a = \mathbf{n}_{wa} \end{aligned}$$

where $\boldsymbol{\Omega}(\boldsymbol{\omega}) = \begin{bmatrix} -\lfloor \boldsymbol{\omega} \rfloor & \boldsymbol{\omega} \\ -\boldsymbol{\omega}^\top & 0 \end{bmatrix}$, $\mathbf{n}_{wg}$ and $\mathbf{n}_{wa}$ are zero-mean white Gaussian noises driving $\mathbf{b}_g$ and $\mathbf{b}_a$, respectively, and the known global gravity assumes ${}^{G}\mathbf{g} = [0 \; 0 \; 9.81]^\top$, while the rest of the states have zero dynamics. The integration of IMU dynamics (26) from time step t_k to t_{k+1} is computed [17]:

$${}^{I_{k+1}}_{G}\mathbf{R} = \Delta\mathbf{R}_k^\top {}^{I_k}_{G}\mathbf{R} \tag{27}$$

$${}^{G}\mathbf{p}_{I_{k+1}} = {}^{G}\mathbf{p}_{I_k} + {}^{G}\mathbf{v}_{I_k}\delta t_k + {}^{I_k}_{G}\mathbf{R}^\top \Delta\mathbf{p}_k - \frac{1}{2}{}^{G}\mathbf{g}\delta t_k^2 \tag{28}$$

$${}^{G}\mathbf{v}_{I_{k+1}} = {}^{G}\mathbf{v}_{I_k} + {}^{I_k}_{G}\mathbf{R}^\top \Delta\mathbf{v}_k - {}^{G}\mathbf{g}\delta t_k \tag{29}$$

$$\mathbf{b}_{g_{k+1}} = \mathbf{b}_{g_k} + \int_{t_k}^{t_{k+1}} \mathbf{n}_{wg} d\tau \tag{30}$$

$$\mathbf{b}_{a_{k+1}} = \mathbf{b}_{a_k} + \int_{t_k}^{t_{k+1}} \mathbf{n}_{wa} d\tau \tag{31}$$

where $\delta t_k = t_{k+1} - t_k$, and the integration quantities are:

$$\Delta\mathbf{R}_k \triangleq {}^{I_k}_{I_{k+1}}\mathbf{R} = \exp\left(\int_{t_k}^{t_{k+1}} {}^{I_\tau}\boldsymbol{\omega} d\tau\right) \tag{32}$$

$$\Delta\mathbf{p}_k \triangleq \int_{t_k}^{t_{k+1}} \int_{t_k}^{s} {}^{I_k}_{I_\tau}\mathbf{R} {}^{I_\tau}\mathbf{a} d\tau ds \tag{33}$$

$$\Delta\mathbf{v}_k \triangleq \int_{t_k}^{t_{k+1}} {}^{I_k}_{I_\tau}\mathbf{R} {}^{I_\tau}\mathbf{a} d\tau \tag{34}$$

where $\exp(\cdot)$ is the $SO(3)$ matrix exponential [12]. Assuming constant ${}^{I_k}\hat{\boldsymbol{\omega}}$ and ${}^{I_k}\hat{\mathbf{a}}$ within the time interval, we approximate $\Delta\hat{\mathbf{R}}_k$, $\Delta\hat{\mathbf{p}}_k$ and $\Delta\hat{\mathbf{v}}_k$ as:

$$\Delta\hat{\mathbf{R}}_k \simeq \exp\left({}^{I_k}\hat{\boldsymbol{\omega}}\delta t_k\right) \tag{35}$$

$$\Delta\hat{\mathbf{p}}_k \simeq \left(\int_{t_k}^{t_{k+1}} \int_{t_k}^{s} {}^{I_k}_{I_\tau}\hat{\mathbf{R}} d\tau ds\right) \cdot {}^{I_k}\hat{\mathbf{a}} \triangleq \boldsymbol{\Xi}_2 \cdot {}^{I_k}\hat{\mathbf{a}} \tag{36}$$

$$\Delta\hat{\mathbf{v}}_k \simeq \left(\int_{t_k}^{t_{k+1}} {}^{I_k}_{I_\tau}\hat{\mathbf{R}} d\tau\right) \cdot {}^{I_k}\hat{\mathbf{a}} \triangleq \boldsymbol{\Xi}_1 \cdot {}^{I_k}\hat{\mathbf{a}} \tag{37}$$

The terms $\boldsymbol{\Xi}_1$ and $\boldsymbol{\Xi}_2$ are defined as integration components which can be evaluated either analytically [43] or numerically using the Runge–Kutta fourth-order (RK4) method. ${}^{I_k}\hat{\boldsymbol{\omega}}$ and ${}^{I_k}\hat{\mathbf{a}}$ are computed as (note that we drop the timestamp k for simplicity):

$${}^{I}\hat{\boldsymbol{\omega}} = {}^{I}_{w}\hat{\mathbf{R}}\hat{\mathbf{D}}_w {}^{w}\hat{\boldsymbol{\omega}}, \; {}^{I}\hat{\mathbf{a}} = {}^{I}_{a}\hat{\mathbf{R}}\hat{\mathbf{D}}_a {}^{a}\hat{\mathbf{a}} \tag{38}$$

$${}^{w}\hat{\boldsymbol{\omega}} = {}^{w}\boldsymbol{\omega}_m - \hat{\mathbf{T}}_g {}^{I}\hat{\mathbf{a}} - \hat{\mathbf{b}}_g \triangleq \begin{bmatrix} {}^{w}\hat{w}_1 & {}^{w}\hat{w}_2 & {}^{w}\hat{w}_3 \end{bmatrix}^\top \tag{39}$$

$${}^{a}\hat{\mathbf{a}} = {}^{a}\mathbf{a}_m - \hat{\mathbf{b}}_a \triangleq \begin{bmatrix} {}^{a}\hat{a}_1 & {}^{a}\hat{a}_2 & {}^{a}\hat{a}_3 \end{bmatrix}^\top \tag{40}$$

where ${}^{I}_{w}\hat{\mathbf{R}} = \mathbf{I}_3$ for $\mathtt{imu22}\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}, \mathbf{T}_{g9}\}$.

C. Linearized System Model

The IMU integration components [see Eq. (32)-(34)] can be linearized as:

$$\Delta\mathbf{R}_k = \Delta\hat{\mathbf{R}}_k \Delta\tilde{\mathbf{R}}_k \triangleq \Delta\hat{\mathbf{R}}_k \exp\left(\mathbf{J}_r(\Delta\hat{\boldsymbol{\theta}}_k){}^{I_k}\tilde{\boldsymbol{\omega}}\delta t_k\right) \tag{41}$$

$$\Delta\mathbf{p}_k = \Delta\hat{\mathbf{p}}_k + \Delta\tilde{\mathbf{p}}_k \triangleq \Delta\hat{\mathbf{p}}_k - \boldsymbol{\Xi}_4 {}^{I_k}\tilde{\boldsymbol{\omega}} + \boldsymbol{\Xi}_2 {}^{I_k}\tilde{\mathbf{a}} \tag{42}$$

$$\Delta\mathbf{v}_k = \Delta\hat{\mathbf{v}}_k + \Delta\tilde{\mathbf{v}}_k \triangleq \Delta\hat{\mathbf{v}}_k - \boldsymbol{\Xi}_3 {}^{I_k}\tilde{\boldsymbol{\omega}} + \boldsymbol{\Xi}_1 {}^{I_k}\tilde{\mathbf{a}} \tag{43}$$

where $\mathbf{J}_r(\Delta\hat{\boldsymbol{\theta}}_k) \triangleq \mathbf{J}_r\left({}^{I_k}\hat{\boldsymbol{\omega}}\delta t_k\right)$ denotes the right Jacobian of $SO(3)$ [12]. The derivation and the definitions of ${}^{I_k}\tilde{\boldsymbol{\omega}}$ and ${}^{I_k}\tilde{\mathbf{a}}$ can be found in Appendix A. The integrated components $\boldsymbol{\Xi}_3$ and $\boldsymbol{\Xi}_4$ are defined as:

$$\boldsymbol{\Xi}_3 \triangleq \int_{t_k}^{t_{k+1}} {}^{I_k}_{I_\tau}\mathbf{R} \lfloor {}^{I_\tau}\mathbf{a} \rfloor \mathbf{J}_r\left({}^{I_k}\boldsymbol{\omega}\delta\tau\right) \delta\tau d\tau \tag{44}$$

$$\boldsymbol{\Xi}_4 \triangleq \int_{t_k}^{t_{k+1}} \int_{t_k}^{s} {}^{I_k}_{I_\tau}\mathbf{R} \lfloor {}^{I_\tau}\mathbf{a} \rfloor \mathbf{J}_r\left({}^{I_k}\boldsymbol{\omega}\delta\tau\right) \delta\tau d\tau ds \tag{45}$$

where $\delta\tau = t_\tau - t_k$. As such, the linearized error-state system for *imu22* is:

$$\tilde{\mathbf{x}}_{I_{k+1}} \simeq \mathbf{\Phi}_{I(k+1,k)}\tilde{\mathbf{x}}_{I_k} + \mathbf{G}_{Ik}\mathbf{n}_{dk} \tag{46}$$

$$\mathbf{\Phi}_{I(k+1,k)} = \begin{bmatrix} \mathbf{\Phi}_{nn} & \mathbf{\Phi}_{wa}\mathbf{H}_b & \mathbf{\Phi}_{wa}\mathbf{H}_{in} \\ \mathbf{0}_{6\times 9} & \mathbf{I}_6 & \mathbf{0}_{6\times 24} \\ \mathbf{0}_{24\times 9} & \mathbf{0}_{24\times 6} & \mathbf{I}_{24} \end{bmatrix} \tag{47}$$

$$\mathbf{G}_{Ik} = \begin{bmatrix} \mathbf{\Phi}_{wa}\mathbf{H}_n & \mathbf{0}_{9\times 6} \\ \mathbf{0}_6 & \mathbf{I}_6\delta t_k \\ \mathbf{0}_{24\times 6} & \mathbf{0}_{24\times 6} \end{bmatrix} \tag{48}$$

where $\mathbf{\Phi}_{I(k+1,k)}$ and $\mathbf{G}_{Ik}$ are the state transition matrix and noise Jacobians for the inertial state $\mathbf{x}_I$ dynamics; $\mathbf{H}_b$, $\mathbf{H}_{in}$ and $\mathbf{H}_n$ are Jacobians related to bias, IMU intrinsics and noises, which can be found in Appendix A. $\mathbf{n}_{dk}$ is the discrete-time IMU noises, while $\mathbf{\Phi}_{nn}$ and $\mathbf{\Phi}_{wa}$ can be computed as:

$$\mathbf{\Phi}_{nn} = \begin{bmatrix} \Delta\hat{\mathbf{R}}_k^\top & \mathbf{0}_3 & \mathbf{0}_3 \\ -{}^{I_k}_{G}\hat{\mathbf{R}}^\top\lfloor\Delta\hat{\mathbf{p}}_k\rfloor & \mathbf{I}_3 & \mathbf{I}_3\delta t_k \\ -{}^{I_k}_{G}\hat{\mathbf{R}}^\top\lfloor\Delta\hat{\mathbf{v}}_k\rfloor & \mathbf{0}_3 & \mathbf{I}_3 \end{bmatrix}, \ \mathbf{\Phi}_{wa} = \begin{bmatrix} \mathbf{J}_r(\delta\boldsymbol{\theta}_k)\delta t_k & \mathbf{0}_3 \\ -{}^{I_k}_{G}\hat{\mathbf{R}}^\top\mathbf{\Xi}_4 & {}^{I_k}_{G}\hat{\mathbf{R}}^\top\mathbf{\Xi}_2 \\ -{}^{I_k}_{G}\hat{\mathbf{R}}^\top\mathbf{\Xi}_3 & {}^{I_k}_{G}\hat{\mathbf{R}}^\top\mathbf{\Xi}_1 \end{bmatrix}$$

Without loss of generality, we consider a single 3D feature ${}^G\mathbf{p}_f$ in the state vector $\mathbf{x}_f$. Since there is zero dynamics for $\mathbf{x}_{IC}$, $\mathbf{x}_{Cin}$ and $\mathbf{x}_f$, we can write the state transition matrix for the whole state vector $\mathbf{x}$ as a block-diagonal matrix [see Eq. (20)] as:

$$\mathbf{\Phi}_{k+1,k} = \mathbf{diag}\{\mathbf{\Phi}_{I(k+1,k)},\ \mathbf{\Phi}_{IC},\ \mathbf{\Phi}_{Cin},\ \mathbf{\Phi}_f\} \tag{49}$$

where $\mathbf{\Phi}_{IC} = \mathbf{I}_8$, $\mathbf{\Phi}_{Cin} = \mathbf{I}_8$, and $\mathbf{\Phi}_f = \mathbf{I}_{3n}$.

D. Linearized Measurement Model

The comprehensive camera measurement model $\mathbf{h}_C(\cdot)$ is composed of the distortion function $\mathbf{h}_d(\cdot)$ [see Eq. (11)], the projection function $\mathbf{h}_p(\cdot)$ [see Eq. (15)] and the transformation function $\mathbf{h}_t(\cdot)$ [see Eq. (17)]:

$$\begin{aligned} \mathbf{z}_C &= \mathbf{h}_C(\mathbf{x}) + \mathbf{n}_C && (50)\\ &= \mathbf{h}_d(\mathbf{z}_n, \mathbf{x}_{Cin}) + \mathbf{n}_C && (51)\\ &= \mathbf{h}_d(\mathbf{h}_p({}^{C_k}\mathbf{p}_f), \mathbf{x}_{Cin}) + \mathbf{n}_C && (52)\\ &= \mathbf{h}_d(\mathbf{h}_p(\mathbf{h}_t({}^{C(t)}_{G}\mathbf{R}, {}^G\mathbf{p}_{C(t)}, {}^G\mathbf{p}_f)), \mathbf{x}_{Cin}) + \mathbf{n}_C && (53) \end{aligned}$$

To perform observability analysis and build linearized state estimators, we need to linearize this complicated visual measurement model, which is given by:

$$\tilde{\mathbf{z}}_C \simeq \mathbf{H}_C\tilde{\mathbf{x}} + \mathbf{n}_C \tag{54}$$

where $\tilde{\mathbf{z}}_C \triangleq \mathbf{z}_C - \mathbf{h}_C(\hat{\mathbf{x}})$ and $\mathbf{H}_C \triangleq \frac{\partial\tilde{\mathbf{z}}_C}{\partial\tilde{\mathbf{x}}}$. We get the following Jacobian matrix with the chain rule of differentiation:

$$\begin{aligned} \mathbf{H}_C &= \begin{bmatrix} \frac{\partial\tilde{\mathbf{z}}_C}{\partial\tilde{\mathbf{x}}_I} & \frac{\partial\tilde{\mathbf{z}}_C}{\partial\tilde{\mathbf{x}}_{IC}} & \frac{\partial\tilde{\mathbf{z}}_C}{\partial\tilde{\mathbf{x}}_{Cin}} & \frac{\partial\tilde{\mathbf{z}}_C}{\partial\tilde{\mathbf{x}}_f} \end{bmatrix} && (55)\\ &= \begin{bmatrix} \mathbf{H}_{\mathbf{p}_f}\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_I} & \mathbf{H}_{\mathbf{p}_f}\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_{IC}} & \frac{\partial\tilde{\mathbf{z}}_C}{\partial\tilde{\mathbf{x}}_{Cin}} & \mathbf{H}_{\mathbf{p}_f}\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_f} \end{bmatrix} \end{aligned}$$

where $\mathbf{H}_{\mathbf{p}_f} = \frac{\partial\tilde{\mathbf{z}}_C}{\partial\tilde{\mathbf{z}}_n}\frac{\partial\tilde{\mathbf{z}}_n}{\partial{}^C\tilde{\mathbf{p}}_f}$. All the pertinent matrices $\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_I}$, $\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_{IC}}$, $\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_f}$ and $\mathbf{H}_{\mathbf{p}_f}$ are computed in Appendix B.

V. Observability Analysis

Observability analysis plays an important role in determining whether or not the states are estimable given measurements. Observability analysis can also be leveraged to identify degenerate motions that might negatively affect estimation performance [34]. While the observability analysis of VINS has been well studied [37], the observability properties and degenerate motions of VINS with full self-calibration – in particular, IMU and camera intrinsic calibration – have not been sufficiently investigated.

To this end, following [37], we construct the observability matrix for the deterministic (noise-free) linearized VINS models as follows:

$$\mathcal{O} = \begin{bmatrix} \mathcal{O}_1 \\ \mathcal{O}_2 \\ \vdots \\ \mathcal{O}_k \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{C1}\mathbf{\Phi}_{1,1} \\ \mathbf{H}_{C2}\mathbf{\Phi}_{2,1} \\ \vdots \\ \mathbf{H}_{Ck}\mathbf{\Phi}_{k,1} \end{bmatrix} \tag{56}$$

The k-th row of $\mathcal{O}$ is written as:

$$\mathcal{O}_k = \begin{bmatrix} \mathbf{M}_n & \mathbf{M}_b & \mathbf{M}_{in} & \mathbf{M}_{IC} & \mathbf{M}_{Cin} & \mathbf{M}_f \end{bmatrix} \tag{57}$$

where $\mathbf{M}_n$, $\mathbf{M}_b$, $\mathbf{M}_{in}$, $\mathbf{M}_{IC}$, $\mathbf{M}_{Cin}$ and $\mathbf{M}_f$ represent the matrix block relating to the IMU navigation, biases, IMU intrinsics, IMU-camera extrinsics, camera intrinsics and feature states [see Eq. (20)], with detailed derivations presented in Appendix C. We now look to find the unobservable subspace $\mathbf{N}$ such that $\mathcal{O}\mathbf{N} = \mathbf{0}$. The following results can be proved:

Lemma 1. *Given fully excited motions, monocular VINS system with online calibration of IMU intrinsics $\mathbf{x}_{in}$, camera intrinsics $\mathbf{x}_{Cin}$ and IMU-camera spatial-temporal parameters $\mathbf{x}_{IC}$ (including RS readout time) has 4 unobservable directions, which relate to the global yaw and global translation.*

$$\mathbf{N} = \begin{bmatrix} {}^{I_1}_{G}\hat{\mathbf{R}}{}^G\mathbf{g} & \mathbf{0}_3 \\ -\lfloor{}^G\hat{\mathbf{p}}_{I_1}\rfloor{}^G\mathbf{g} & \mathbf{I}_3 \\ -\lfloor{}^G\hat{\mathbf{v}}_{I_1}\rfloor{}^G\mathbf{g} & \mathbf{0}_3 \\ \mathbf{0}_{46\times 1} & \mathbf{0}_{46\times 3} \\ -\lfloor{}^G\hat{\mathbf{p}}_f\rfloor{}^G\mathbf{g} & \mathbf{I}_3 \end{bmatrix} \tag{58}$$

Proof. See Appendix D. □

It is clear from Appendix C that the terms $\mathbf{M}_{in}$ [see Eq. (103)] and $\mathbf{M}_{IC}$ [see Eq. (104)] of the observability matrix – corresponding to IMU intrinsics $\mathbf{x}_{in}$ and IMU-camera spatial-temporal parameters $\mathbf{x}_{IC}$ (including RS effects) – contain ${}^w\hat{\boldsymbol{\omega}}$, ${}^a\hat{\mathbf{a}}$, ${}^I\hat{\boldsymbol{\omega}}$ and ${}^G\hat{\mathbf{v}}_I$, which represent the sensor platform motion. This implies that $\mathbf{M}_{in}$ and $\mathbf{M}_{IC}$ are motion-dependent and time-varying. From the numerical simulations of VINS with a monocular RS camera and an IMU shown in Fig. 3, we can confirm that all these calibration parameters are observable and can be estimated given fully-excited motions. Note that other IMU intrinsic model variants besides *imu22* also observable given fully-excited motions, while their derivations and simulation results are omitted for brevity.

Similarly, the camera intrinsics, $\mathbf{M}_{Cin}$, are mainly affected by the environmental structure (the u and v measurements of the 3D point features). The camera intrinsic parameters are observable for most motion cases, even for under-actuated motions (e.g., planar motion), which is validated by our simulation results shown in Fig. 3-7. Note that these results also hold for the *equidist* camera distortion model, which however are again omitted here but can be found in our companion technical report [44].

VI. Degenerate Motion Analysis

While the observability properties found in the preceding section hold with *fully-excited* motions, this may not always be the case in reality. As such, identifying *degenerate* motion profiles, which cause extra unobservable directions for either the navigation or calibration parameters, is of practical importance. As the degenerate motion analysis of the navigation state for VINS has been studied in the prior work [35], [45], [46], we here focus only on motions that cause the calibration parameters to become unobservable.

A. IMU Intrinsic Parameters

A selection of basic motion types, which can cause the IMU intrinsics to become unobservable for imu22$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^I_a\mathbf{R}, \mathbf{T}_{g9}\}$, are identified. Note that similar results hold for other IMU model variants.

1) Degenerate motions for $\mathbf{D}_w$*:* As the gyroscope related IMU intrinsics $\mathbf{D}_w$ are coupled with gyroscope bias $\mathbf{b}_g$ and the angular velocity readings ${}^w\boldsymbol{\omega}$ from the IMU, we have the following results:

Lemma 2. *If any component of* ${}^w\boldsymbol{\omega}$ *(including* ${}^w\omega_1$*,* ${}^w\omega_2$*,* ${}^w\omega_3$*) is constant, then* $\mathbf{D}_w$ *will become unobservable.*

Proof. If ${}^w w_1$ is constant, d_{w1} will be unobservable with unobservable directions as:

$$\mathbf{N}_{w1} = \begin{bmatrix} \mathbf{0}_{1\times 9} & (\hat{\mathbf{D}}_w^{-1} {}^I_w\hat{\mathbf{R}}^\top \mathbf{e}_1)^\top {}^w w_1 & \mathbf{0}_{1\times 3} & 1 & \mathbf{0}_{1\times 42} \end{bmatrix}^\top \tag{59}$$

If ${}^w w_2$ is constant, d_{w2} and d_{w3} will be unobservable with unobservable directions as:

$$\mathbf{N}_{w2} = \begin{bmatrix} \mathbf{0}_{1\times 9} & (\hat{\mathbf{D}}_w^{-1} {}^I_w\hat{\mathbf{R}}^\top \mathbf{e}_1)^\top {}^w w_2 & \mathbf{0}_{1\times 4} & 1 & \mathbf{0}_{1\times 41} \\ \mathbf{0}_{1\times 9} & (\hat{\mathbf{D}}_w^{-1} {}^I_w\hat{\mathbf{R}}^\top \mathbf{e}_2)^\top {}^w w_2 & \mathbf{0}_{1\times 5} & 1 & \mathbf{0}_{1\times 40} \end{bmatrix}^\top \tag{60}$$

If ${}^w w_3$ is constant, d_{w4}, d_{w5} and d_{w6} are unobservable with unobservable directions as:

$$\mathbf{N}_{w3} = \begin{bmatrix} \mathbf{0}_{1\times 9} & (\hat{\mathbf{D}}_w^{-1} {}^I_w\hat{\mathbf{R}}^\top \mathbf{e}_1)^\top {}^w w_3 & \mathbf{0}_{1\times 6} & 1 & \mathbf{0}_{1\times 39} \\ \mathbf{0}_{1\times 9} & (\hat{\mathbf{D}}_w^{-1} {}^I_w\hat{\mathbf{R}}^\top \mathbf{e}_2)^\top {}^w w_3 & \mathbf{0}_{1\times 7} & 1 & \mathbf{0}_{1\times 38} \\ \mathbf{0}_{1\times 9} & (\hat{\mathbf{D}}_w^{-1} {}^I_w\hat{\mathbf{R}}^\top \mathbf{e}_3)^\top {}^w w_3 & \mathbf{0}_{1\times 8} & 1 & \mathbf{0}_{1\times 37} \end{bmatrix}^\top \tag{61}$$

□

2) Degenerate motions for $\mathbf{D}_a$*:* Similarly, as ${}^a\mathbf{a}$ can affect the observability property for the accelerometer related IMU intrinsics $\mathbf{D}_a$, we have:

Lemma 3. *If any component of* ${}^a\mathbf{a}$ *(including* ${}^a a_1$*,* ${}^a a_2$ *and* ${}^a a_3$*) is constant, then* $\mathbf{D}_a$ *will become unobservable.*

Proof. If ${}^a a_1$ is constant, d_{a1}, pitch and yaw of ${}^I_a\mathbf{R}$ are unobservable with unobservable directions as:

$$\mathbf{N}_{a1} = \begin{bmatrix} \mathbf{0}_{12\times 1} & \mathbf{0}_{12\times 1} & \mathbf{0}_{12\times 1} \\ \hat{\mathbf{D}}_a^{-1}\mathbf{e}_1 {}^a a_1 & \hat{\mathbf{D}}_a^{-1}\mathbf{e}_2 \hat{d}_{a1} {}^a a_1 & \hat{\mathbf{D}}_a^{-1}\mathbf{e}_3 \hat{d}_{a1}\hat{d}_{a3} {}^a a_1 \\ \mathbf{0}_{6\times 1} & \mathbf{0}_{6\times 1} & \mathbf{0}_{6\times 1} \\ 1 & 0 & 0 \\ 0 & \hat{d}_{a3} & 0 \\ 0 & -\hat{d}_{a2} & 0 \\ 0 & \hat{d}_{a5} & \hat{d}_{a6}\hat{d}_{a3} \\ 0 & -\hat{d}_{a4} & -\hat{d}_{a2}\hat{d}_{a6} \\ 0 & 0 & \hat{d}_{a2}\hat{d}_{a5} - \hat{d}_{a4}\hat{d}_{a3} \\ \mathbf{0}_{3\times 1} & -{}^I_a\hat{\mathbf{R}}\mathbf{e}_3 & {}^I_a\hat{\mathbf{R}}(\mathbf{e}_1\hat{d}_{a2} + \mathbf{e}_2\hat{d}_{a3}) \\ \mathbf{0}_{28\times 1} & \mathbf{0}_{28\times 1} & \mathbf{0}_{28\times 1} \end{bmatrix} \tag{62}$$

TABLE II: Degenerate motions for imu22$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^I_a\mathbf{R}, \mathbf{T}_{g9}\}$ intrinsic parameters.

Motion Types	Dim.	Unobservable Parameters
constant ${}^w\omega_1$	1	d_{w1}
constant ${}^w\omega_2$	2	d_{w2}, d_{w3}
constant ${}^w\omega_3$	3	d_{w4}, d_{w5}, d_{w6}
constant ${}^a a_1$	3	d_{a1}, pitch and yaw of ${}^I_a\mathbf{R}$
constant ${}^a a_2$	3	d_{a2}, d_{a3}, roll of ${}^I_a\mathbf{R}$
constant ${}^a a_3$	3	d_{a4}, d_{a5}, d_{a6}
constant ${}^I a_1$	3	t_{g1}, t_{g2}, t_{g3}
constant ${}^I a_2$	3	t_{g4}, t_{g5}, t_{g6}
constant ${}^I a_3$	3	t_{g7}, t_{g8}, t_{g9}

If ${}^a a_2$ is constant, d_{a2}, d_{a3} and roll of ${}^I_a\mathbf{R}$ are unobservable with unobservable directions as:

$$\mathbf{N}_{a2} = \begin{bmatrix} \mathbf{0}_{12\times 1} & \mathbf{0}_{12\times 1} & \mathbf{0}_{12\times 1} \\ \hat{\mathbf{D}}_a^{-1}\mathbf{e}_1 {}^a a_2 & \hat{\mathbf{D}}_a^{-1}\mathbf{e}_2 {}^a a_2 & \hat{\mathbf{D}}_a^{-1}\mathbf{e}_3 \hat{d}_{a3} {}^a a_2 \\ \mathbf{0}_{6\times 1} & \mathbf{0}_{6\times 1} & \mathbf{0}_{6\times 1} \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \hat{d}_{a6} \\ 0 & 0 & -\hat{d}_{a5} \\ \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 1} & -{}^I_a\hat{\mathbf{R}}\mathbf{e_1} \\ \mathbf{0}_{28\times 1} & \mathbf{0}_{28\times 1} & \mathbf{0}_{28\times 1} \end{bmatrix} \tag{63}$$

If ${}^a a_3$ is constant, d_{a4}, d_{a5} and d_{a6} are unobservable with unobservable directions as:

$$\mathbf{N}_{a3} = \begin{bmatrix} \mathbf{0}_{1\times 12} & (\hat{\mathbf{D}}_a^{-1}\mathbf{e}_1)^\top {}^a a_3 & \mathbf{0}_{1\times 9} & 1 & \mathbf{0}_{1\times 33} \\ \mathbf{0}_{1\times 12} & (\hat{\mathbf{D}}_a^{-1}\mathbf{e}_2)^\top {}^a a_3 & \mathbf{0}_{1\times 10} & 1 & \mathbf{0}_{1\times 32} \\ \mathbf{0}_{1\times 12} & (\hat{\mathbf{D}}_a^{-1}\mathbf{e}_3)^\top {}^a a_3 & \mathbf{0}_{1\times 11} & 1 & \mathbf{0}_{1\times 31} \end{bmatrix}^\top \tag{64}$$

□

3) Degenerate motions for $\mathbf{T}_g$*:* As ${}^I\mathbf{a}$ (the acceleration in IMU frame) can affect the observability property for the g-sensitivity $\mathbf{T}_g$, by close inspection of special configurations for ${}^I\mathbf{a}$, we have:

Lemma 4. *If any component of* ${}^I\mathbf{a}$ *(including* ${}^I a_1$*,* ${}^I a_2$ *and* ${}^I a_3$*) is constant, then* $\mathbf{T}_g$ *will become unobservable.*

Proof. If ${}^I a_1$ is constant, t_{g1}, t_{g2} and t_{g3} are unobservable with unobservable directions as:

$$\mathbf{N}_{g1} = \begin{bmatrix} \mathbf{0}_{3\times 9} & \mathbf{I}_3 {}^I a_1 & \mathbf{0}_{3\times 18} & -\mathbf{I}_3 & \mathbf{0}_{3\times 25} \end{bmatrix}^\top \tag{65}$$

If ${}^I a_2$ is constant, t_{g4}, t_{g5} and t_{g6} are unobservable with unobservable directions as:

$$\mathbf{N}_{g2} = \begin{bmatrix} \mathbf{0}_{3\times 9} & \mathbf{I}_3 {}^I a_2 & \mathbf{0}_{3\times 21} & -\mathbf{I}_3 & \mathbf{0}_{3\times 22} \end{bmatrix}^\top \tag{66}$$

If ${}^I a_3$ is constant, t_{g7}, t_{g8} and t_{g9} are unobservable with unobservable directions as:

$$\mathbf{N}_{g3} = \begin{bmatrix} \mathbf{0}_{3\times 9} & \mathbf{I}_3 {}^I a_3 & \mathbf{0}_{3\times 24} & -\mathbf{I}_3 & \mathbf{0}_{3\times 19} \end{bmatrix}^\top \tag{67}$$

□

4) Remarks: It is evident from the above analysis that the IMU intrinsic calibration is sensitive to sensor motion and thus all 6 axes need to be excited to ensure all of them can be calibrated. These findings are summarized in Table II. It should be noted that any combination of these primitive motions is still degenerate and causes all related parameters to become unobservable (e.g., planar motion with constant acceleration). It is also important to mention that it is common that ${}^I_a\mathbf{R} \simeq \mathbf{I}_3$

TABLE III: Degenerate motions for IMU-camera spatial-temporal calibration.

Motion Types	Unobservable Parameters	Observable
pure translation	${}^C\mathbf{p}_I$	${}^C_I\mathbf{R}$, t_d, t_r
one-axis rotation	${}^C\mathbf{p}_I$ along rotation axis	${}^C_I\mathbf{R}$, t_d, t_r
constant ${}^I\boldsymbol{\omega}$ constant ${}^I\mathbf{v}$	t_d and ${}^C\mathbf{p}_I$ along rotation axis	${}^C_I\mathbf{R}$, t_r
constant ${}^I\boldsymbol{\omega}$ constant ${}^G\mathbf{a}$	t_d and ${}^C\mathbf{p}_I$ along rotation axis	${}^C_I\mathbf{R}$, t_r

and $\mathbf{D}_a \simeq \mathbf{I}_3$ for most IMUs, and thus, ${}^a\hat{\mathbf{a}} \simeq {}^I\mathbf{a}$. As such, the degenerate motions for $\mathbf{D}_a$ might also lead to the calibration failures of $\mathbf{T}_g$, and vice-versa. Again, this degenerate motion analysis can be extended to other model variants, which is omitted here for brevity.

B. IMU-Camera Spatial-Temporal Parameters

In our previous work [16] which investigated four commonly-seen degenerate motions of VINS with only IMU-camera spatial-temporal calibration, we here show these degenerate motions hold true for VINS with full-parameter calibration:

Lemma 5. *The IMU-camera spatial-temporal calibration will become unobservable, if the sensor platform undergoes the following degenerate motions:*

- *Pure translation*
- *One-axis rotation*
- *Constant local angular and linear velocity*
- *Constant local angular velocity and global linear acceleration*

Proof. If the system undergoes pure translation (no rotation), the translation part ${}^C\mathbf{p}_I$ of the spatial calibration will be unobservable and is:

$$\mathbf{N}_{pt} = \begin{bmatrix}\mathbf{0}_{3\times45} & \mathbf{I}_3 & \mathbf{0}_{3\times10} & -({}^G_{I_1}\hat{\mathbf{R}}{}^I_C\hat{\mathbf{R}})^\top\end{bmatrix}^\top \tag{68}$$

If the system undergoes random (general) translation but with only one-axis rotation, the translation calibration ${}^C\mathbf{p}_I$ along the rotation axis will be unobservable, with the following unobservable direction:

$$\mathbf{N}_{oa} = \begin{bmatrix}\mathbf{0}_{1\times45} & ({}^C_I\hat{\mathbf{R}}{}^I\hat{\mathbf{k}})^\top & 0_{1\times10} & -({}^G_{I_1}\hat{\mathbf{R}}{}^I\hat{\mathbf{k}})^\top\end{bmatrix}^\top \tag{69}$$

where ${}^I\mathbf{k}$ is the constant rotation axis in the IMU frame $\{I\}$. If the VINS undergoes constant local angular velocity ${}^I\boldsymbol{\omega}$ and linear velocity ${}^I\mathbf{v}$, the time offset t_d will be unobservable with the following unobservable direction:

$$\mathbf{N}_{t1}=\begin{bmatrix}\mathbf{0}_{1\times42} & ({}^C_I\hat{\mathbf{R}}{}^I\hat{\boldsymbol{\omega}})^\top & -({}^C_I\hat{\mathbf{R}}{}^I\hat{\mathbf{v}})^\top & -1 & \mathbf{0}_{1\times12}\end{bmatrix}^\top \tag{70}$$

If the VINS undergoes constant local angular velocity ${}^I\boldsymbol{\omega}$ and global acceleration ${}^G\mathbf{a}$, the time offset t_d will be unobservable with the following unobservable direction:

$$\mathbf{N}_{t2} = \tag{71}$$

$$\begin{bmatrix}\mathbf{0}_{1\times6} & {}^G\hat{\mathbf{a}} & \mathbf{0}_{1\times30} & ({}^C_I\hat{\mathbf{R}}{}^I\hat{\boldsymbol{\omega}})^\top & \mathbf{0}_{1\times3} & -1 & \mathbf{0}_{1\times9} & -({}^G\hat{\mathbf{v}}_{I_1})^\top\end{bmatrix}^\top$$

□

Table III summarizes these degenerate motions for completeness. It is important to note that unlike t_d (whose Jacobian is mainly affected by the sensor motion), the Jacobian for RS readout time, t_r, is also affected by the feature observations due to the term $\frac{m}{M}$ [see Eq. (96)], and is observable, as hundreds of features can be observed from different image rows during exploration.

C. Camera Intrinsic Parameters

As mentioned before, the camera intrinsics are mainly affected by the observed feature structure. By investigating special feature configurations, we find the following degenerate case for camera calibration when using a *radtan* distortion model:

Lemma 6. *The camera intrinsics will become unobservable if the following conditions are satisfied:*

- *The features keep the same depth relative to the camera (e.g., ${}^Cz_f$ is constant).*
- *The camera moves with one-axis rotation and the rotation axis is defined as ${}^C\mathbf{k} = \mathbf{e}_3$.*

Proof. The camera focal length f_u, f_v, the camera distortion model k_1, k_2, p_1 and p_2 will become unobservable along with the unobservable direction:

$$\mathbf{N}_{Cin} = \begin{bmatrix}\mathbf{0}_{1\times47} & \hat{f}_u & \hat{f}_v & \mathbf{0}_{1\times2} & 2\hat{k}_1 & 4\hat{k}_2 & \hat{p}_1 & \hat{p}_2 & {}^G\mathbf{k}^\top\end{bmatrix}^\top \tag{72}$$

with ${}^G\mathbf{k} = {}^G_{I_1}\hat{\mathbf{R}}{}^I_C\hat{\mathbf{R}}{}^C\mathbf{k}{}^Cz_f$. □

As an example, if a ground vehicle is performing planar motion with an upward-facing camera only observing features from the ceilings, the above two conditions will hold and thus the camera intrinsics with *radtan* distortion model will be unobservable. Nevertheless, since it is common to observe hundreds of features, it might be rarely the case that every feature maintains the same relative depth, ${}^Cz_f$, to the camera, and thus, this degeneracy may not happen in practice if features are tracked uniformly throughout images.

It is interesting to point out that this degenerate case does not hold for camera models with *equidist* distortion. By noting that the following *equidist* model dislikes the *radtan* distortion model [see Eq. (14)], it is not difficult to verify that the above *radtan* unobservable subspace [see Eq. (72)] is no longer valid if using the *equidist* model:

$$\begin{bmatrix}u_d \\ v_d\end{bmatrix} = \begin{bmatrix}\frac{u_n}{r}\theta(1+k_1\theta^2+k_2\theta^4+p_1\theta^6+p_2\theta^8) \\ \frac{v_n}{r}\theta(1+k_1\theta^2+k_2\theta^4+p_1\theta^6+p_2\theta^8)\end{bmatrix} \tag{73}$$

where $\theta = \operatorname{atan}(r)$. For example, in an extreme case, if distortion parameters are all zeros (i.e., $k_1 = k_2 = p_1 = p_2 = 0$), images with the *radtan* model become distortion-free, i.e., $[u_d\ v_d]^\top = [u_n\ v_n]^\top$, while the *equidist* model still possesses the radial distortion, i.e., $[u_d\ v_d]^\top = [\frac{u_n}{r}\theta\ \frac{v_n}{r}\theta]^\top$. In fact, we are unable to analytically find a similar unobservable subspace for the camera intrinsics with *equidist* distortion even when all features have the same relative depth.

VII. Visual-Inertial Estimator Design

Leveraging our MSCKF-based VINS estimator [10], the proposed estimator extends the state vector $\mathbf{x}_k$ at time step k to include the current IMU state $\mathbf{x}_{I_k}$, a sliding window of cloned IMU poses $\mathbf{x}_c$, the camera calibration parameters ($\mathbf{x}_{IC}$ and $\mathbf{x}_{Cin}$) and feature state $\mathbf{x}_f$.

$$\mathbf{x}_k = \begin{bmatrix}\mathbf{x}_{I_k}^\top & \mathbf{x}_c^\top & \mathbf{x}_{IC}^\top & \mathbf{x}_{Cin}^\top & \mathbf{x}_f^\top\end{bmatrix}^\top \tag{74}$$

$$\mathbf{x}_c = \begin{bmatrix}{}^{I_{ck-1}}_{G}\bar{q}^\top & {}^{G}\mathbf{p}_{I_{ck-1}}^\top & \cdots & {}^{I_{ck-n}}_{G}\bar{q}^\top & {}^{G}\mathbf{p}_{I_{ck-n}}^\top\end{bmatrix}^\top \tag{75}$$

where $\mathbf{x}_I$, $\mathbf{x}_{IC}$, $\mathbf{x}_{Cin}$ and $\mathbf{x}_f$ are the same as Eq. (20), $\mathbf{x}_c$ denotes the sliding window containing n cloned IMU poses with index from $ck-n$ to $ck-1$. Note that the IMU intrinsics $\mathbf{x}_{in}$ are contained in the current IMU state $\mathbf{x}_{I_k}$.

As $\mathbf{x}_c$, $\mathbf{x}_{IC}$, $\mathbf{x}_{Cin}$ and $\mathbf{x}_f$ have zero dynamics, we only propagate the estimate and covariance of the next IMU state based on Eq. (27)-(31) and Eq. (46), which all incorporate the IMU intrinsics $\mathbf{x}_{in}$.

As in [6], we handle the IMU-camera time offset t_d when we clone the "true" IMU pose corresponding to image measurements. For example, if we clone the current IMU pose $\{{}^{I_k}_{G}\bar{q}, {}^{G}\mathbf{p}_{I_k}\}$ into the sliding window as $\{{}^{I_{ck}}_{G}\bar{q}, {}^{G}\mathbf{p}_{I_{ck}}\}$ using:

$${}^{G}_{I_{ck}}\mathbf{R} \simeq {}^{G}_{I_k}\mathbf{R}\exp({}^{I_k}\hat{\boldsymbol{\omega}}\tilde{t}_d) \tag{76}$$

$${}^{G}\mathbf{p}_{I_{ck}} \simeq {}^{G}\mathbf{p}_{I_k} + {}^{G}\mathbf{v}_{I_k}\tilde{t}_d \tag{77}$$

with the linearized clone Jacobians as:

$$\begin{bmatrix}\delta\boldsymbol{\theta}_{I_{ck}} \\ {}^{G}\tilde{\mathbf{p}}_{I_{ck}}\end{bmatrix} \simeq \begin{bmatrix}\mathbf{I}_3 & \mathbf{0}_3 & {}^{I_k}\hat{\boldsymbol{\omega}} \\ \mathbf{0}_3 & \mathbf{I}_3 & {}^{G}\hat{\mathbf{v}}_{I_k}\end{bmatrix}\begin{bmatrix}\delta\boldsymbol{\theta}_{I_k} \\ {}^{G}\tilde{\mathbf{p}}_{I_k} \\ \tilde{t}_d\end{bmatrix} \tag{78}$$

Both $\mathbf{x}_{in}$ and t_d will be updated through correlations when visual feature measurements are present.

We utilize first-estimates Jacobians (FEJ) [6], [34] to preserve the system unobservable subspace and improve the estimator consistency. We directly model the camera intrinsic and IMU-camera spatial calibration through the visual measurement functions [see Eq. (50)] and update them in the filter with Jacobians in Eq. (55).

For the RS cameras, the feature measurements from different image rows are captured at different timestamps. This indicates that we cannot directly find a cloned pose in the sliding window for $\{{}^{G}_{I(t)}\mathbf{R}, {}^{G}\mathbf{p}_{I(t)}\}$ shown in Eq. (17). Therefore, for the readout time calibration, we model the feature measurement affected by RS effects through pose interpolation [7], [11]. For example, if the feature measurement is in the m-th row with total M rows in an image, we can find two bounding clones $ci-1$ and ci based on the measurement time t. Hence, the corresponding time t is between two clones within the sliding window, that is: $t_{ck-n} \le t_{ci-1} \le t \le t_{ci} \le t_{ck}$. We can then find the *virtual* IMU pose $\{{}^{G}_{I(t)}\mathbf{R}, {}^{G}\mathbf{p}_{I(t)}\}$ between clones $ci-1$ and ci with:

$$\lambda = (t_I + \frac{m}{M}t_r - t_{ci-1})/(t_{ci} - t_{ci-1}) \tag{79}$$

$${}^{G}_{I(t)}\mathbf{R} = {}^{G}_{I_{ci-1}}\mathbf{R}\exp\left(\lambda\log\left({}^{G}_{I_{ci-1}}\mathbf{R}^\top{}^{G}_{I_{ci}}\mathbf{R}\right)\right) \tag{80}$$

$${}^{G}\mathbf{p}_{I(t)} = (1-\lambda){}^{G}\mathbf{p}_{I_{ci-1}} + \lambda{}^{G}\mathbf{p}_{I_{ci}} \tag{81}$$

To summarize, feature measurements which occur at different rows of the image can be related to the state vector defined in Eq. (74) through the above linear pose interpolation. This measurement function can then be linearized for use in the EKF update [47]. Note that a higher-order polynomial pose interpolation is used by [11] and can be utilized if necessary.

TABLE IV: Simulation parameters and prior σs that perturbations were drawn from.

Parameter	Value	Parameter	Value
IMU Scale	0.006	IMU Skew	0.006
Rot. atoI (rad)	0.008	Rot. wtoI (rad)	0.008
Gyro. Noise ($\mathrm{rad\,s^{-1}\sqrt{Hz^{-1}}}$)	1.6968e-04	Gyro. Bias ($\mathrm{rad\,s^{-2}\sqrt{Hz^{-1}}}$)	1.9393e-05
Accel. Noise ($\mathrm{m\,s^{-2}\sqrt{Hz^{-1}}}$)	0.002	Accel. Bias ($\mathrm{m\,s^{-3}\sqrt{Hz^{-1}}}$)	0.003
Focal Len. (px/m)	1.0	Cam. Center (px)	1.0
d1 and d2	0.008	d3 and d4	0.002
Rot. CtoI (rad)	0.010	Pos. IinC (m)	0.010
Readout Time (ms)	0.5	Timeoff (s)	0.005
Cam Freq. (Hz)	20	IMU Freq. (Hz)	400

VIII. Simulation Analysis

The proposed estimator is implemented within OpenVINS [10], which contains a visual-inertial simulator and a real-time modular sliding window EKF-based VINS estimator. The basic configurations for our simulator are listed in Table IV. To simulate RS visual bearing measurements, we follow the logic presented by [6] and [11]. Static environmental features are first generated along the trajectory at random depths and bearings. Then, for a given imaging time, we project each feature in view into the current image frame using the true camera intrinsic and distortion model and find the corresponding observation row. Given this projected row and image time, we can find the pose at which that RS row should have been exposed. We can then re-project this feature into the new pose and iterate until the projected row does not change (which typically requires 2-3 iterations). We now have a feature measurement which occurs at the correct pose given its RS row. This measurement is then corrupted with white noise. The imaging timestamp corresponding to the starting row is then shifted by the true IMU-camera time offset t_d to simulate cross-sensor delay.

It is important to note that, in the following paper, we only present the most prominent results due to space limits, while comprehensive simulation and experimental results can be found in our companion technical report [44].

A. Simulation with Fully-Excited Motion

We first perform a simulation with full calibration on a fully-excited trajectory. Note that the perturbations added to initial calibration are similar or larger than real-world situations. For example, the perturbations to IMU scale scalars can be as large as 0.02, which is even larger than the real world results (see Fig. 14). The perturbations to each component of camera translation can be as large as $2.5\,\mathrm{cm}$, which is also challenging. The trajectory, shown in the left of Fig. 2, is designed based on *tum_corridor* sequence of TUM visual-inertial dataset with full excitation of all 6 axes and provides a realistic 3D hand-held motion [48]. From the results shown in Fig. 3, the estimation errors and 3σ bounds for the calibration parameters (including $\mathtt{imu22}\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}, \mathbf{T}_{g9}\}$ and *radtan*) can converge quite nicely, verifying that the analysis for general motions holds true. We plot results from six different realizations of the initial calibration guesses based on the specified priors, and it is clear that the estimates for all these calibration parameters are able to converge from different initial guesses to near the ground truth. Each parameter is able to "gain" information since their 3σ bounds shrink. These results verify our Lemma 1 that all these calibration parameters are observable given a fully-excited motion.

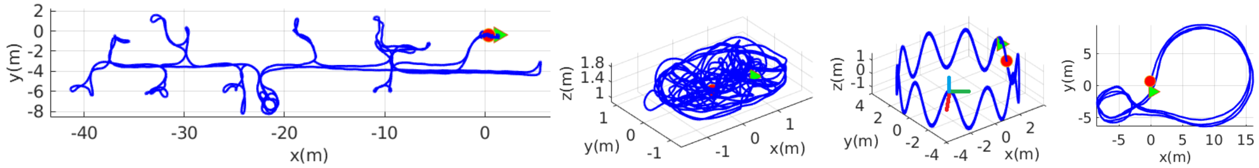


Fig. 2: Simulated trajectories. Left: *tum_corridor* with fully excited 3D motion; Middle left: *tum_room* with 1 axis rotation and 3D translation; Middle right: *sine_3d* with constant centripetal acceleration along local IMU x-axis (red-axis); Right: *udel_gore* planar motion with constant z and only yaw rotation. The green triangle and red circle denote the beginning and ending of these trajectories, respectively.

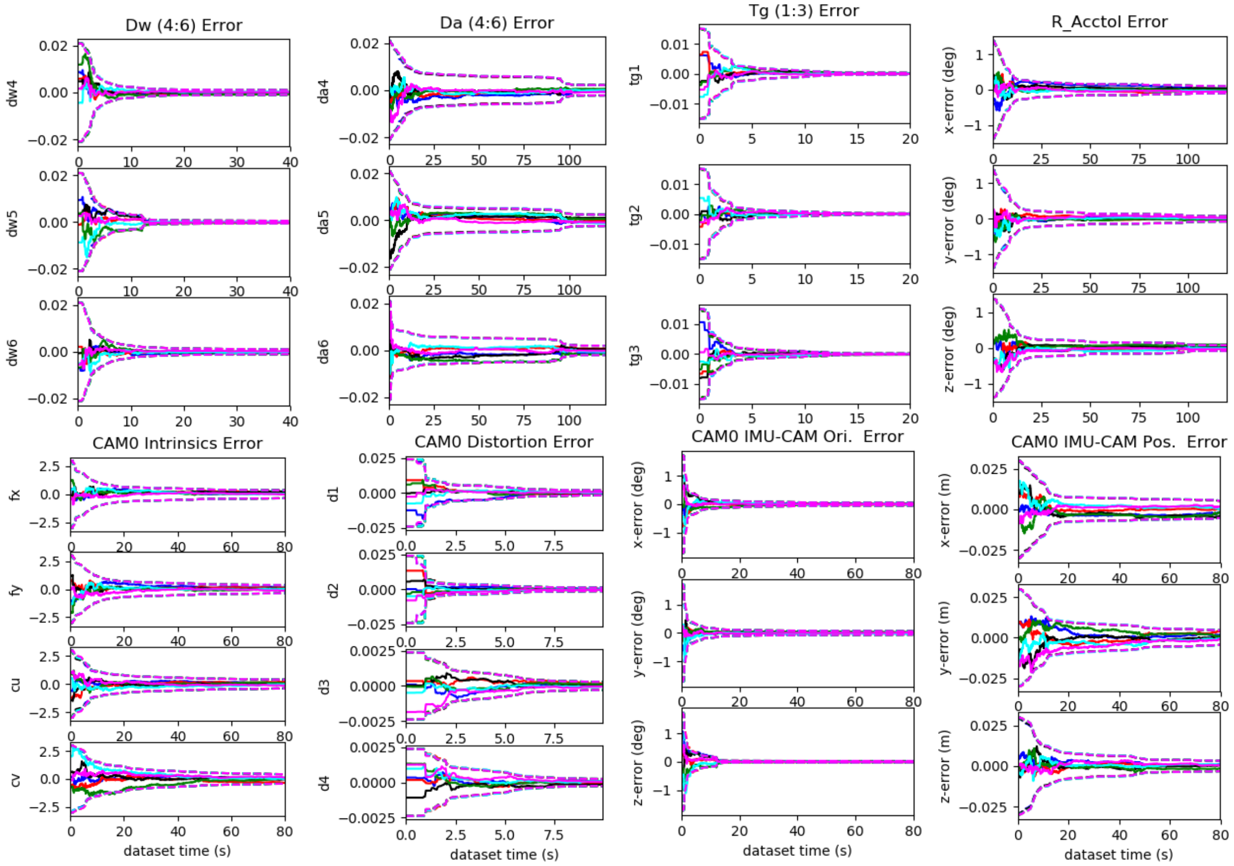


Fig. 3: Evaluation on *tum_corridor* with fully excited motion (using *imu22*$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}, \mathbf{T}_{g9}\}$ and *radtan*). 3σ bounds (dotted lines) and estimation errors (solid lines) for six different runs (different colors) are shown.

B. Sensitivities to Perturbations

The next question is how robust the system is to the initial perturbations and whether the use of online sensor calibration enables improvements in robustness and accuracy. Shown in Fig. 4, for each of the different calibration parameters we perturb with different noise level (following Gaussian distribution with σ from x-axis of each plot) on the *tum_corridor* trajectory (note that we also change the initial prior provided to the filter as its distribution has changed). For example, when perturbing the IMU accelerometer scale parameter, the d_{a1}, d_{a3} and d_{a6} from the diagonal of $\mathbf{D}_{a6}$ are all perturbed. The perturbed values might be as large as the 3σ indicated from the x-axis.

We can see that the proposed estimator is relatively invariant to the initial inaccuracies of the parameters and is, in general, able to output a near constant trajectory error. A filter, which does not perform this online estimation, has its trajectory estimation error quickly increase to non-usable levels. It is interesting to see that even small perturbations to calibration parameters can cause huge trajectory errors which further verifies the motivations to perform online calibration.

C. Degenerate Motion Verification

We now verify the identified degenerate motions and present simulation results for three special motions. In all simulations, we perform full-parameter calibration. The trajectories[2] shown in Fig. 2 are created as follows:

- One-axis rotation with a modified *tum_room* trajectory, see middle left, which removes roll and pitch changes and creates a yaw-only dataset but still with 3D translation.
- Constant local ${}^{I}a_x$ with modified *sine_3d*, see middle right, for which we have a constant global pitch and make the global yaw rotation tangent to the trajectory in the x-y plane (gives constant centripetal acceleration along local IMU x-axis).

[2] A demo video can be found at: https://youtu.be/MP4ADABtqXQ

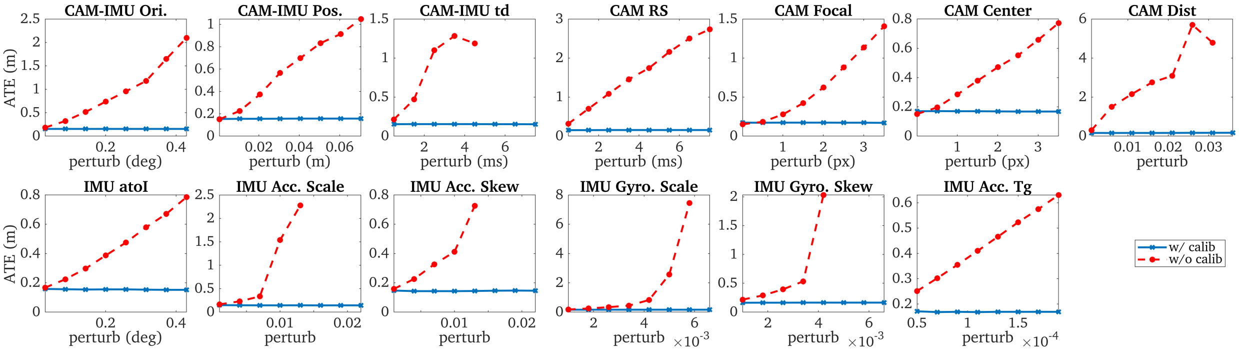


Fig. 4: Average absolute trajectory errors (ATE) over five runs using $\texttt{imu22}\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^I_a\mathbf{R}, \mathbf{T}_{g9}\}$ and *radtan* on the *tum_corridor* with full 3D motion given different levels of perturbation. Only the calibrated parameter was perturbed while other parameters were initialized to their true values and not estimated. ATE above eight meters were not reported and can be considered as divergence.

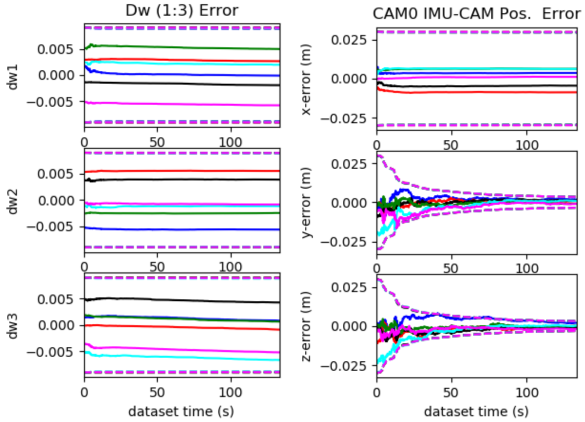


Fig. 5: Evaluation on *tum_room* with one-axis rotation using $\texttt{imu22}\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^I_a\mathbf{R}, \mathbf{T}_{g9}\}$. Note that the estimation errors and 3σ bounds for d_{w1}, d_{w2}, d_{w3} and the IMU-camera position calibration along the rotation axis can not converge.

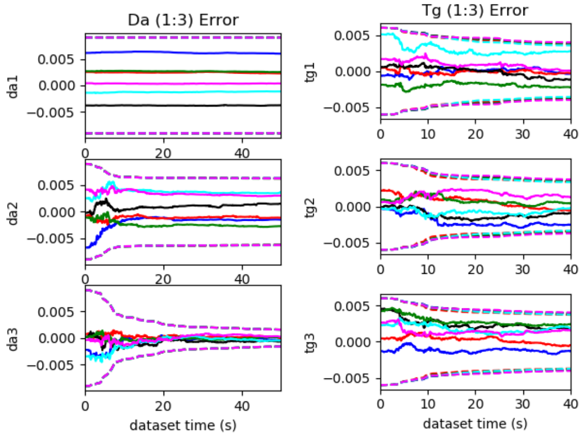


Fig. 6: Evaluation on the *sine_3d* with constant acceleration along x-axis using $\texttt{imu22}\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^I_a\mathbf{R}, \mathbf{T}_{g9}\}$. The estimation errors and 3σ bounds for d_{a1}, pitch and yaw of ${}^I_a\mathbf{R}$ cannot converge. Note that t_{g1}, t_{g2} and t_{g3} also converge very slow.

- Planar motion with modified *udel_gore*, see right, which removes roll and pitch changes and all poses are projected to the x-y plane by removing z (planar motion in the global x-y plane).

1) One-axis rotation motion: Shown in Fig. 5, the first 3 parameters (d_{w1}, d_{w2} and d_{w3}) for $\mathbf{D}_w$ do not converge at all (the 3σ bounds are almost straight lines), which matches our analysis, see Table II. These parameters should be unobservable in the case of one-axis rotation with ${}^w w_x$ (roll) and ${}^w w_y$ (pitch) are constant. Additionally, the translation between IMU and camera does not converge either. The x-error of the IMU-camera translation does not converge at all, reinforcing the undesirability of degenerate motions and verifies the analysis in Table III.

2) Constant local acceleration motion: The results shown in Fig. 6, where we have enforced that the local acceleration along the x-axis, a_x, is constant. The d_{a1}, and pitch and yaw of ${}^I_a\mathbf{R}$ does not converge, thus validating our analysis shown in Table II. Note that in the simulation, we have set ${}^I_a\mathbf{R} \simeq \mathbf{I}_3$ and $\mathbf{D}_a \simeq \mathbf{I}_3$. Hence, ${}^a\hat{\mathbf{a}} \simeq {}^I\mathbf{a}$ and ${}^I a_x$ is also near constant. Therefore, three terms of g-sensitivity (t_{g1}, t_{g2} and t_{g2}) are also unobservable and converge much slower than other terms.

3) Planar motion: Shown in Fig. 7, with one-axis rotation (yaw axis) for planar motion the d_{w1}, d_{w2} and d_{w3} for $\mathbf{D}_w$ and the IMU-camera translation are unobservable and do not converge. Since the ${}^I a_z$ is constant, the last three terms of g-sensitivity (t_{g7}, t_{g8} and t_{g9}) become unobservable and cannot converge. Both these results verify our analysis shown in Tables II and III. Additionally, this trajectory is quite smooth with small excitation of linear acceleration, hence, the terms of $\mathbf{D}_a$ and ${}^I_a\mathbf{R}$ in general converge much slower than the fully excited motion case.

D. Simulated Over Parametrization

We now look to investigate the impact of poor choice of calibration parameters which *over parameterizes* the IMU intrinsics. The $\texttt{imu5}\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^I_w\mathbf{R}, {}^I_a\mathbf{R}\}$ model, see Table I, is an over parametrization since we calibrate both 9 parameters for gyroscope and accelerometer. This causes the IMU-camera orientation to be affected since the intermediate inertial frame $\{I\}$ is not constrained. If we change the relative rotation from $\{I\}$ to $\{C\}$, then this perturbed rotation can be absorbed into the $\{a\}$ to $\{I\}$ and $\{w\}$ to $\{I\}$ terms. Thus, it means

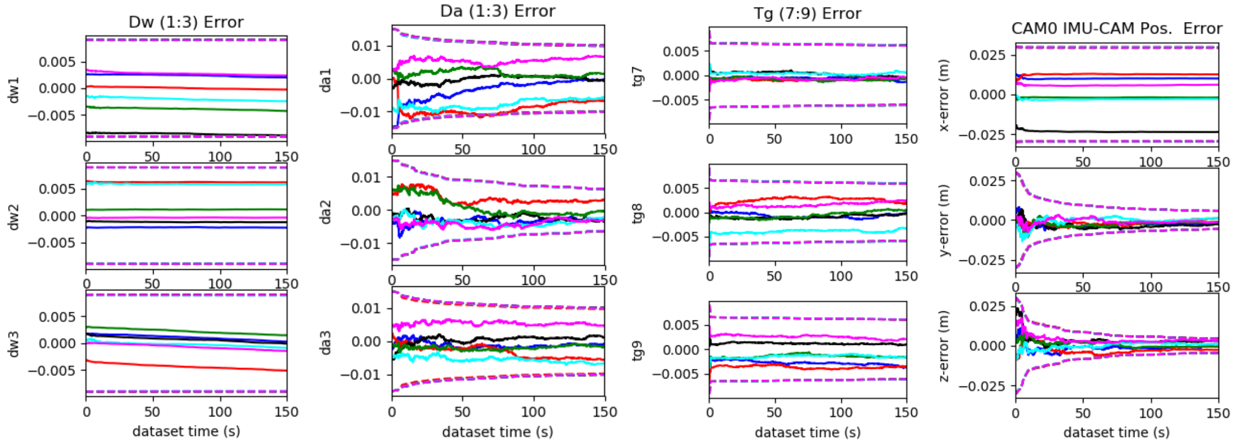


Fig. 7: Evaluation on *udel_gore* with planar motion using imu22$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}, \mathbf{T}_{g9}\}$ and *radtan*. With planar motion, the estimation errors and 3σ bounds of d_{w1}, d_{w2}, d_{w3}, t_{g7}, t_{g8}, t_{g9} and the IMU-camera position cannot converge. Due to lack of motion excitation, the parameters of $\mathbf{D}_a$ and ${}^{I}_{a}\mathbf{R}$ converge much slower than the other motion cases.

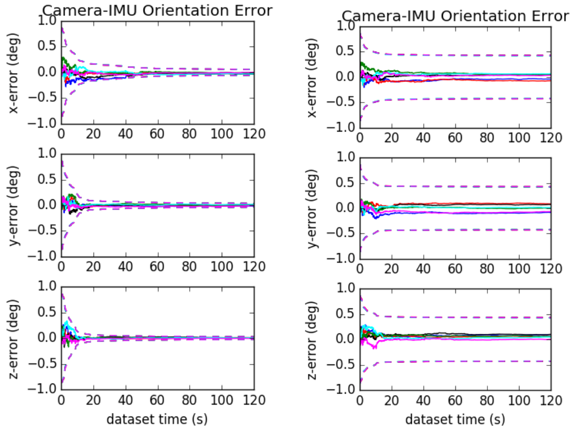


Fig. 8: Camera to IMU orientation errors when using IMU imu2$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}\}$ (left) and the over parameterized imu5$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{w}\mathbf{R}, {}^{I}_{a}\mathbf{R}\}$ (right). Note that only the IMU intrinsics and relative pose between IMU and camera were calibrated online.

we have an extra unobservable directions for the rotation not constrained by our measurements. We compare imu5$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{w}\mathbf{R}, {}^{I}_{a}\mathbf{R}\}$ model to its close equivalent imu2$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}\}$ model in Fig. 8. We can see that even though the trajectory fully excites the sensor platform, the convergence of ${}^{C}_{I}\mathbf{R}$ becomes much worse if we calibrate IMU-camera extrinsics and all 18 parameters for imu5$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{w}\mathbf{R}, {}^{I}_{a}\mathbf{R}\}$ even when the *same* priors and measurements are used. This further motivates the use of minimal calibration parameters to ensure fast and robust convergence of all states.

IX. Real-World Experimental Validation on TUM RS VIO Datasets

The proposed algorithm is first evaluated on the TUM RS VIO dataset [26], which contains a time-synchronized stereo pair of two uEye UI-3241LE-M-GL cameras (left: global-shutter and right: rolling-shutter) and a Bosch BMI160 IMU. When collecting data, the cameras were operated at 20 Hz while the IMU operated at 200 Hz and an OptiTrack system captured the ground truth motion. The dataset is provided in both "raw" and "calibrated" formats. The "calibrated" dataset has IMU intrinsic corrections pre-applied to the "raw" dataset. We evaluate our proposed system by using the right (RS) camera directly with the raw datasets, which has much noisier measurements with varying sensing rates than the "calibrated" ones. Hence, the raw datasets are more challenging compared to the calibrated datasets. We re-calibrated the camera intrinsics and IMU-camera spacial-temporal parameters using the raw calibration datasets. Note that we set the initial values for $\mathbf{D}_a$, $\mathbf{D}_w$, ${}^{I}_{a}\mathbf{R}$ and ${}^{I}_{w}\mathbf{R}$ as identity and $\mathbf{T}_g$ as zeros, while the initial readout time for the whole RS image is set to 20 ms as prior calibration. All IMU intrinsic models listed in Table I were run with and without RS calibration. The results are presented in the following sections.

TABLE V: Averaged ATE of five runs over all eight sequences of the TUM RS VIO datasets with full-parameter calibration.

ATE	imu0	imu1	imu2	imu3	imu4
Ori. (deg) Pos. (m)	72.994 363.610	2.574 0.092	2.679 0.094	2.590 0.093	2.205 0.076
ATE	imu5	imu11	imu12	imu13	imu14
Ori. (deg) Pos. (m)	3.418 0.149	2.422 0.074	2.778 0.098	2.510 0.075	2.524 0.084

A. RS Self-Calibration

The results are shown in Fig. 10 and 11, with and without RS readout calibration, respectively. It is clear that the systems without RS readout time calibration and without IMU intrinsic calibration (imu0{no intrinsics}, imu31$\{\mathbf{D}_{a9}\}$ - imu34$\{\mathbf{T}_{g9}\}$) are unstable and diverge with large pose errors. With IMU intrinsic calibration (imu1$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{w}\mathbf{R}\}$ - imu24$\{\mathbf{D}_{w6}, \mathbf{D}_{a9}, \mathbf{T}_{g9}\}$) but without RS calibration, the system still fails for certain datasets, while online readout time calibration will greatly improve the system robustness for RS cameras. The final estimated RS readout time for each image is around 30 ms, which means given the image resolution of 1280×1024, the row readout time should be

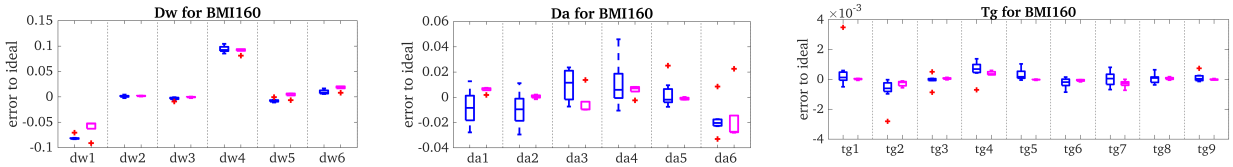


Fig. 9: IMU intrinsic evaluation of Bosch BMI160 IMU used in TUM RS VIO datasets using the proposed method (blue, left) and Kalibr (magenta, right) relative to the "ideal" sensor intrinsics. Red + denotes outliers.

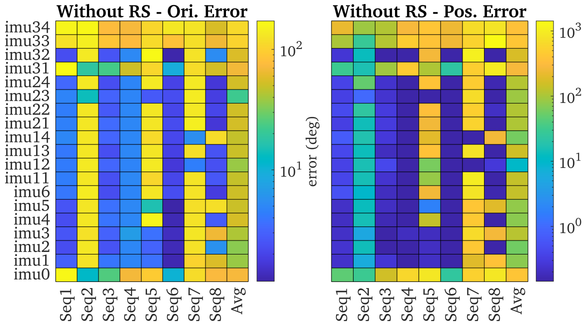


Fig. 10: Results on TUM RS VIO dataset *without* readout time calibration, with different IMU intrinsic models. The averaged ATE of five runs in degree (top) and meters (bottom) are provided. Note that the camera intrinsics, and IMU-camera spatial-temporal parameters are calibrated.

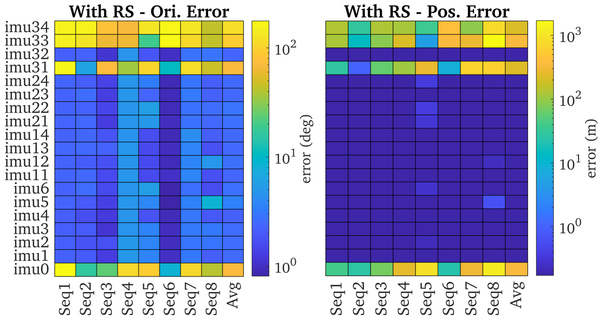


Fig. 11: Results on TUM RS VIO dataset *with* readout time calibration, with different IMU intrinsic models. The averaged ATE of five runs in degree (top) and meters (bottom) are provided. Note that camera intrinsics, and IMU-camera spatial-temporal parameters are calibrated.

around 29 µs, which matches with values provided by the camera manufacturer [26].

B. IMU Intrinsic Self-Calibration

We focus on the results in Fig. 11 which has RS enabled. It is clear from the performance of imu0{no intrinsics} that the BMI160 IMU will cause large trajectory errors without IMU intrinsic calibration, while performing intrinsic calibration will achieve accuracy more than an order of magnitude. Table V shows the average error over all sequences for 10 IMU models. It can be seen that the imu5$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a6}$, ${}^{I}_{w}\mathbf{R}$, ${}^{I}_{a}\mathbf{R}\}$ model which over parameterizes the intrinsics has worst accuracy in both orientation and position trajectory estimates, while the accuracy of the other models (imu1$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a6}$, ${}^{I}_{w}\mathbf{R}\}$ - imu4$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a9}\}$, imu11$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a6}$, ${}^{I}_{w}\mathbf{R}$, $\mathbf{T}_{g6}\}$ - imu14$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a9}$, $\mathbf{T}_{g6}\}$)) is comparable to each other (similar accuracy level). We further do an ablation study with models imu31$\{\mathbf{D}_{a9}\}$ - imu34$\{\mathbf{T}_{g9}\}$ to find the individual impact of each of the IMU intrinsic parameters. We can see that the imu32$\{\mathbf{D}_{w9}\}$ model, which estimates $\mathbf{D}_{w9}$, has large accuracy gains over the other three. This indicates that the readings from gyroscope of BMI160 are very noisy. The calibration of $\mathbf{D}_{w9}$ dominates the performance of this VINS system. Through these results, we show that online IMU intrinsic calibration can enhance both the system robustness and accuracy.

C. Comparison to Kalibr Calibration

We run Kalibr's offline calibration with *scale-misalignment*[3] IMU model on five calibration datasets (see [26], [48]) for the Bosch BMI160 IMU and treat these results as reference values when evaluating the proposed *online* calibration system. The Kalibr calibration datasets were collected with the stereo camera pair both operating in the global shutter mode along with an AprilTag board [27]. By contrast, the proposed system is run with only one camera of the above stereo pair — the right camera, which is set to rolling shutter mode — without AprilTags on the 8 data sequences but with the same IMU sensor [26]. Note that imu6$\{\mathbf{D}'_{w6}$, $\mathbf{D}'_{a6}$, ${}^{I}_{w}\mathbf{R}$, $\mathbf{T}_{g9}\}$, which is equivalent to the *scale-misalignment* IMU model of Kalibr, is used. For the evaluation, we directly report the estimation errors of $\mathbf{D}'_w = (\mathbf{T}'_w)^{-1}$, $\mathbf{D}'_a = (\mathbf{T}'_a)^{-1}$ and ${}^{w}_{I}\mathbf{R} = {}^{I}_{w}\mathbf{R}^{\top}$ for the Kalibr and our proposed system.

As shown in Fig. 9, even though we run on more challenging datasets, our proposed system can still achieve reasonable calibration results for $\mathbf{D}'_w$, $\mathbf{D}'_a$ and $\mathbf{T}_g$. The values of g-sensitivity $\mathbf{T}_g$ of the BMI160 IMU are generally one or two orders smaller than the other IMU intrinsics in magnitude. This matches the results presented in Fig. 11, for which the estimation errors of imu1$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a6}$, ${}^{I}_{w}\mathbf{R}\}$ - imu4$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a9}\}$ (without g-sensitivity) are similar to those of imu11$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a6}$, ${}^{I}_{w}\mathbf{R}$, $\mathbf{T}_{g6}\}$ - imu14$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a9}$, $\mathbf{T}_{g6}\}$) (with 6-DoF g-sensitivity) and imu21$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a6}$, ${}^{I}_{w}\mathbf{R}$, $\mathbf{T}_{g9}\}$ - imu24$\{\mathbf{D}_{w6}$, $\mathbf{D}_{a9}$, $\mathbf{T}_{g9}\}$ (with 9-DoF g-sensitivity). This means that the proposed system performance is less sensitive to g-sensitivity, no matter 0, 6 or 9 parameterization.

The values of scale-misalignment for gyroscope $\mathbf{D}'_w$ are much larger than those of $\mathbf{D}'_a$ and $\mathbf{T}_g$. This again confirms that the calibration of $\mathbf{D}'_w$ dominates the performance.

X. Real-World Experimental Validation on Self-Made VI-Rig

The proposed self-calibration system is further evaluated on a self-made visual-inertial sensor rig (VI-Rig, shown in Fig. 12), which contains multiple IMU and camera sensors.

[3] https://github.com/ethz-asl/kalibr/wiki/Multi-IMU-and-IMU-intrinsic-calibration

Specifically, it contains a MS-GX5-25, MS-GX5-35, Xsens MTi 100, FLIR blackfly camera and RealSense T265 tracking camera which contains an integrated BMI055 IMU along with a fisheye stereo camera. All cameras are not rolling shutter to ensure fair comparison against the baseline Kalibr [27] which only supports IMU-camera calibration with global shutter cameras. A total of 10 datasets were collected with an AprilTag board, on which both the proposed system and the Kalibr calibration toolbox were run to report repeatability statistics and evaluate real-world performances of both systems. During data collection, all 6-axis motion of VI-Rig were excited to avoid degenerate motions for calibration parameters.

To provide a fair comparison, we modified the front-end of the proposed system to directly and only use the same AprilTag detection as Kalibr. Additionally, while the proposed system was only run with one of the four IMUs and either the Blackfly or left T265 Realsense camera, Kalibr used all the available sensors to ensure the highest and most consistent performance (4 IMUs and 3 cameras). The $\texttt{imu6}\{\mathbf{D}'_{w6}, \mathbf{D}'_{a6}, {}^I_w\mathbf{R}, \mathbf{T}_{g9}\}$ model is used during evaluation, which is equivalent to the *scale-misalignment* IMU model of Kalibr. We define the "ideal" IMU sensor intrinsics as $\mathbf{D}'_w = \mathbf{D}'_a = \mathbf{I}_3$, ${}^I_w\mathbf{R} = {}^I_a\mathbf{R} = \mathbf{I}_3$ and $\mathbf{T}_g = \mathbf{0}_3$ if factory or offline calibration has been pre-applied. Generally, these values are what the users expect for a ready-to-use IMU, and are the initial values that the proposed estimator starts from. The quality of each IMU can be evaluated by how close the converged calibrated values are to these "ideal" values.

A. IMU-Camera Spatiotemporal Extrinsics and Intrinsics

The convergence of camera related parameters are investigated. The results shown in Fig. 13 demonstrate that the proposed system is able to calibrate the spatial-temporal parameters with both high repeatability and accuracy relative to the offline Kalibr calibration baseline. Additionally shown is the convergence of camera intrinsics estimated by the proposed algorithm relative to the Kalibr static calibration results which are fixed during their IMU-camera calibration.

B. IMU Intrinsic Parameters

As shown in Fig. 14, the average calibration errors of the proposed system are quite close to the results of Kalibr, and the proposed system demonstrates better repeatability than Kalibr, as our calibration errors have smaller variances and less outliers. In general, concerning the IMUs presented throughout the paper (see Fig. 9 and 14), we have:

- The MS-GX5-25, MS-GX5-35 and Xsens MTi-100 IMU are more close to "ideal" IMU than T265 IMU and BMI160 IMU. This is reasonable since both the MicroStrain and Xsens IMU are more expensive high-end IMUs with likely more sophisticate factory calibration.
- For each IMU, the g-sensitivity terms are, in general, much smaller than the other terms of the IMU intrinsic model. This suggests that the g-sensitivity should not have significant effects on system performance. This is likely due to the levels of achievable acceleration magnitudes in hand-held motions.
- The BMI160 IMU (Fig. 9), has a much more significant gyroscope calibration, $\mathbf{D}'_w$, compared to its accelerometer calibration and other IMUs. Thus the BMI160 can see large accuracy gains from only calibrating $\mathbf{D}'_w$, while for other IMUs, the calibration of $\mathbf{D}'_a$ should be more impactful.

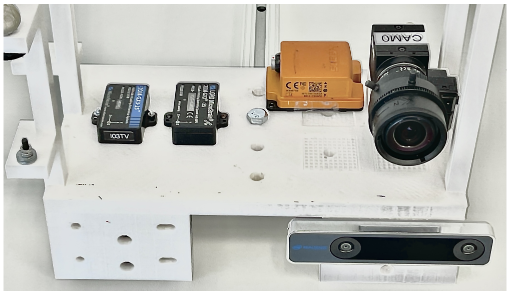


Fig. 12: Visual-Inertial Sensor Rig contains a MS-GX5-25 IMU, MS-GX5-35, Xsens MTi 100, FLIR Blackfly camera and RealSense T265 tracking camera (containing an integrated IMU and a fisheye stereo camera).

C. Timing Evaluation

The measurements from MS-GX5-25 IMU and the left camera of T265 are run for timing evaluation on the 10 recorded datasets. In order to get more realistic timing evaluation, no AprilTags are detected and only the natural features tracked from images are used. The average execution time of the proposed system with online calibration is $22.4\,\mathrm{ms}$ per frame, which shows relatively small increases than $18.8\,\mathrm{ms}$, which is the average running time without online calibration.

XI. Real-World Experimental Validation of Degenerate Motion Demonstration and Analysis

The proposed system is also evaluated on a collection of real-world datasets which exhibit varying degrees of degenerate motions. The g-sensitivity is not estimated since it has been shown that it is not significant for VINS performance. Note again that more real-world results can be found in our companion technical report [44].

A. EuRoC MAV: Under-Actuated Motion

The EuRoC MAV dataset [2] contains a series of trajectories from a MAV and provides $20\,\mathrm{Hz}$ grayscale stereo images, $200\,\mathrm{Hz}$ inertial readings, and an external groundtruth pose from a motion capture system. The proposed estimator is run with just the left camera on each of the Vicon room datasets and report the results in Table VI. It can be seen that $\texttt{imu0}\{$no intrinsics$\}$ model, for which IMU intrinsics is not calibrated, outperforms the methods which additionally estimate the IMU intrinsics. This makes sense since the IMU intrinsics suffer from a large number of degenerate motions which can be expected for the under-actuated MAV platform. Additionally, we believe that this is specifically caused by the MAV being unable to fully excite its 6-DoF motion for a given small time interval and thus undergoes (nearly) degenerate motions locally throughout the whole trajectory, hurting the sliding-window filter.

In order to verify the above reasoning, we use the groundtruth trajectories of *EuRoc V1_02* and *tum_room1* (with

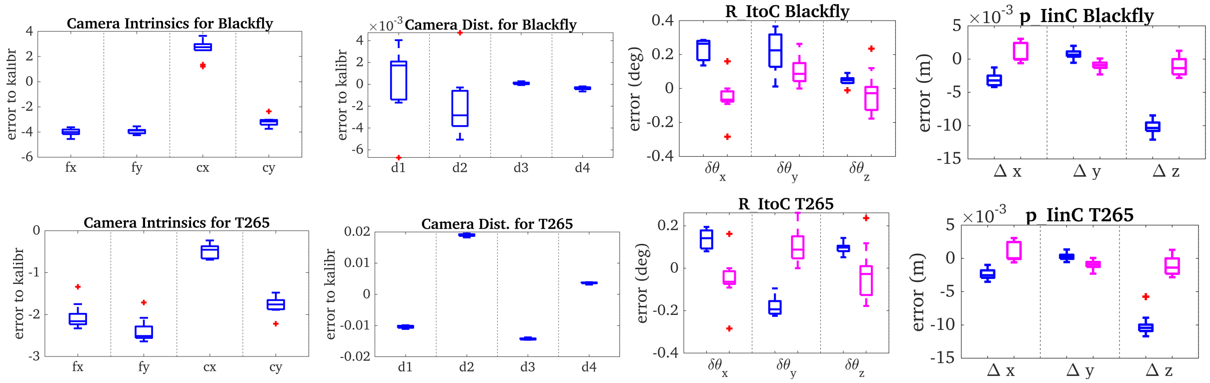


Fig. 13: Evaluation of Kalibr and the proposed methods, while for camera intrinsics only the proposed is reported since Kalibr fixes this during optimization. Kalibr (magenta, right in each group) was run with all cameras and IMUs available over 10 datasets, while the proposed system (blue, left) was run with either the Blackfly camera or left T265 fisheye and the MS-GX5-25 IMU resulting in 10 runs for each. Red + denotes outliers.

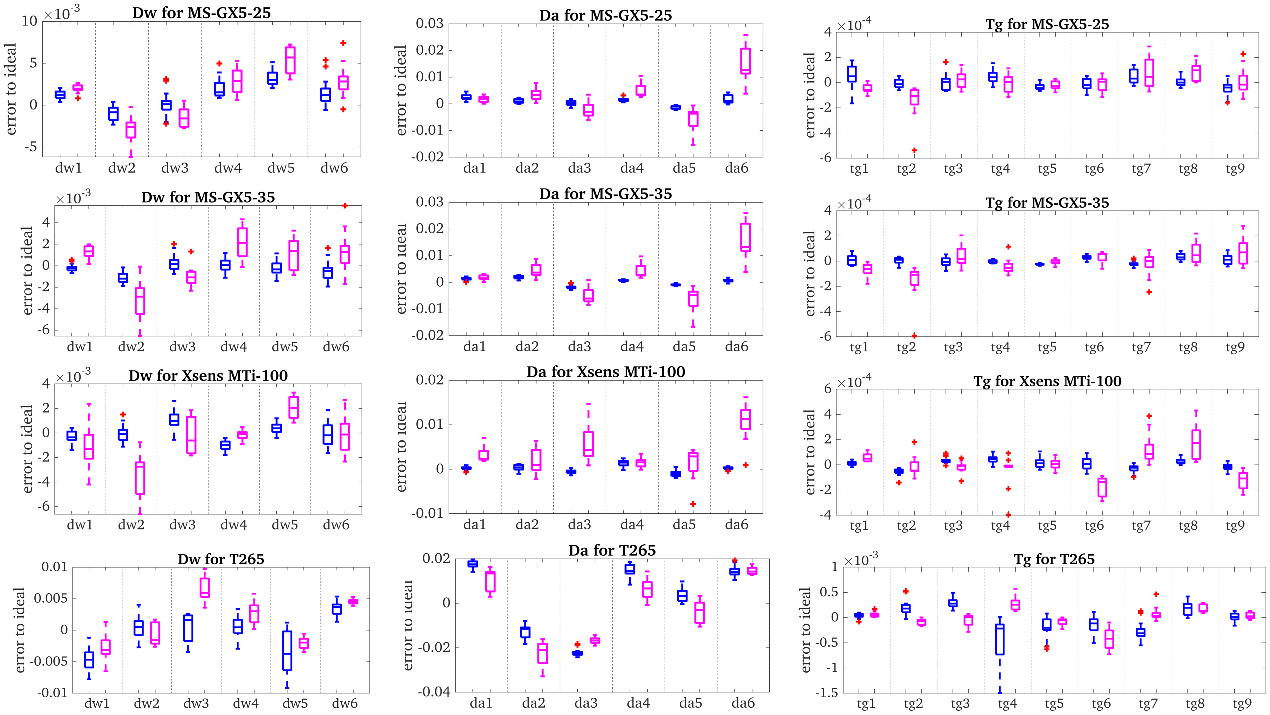


Fig. 14: Comparison of the proposed method with imu6$\{\mathbf{D}'_{w6}, \mathbf{D}'_{a6}, {}^{I}_{w}\mathbf{R}, \mathbf{T}_{g9}\}$ and Kalibr relative to the "ideal" sensor intrinsics. Kalibr (magenta, right in each group) was run with all cameras and IMUs available over 10 datasets, while the proposed system (blue, left) was run with either the Blackfly camera or left T265 fisheye resulting in 20 runs. Red + denotes outliers.

fully-excited motions as comparison) to *simulate* synthetic inertial and visual feature measurements (see Section VIII) and evaluate our system with these simulated data. Fig. 15 shows four different runs with estimation errors and 3σ bounds for $\mathbf{D}_a$. It is clear that the motion of sensor on the *EuRoc V1_02* trajectory (right) is mildly excited within local window, causing poor convergence of the $\mathbf{D}_a$ with relatively slower convergence of 3σ bounds as compared to the *tum_room1* (left). This verifies that the online IMU intrinsic calibration will benefit VINS with fully-excited motion (e.g., the *tum_room1* trajectory) and might not be a good option for under-actuated motions such as the *EuRoc V1_02* trajectory.

B. VI-Rig Planar Motion Datasets

We also evaluate on 4 datasets collected with VI-Rig (shown in Fig. 12) under planar motion. In this evaluation, MS-GX5-25 and the left camera of T265 are used. When collecting data, we put the VI-Rig on a chair and moved ensuring that the VI-Rig is performing planar motion with global yaw as rotation axis, which is also the y-axis (pointing downward) of the camera. We calibrate all parameters using imu2$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}\}$ and *equidist* when running the system. Since T265 is a global shutter camera, the readout time is zero.

The calibration results for the translation parameters ${}^{C}\mathbf{p}_{I}$, time offset t_d, readout time t_r and the $\mathbf{D}_w$ are shown in Fig. 16. All the temporal calibration can converge well to the

TABLE VI: Absolute Trajectory Error (ATE) on EuRoC MAV Vicon room sequences (with units degrees/meters).

IMU Model	V1_01_easy	V1_02_medium	V1_03_difficult	V2_01_easy	V2_02_medium	V2_03_difficult	Average
imu0{no intrinsics}	0.657 / 0.043	1.805 / 0.060	2.437 / 0.069	0.869 / 0.109	1.373 / 0.080	1.277 / 0.180	1.403 / 0.090
imu1$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{w}\mathbf{R}\}$	0.601 / 0.055	1.924 / 0.065	2.334 / 0.073	1.201 / 0.115	1.342 / 0.086	1.710 / 0.168	1.519 / 0.094
imu2$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}\}$	0.552 / 0.054	1.990 / 0.062	2.197 / 0.083	0.960 / 0.107	1.453 / 0.085	1.666 / 0.216	1.470 / 0.101
imu3$\{\mathbf{D}_{w9}, \mathbf{D}_{a6}\}$	0.606 / 0.055	1.905 / 0.065	2.359 / 0.073	1.180 / 0.114	1.335 / 0.088	1.640 / 0.167	1.504 / 0.094
imu4$\{\mathbf{D}_{w6}, \mathbf{D}_{a9}\}$	0.569 / 0.056	1.969 / 0.069	2.165 / 0.076	0.846 / 0.127	1.636 / 0.094	1.577 / 0.195	1.461 / 0.103

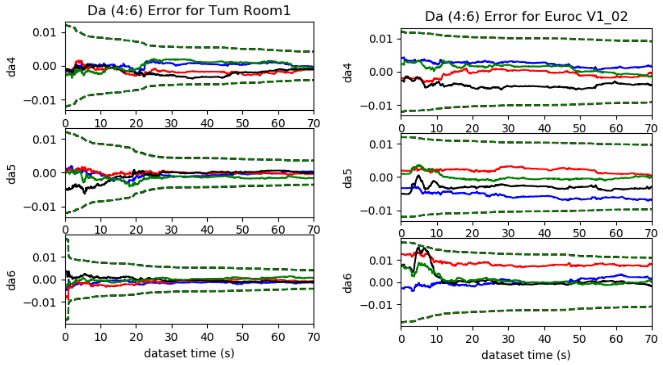


Fig. 15: Evaluation of the proposed system for $\mathbf{D}_a$ with *tum_room1* (left) and *EuRoc V1_02* (right) trajectories using imu2$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}\}$. 3σ (dotted lines) and estimation errors (solid lines) for four different runs (different colors) are drawn. The convergence of d_{a4}, d_{a5} and d_{a6} is poor for the *EuRoc V1_02* due to lack of motion excitation.

reference values based on offline calibration results of Kalibr. Note that t_r converges to almost zero as expected and t_d converges from $15\,\mathrm{ms}$ to $5\,\mathrm{ms}$ with reference values as $7\,\mathrm{ms}$. The final estimation errors are around $2\,\mathrm{ms}$, which is pretty small. While the x and z components of ${}^{C}\mathbf{p}_I$ can also converge well to the reference values with small standard deviations (smaller than $4\,\mathrm{mm}$), the y component diverges with estimation errors more than $5\,\mathrm{cm}$ and the standard deviation reach $3\,\mathrm{cm}$ since it is along the rotation axis of the camera and hence, unobservable. Since the system has only yaw rotation for the IMU sensor, the d_{w1}, d_{w2} and d_{w3} are also unobservable (see Table II), and their calibration results diverge a lot compared to those of d_{w4}, d_{w5} and d_{w6}. This result verifies our degenerate motion analysis for IMU-camera and IMU intrinsic calibration.

As a comparison, we also plot the online calibration results of the proposed system running on another four datasets from Section X with fully excited motions in Fig. 17. We use the same scale to plot the results for both Fig. 16 and 17. It is clear that all these calibration parameters (t_r, t_d, ${}^{C}\mathbf{p}_I$ and $\mathbf{D}_w$) can converge much better in fully excited motions than planar motion.

XII. Discussion and Recommendation

As learnt from the preceding extensive simulation analysis and real-world experimental validations, we generally recommend online self-calibration for VINS, especially in the following scenarios:

- Poor calibration priors are provided.
- Low-end IMUs or cameras are used.
- RS cameras are used.
- The sensor platform undergoes fully-excited motions.

Specifically, as shown in Fig. 10 and 11, if starting with imperfect calibration, the system without online self-calibration is highly likely to fail, as clearly demonstrated in Fig. 4. In comparison, performing online calibration can greatly improve the system robustness and accuracy.

If using high quality IMUs (e.g., ADIS16470) or well-calibrated IMUs (e.g., pre-calibrated IMU data from TUM VI dataset [48]), VINS performance gain might be marginal if performing online IMU calibration. However, it is evident from Fig. 11 that it is necessary to perform online calibration for the low-end IMU (e.g., BMI160) with uncalibrated raw data and RS readout time for improved accuracy and robustness. Note that OpenVINS [10][4] and VINS-Mono [13] assume good IMU intrinsic calibration and thus are unable to work well on the uncalibrated raw datasets. This has also motivated us to perform online self-calibration to lower the technological barriers of VINS.

Interestingly, based on the results from the EuRoC MAV dataset as shown in Table VI, online calibration, especially IMU intrinsic calibration, can hurt the system performance when the robot undergoes underactuated motions. As shown in our degenerate motion analysis, there are a large number of motion types that prohibit accurate calibration of the IMU intrinsics and IMU-camera spatial calibration, while the camera intrinsics and IMU-camera temporal calibration are more robust to different motions. More importantly, in the most commonly-seen motion cases of aerial and ground vehicles, there is usually at least one unobservable direction for calibration, due to these robots traveling with either underactuated 3D or planar motion.

Due to the high likelihood of experiencing degenerate motions for some periods of time, solely based on our analysis and results, we do *not* recommend performing online IMU intrinsic and IMU-camera spatial calibration during real-time operations for most underactuated motions (e.g., planar motion and one-axis rotation for most ground vehicles). The exception to this is the handheld cases (e.g., mobile AR/VR), which often exhibit full 6-DoF motions and thus is recommended to perform online calibration to improve estimation accuracy, especially when low-end IMUs or RS cameras are used. For these applications, we do recommend using an offline batch optimization to obtain an accurate initial calibration for the state estimator and/or keep the calibration parameters (especially intrinsics) fixed if one knows they are going to experience degenerate motions. For online IMU intrinsic calibration, it is not necessary to calibrate the full IMU model and instead one may calibrate only the dominating parameters in the inertial models, for example, $\mathbf{D}_w$ for BMI160 IMU or $\mathbf{D}_a$ for MicroStrain, Xsens and T265 IMUs.

[4] We have open sourced this work's support for IMU intrinsics as a part of OpenVINS [10].

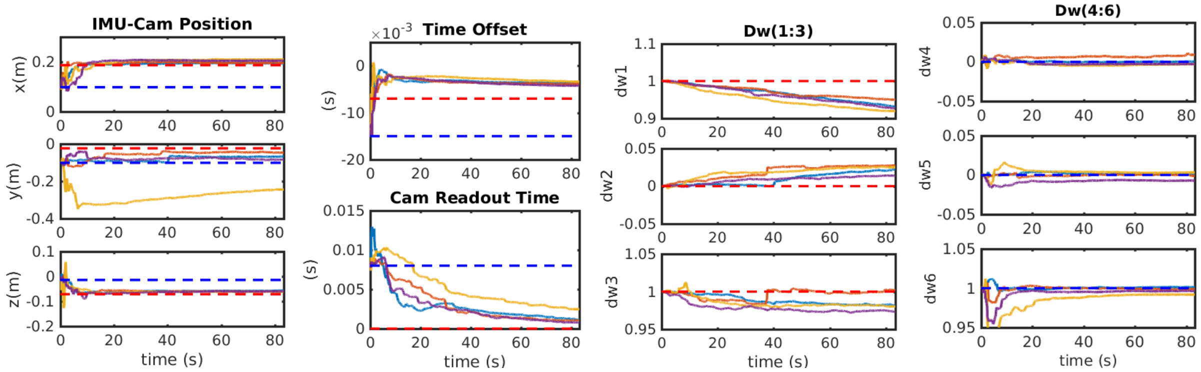


Fig. 16: Calibration results (four VI-Rig planar motion datasets with colored solid lines) for ${}^{C}\mathbf{p}_I$, t_d, t_r and $\mathbf{D}_w$ of the proposed system evaluated using imu2$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}\}$ and *equi-dist*. Red and blue dotted lines denote the reference value from Kalibr and initial (perturbed) values, respectively. The y component of ${}^{C}\mathbf{p}_I$, dw_1, dw_2 and dw_3 diverges.

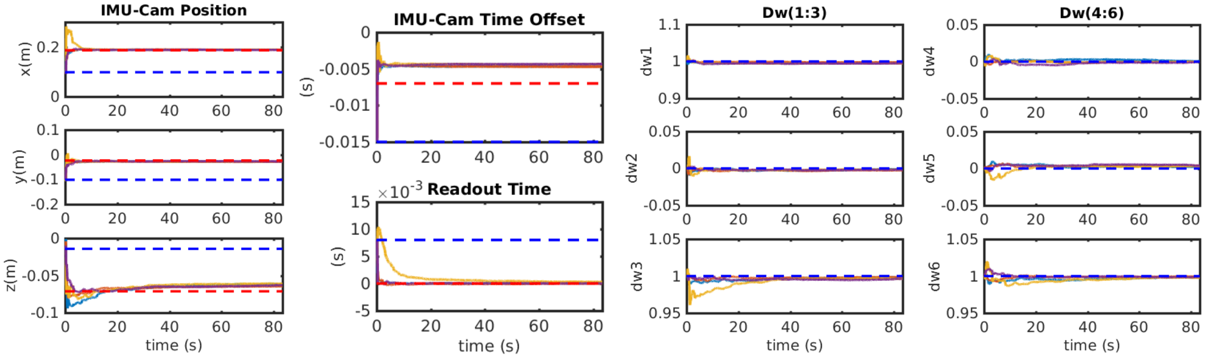


Fig. 17: Calibration results (from four 3D-motion datasets in Section X with colored solid lines) for ${}^{C}\mathbf{p}_I$, t_d, t_r and $\mathbf{D}_w$ of the proposed system evaluated using imu2$\{\mathbf{D}_{w6}, \mathbf{D}_{a6}, {}^{I}_{a}\mathbf{R}\}$ and *equidist*. Red and blue dotted lines denote the reference value from Kalibr and initial (perturbed) values, respectively.

XIII. Conclusions and Future Work

In this paper, we have comprehensively studied the problem of online full-parameter self-calibration for visual-inertial navigation in order to achieve accurate and robust estimation performance. We have first investigated different IMU intrinsic model variants which uses different parameterizations to account for scale correction, axis misalignment and g-sensitivity. These variants encompass commonly-used inertial models in practice. Along with the inertial intrinsics, we have examined the full visual measurement model that accounts for full IMU-camera spatial-temporal parameters including RS readout time. Based on these models, we have performed the observability analysis for linearized VINS with full self-calibration to show that it truly has only 4 unobservable directions corresponding to global yaw and global translation, while all the calibration parameters are observable given fully excited motions, thus reassuring the intuitions assumed in the literature. Moreover, we have for the first time identified the basic degenerate motion patterns for IMU/camera intrinsics, whose combination would still cause unobservable directions.

More importantly, we have developed the MSCKF-based VINS estimator with full self-calibration. With that, we have performed extensive simulation analysis and real-world experimental validations to verify our observability and degenerate motion analysis. Solely based on our analysis and validations, we have offered our self-calibration recommendations. While in general online self-calibration can improve the VINS robustness and accuracy, online IMU intrinsic calibration is risky due to its dependence on the motion profile to ensure observability. For example, in the case of autonomous (ground) vehicles, most trajectories have degenerate motions, thus we do *not* recommending online calibration of IMU intrinsics for under-actuated robots. By contrast, in the case of handheld motions, we found that the estimation of calibration parameters improved performance as expected.

In the future, we will investigate a complete degenerate motion analysis for multi-visual-inertial system along with robust estimation algorithms (e.g., Schmidt-KF [49]) to enable online calibration under degenerate motions.

Acknowledgement

Thanks Dr. Kun Fu for 3D printing of the VI-Rig. Thanks Dr. Kevin Eckenhoff for the rolling shutter implementation from the MIMC-VINS project [11]. Thanks David Schubert and Nikolaus Demmel from TUM for the helpful discussion on the TUM RS VIO datasets [26].

Appendix A
IMU Intrinsic Jacobians

The Jacobians for all the variables that might appear in the IMU models, including ${}^{I}_{w}\mathbf{R}$ and ${}^{I}_{a}\mathbf{R}$, will be derived. More derivations can be found in our companion technical report [44]. To simplify the derivations, we define ${}^{I}\hat{\mathbf{a}}$ and ${}^{I}\tilde{\mathbf{a}}$ as:

$$ {}^{I}\hat{\mathbf{a}} = {}^{I}_{a}\hat{\mathbf{R}}\hat{\mathbf{D}}_a \left({}^{a}\mathbf{a}_m - \hat{\mathbf{b}}_a\right) $$

$$ {}^{I}\tilde{\mathbf{a}} = {}^{I}_{a}\hat{\mathbf{R}}\mathbf{H}_{Da}\tilde{\mathbf{x}}_{Da} + \lfloor {}^{I}\hat{\mathbf{a}} \rfloor \delta\boldsymbol{\theta}_{Ia} - {}^{I}_{a}\hat{\mathbf{R}}\hat{\mathbf{D}}_a\tilde{\mathbf{b}}_a - {}^{I}_{a}\hat{\mathbf{R}}\hat{\mathbf{D}}_a\mathbf{n}_a $$

We define ${}^I\hat{\boldsymbol{\omega}}$ and ${}^I\tilde{\boldsymbol{\omega}}$ as:

$$
\begin{aligned}
{}^I\hat{\boldsymbol{\omega}} &= {}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\left({}^w\boldsymbol{\omega}_m - \hat{\mathbf{T}}_g{}^I\hat{\mathbf{a}} - \hat{\mathbf{b}}_g\right)\\
{}^I\tilde{\boldsymbol{\omega}} &= -{}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\tilde{\mathbf{b}}_g + {}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g{}^I_a\hat{\mathbf{R}}\hat{\mathbf{D}}_a\tilde{\mathbf{b}}_a\\
&\quad + {}^I_w\hat{\mathbf{R}}\mathbf{H}_{Dw}\tilde{\mathbf{x}}_{Dw} - {}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g{}^I_a\hat{\mathbf{R}}\mathbf{H}_{Da}\tilde{\mathbf{x}}_{Da}\\
&\quad + \lfloor{}^I\hat{\boldsymbol{\omega}}\rfloor\delta\boldsymbol{\theta}_{Iw} - {}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g\lfloor{}^I\hat{\mathbf{a}}\rfloor\delta\boldsymbol{\theta}_{Ia}\\
&\quad - {}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\mathbf{H}_{Tg}\tilde{\mathbf{x}}_{Tg} - {}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\mathbf{n}_g\\
&\quad + {}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g{}^I_a\hat{\mathbf{R}}\hat{\mathbf{D}}_a\mathbf{n}_a
\end{aligned}
$$

where we have:

$$\mathbf{H}_{Dw} = \begin{bmatrix}{}^w\hat{w}_1\mathbf{e}_1 & {}^w\hat{w}_2\mathbf{e}_1 & {}^w\hat{w}_2\mathbf{e}_2 & {}^w\hat{w}_3\mathbf{I}_3\end{bmatrix} \tag{82}$$

$$\mathbf{H}_{Da} = \begin{bmatrix}{}^a\hat{a}_1\mathbf{e}_1 & {}^a\hat{a}_2\mathbf{e}_1 & {}^a\hat{a}_2\mathbf{e}_2 & {}^a\hat{a}_3\mathbf{I}_3\end{bmatrix} \tag{83}$$

$$\mathbf{H}_{Tg} = \begin{bmatrix}{}^I\hat{a}_1\mathbf{I}_3 & {}^I\hat{a}_2\mathbf{I}_3 & {}^I\hat{a}_3\mathbf{I}_3\end{bmatrix} \tag{84}$$

By summarizing the above equations, we have:

$$\begin{bmatrix}{}^{I_k}\tilde{\boldsymbol{\omega}}\\ {}^{I_k}\tilde{\mathbf{a}}\end{bmatrix} = \begin{bmatrix}\mathbf{H}_b & \mathbf{H}_{in}\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{x}}_b\\ \tilde{\mathbf{x}}_{in}\end{bmatrix} + \mathbf{H}_n\begin{bmatrix}\mathbf{n}_g\\ \mathbf{n}_a\end{bmatrix} \tag{85}$$

where we have defined:

$$\mathbf{H}_b = \mathbf{H}_n = \begin{bmatrix}-{}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w & {}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g{}^I_a\hat{\mathbf{R}}\hat{\mathbf{D}}_a\\ \mathbf{0}_3 & -{}^I_a\hat{\mathbf{R}}\hat{\mathbf{D}}_a\end{bmatrix} \tag{86}$$

$$\mathbf{H}_{in} = \begin{bmatrix}\mathbf{H}_w & \mathbf{H}_a & \mathbf{H}_{Iw} & \mathbf{H}_{Ia} & \mathbf{H}_g\end{bmatrix} \tag{87}$$

$$\mathbf{H}_w = \begin{bmatrix}{}^I_w\hat{\mathbf{R}}\mathbf{H}_{Dw}\\ \mathbf{0}_3\end{bmatrix},\ \mathbf{H}_a = \begin{bmatrix}-{}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g{}^I_a\hat{\mathbf{R}}\mathbf{H}_{Da}\\ {}^I_a\hat{\mathbf{R}}\mathbf{H}_{Da}\end{bmatrix} \tag{88}$$

$$\mathbf{H}_{Iw} = \begin{bmatrix}\lfloor{}^I\hat{\boldsymbol{\omega}}\rfloor\\ \mathbf{0}_3\end{bmatrix},\ \mathbf{H}_{Ia} = \begin{bmatrix}-{}^I_a\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}\lfloor{}^I\hat{\mathbf{a}}\rfloor\\ \lfloor{}^I\hat{\mathbf{a}}\rfloor\end{bmatrix},\ \mathbf{H}_g = \begin{bmatrix}-{}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\mathbf{H}_{Tg}\\ \mathbf{0}_3\end{bmatrix} \tag{89}$$

$\mathbf{n}_{dI} = [\mathbf{n}_{dg}^\top\ \mathbf{n}_{da}^\top\ \mathbf{n}_{dwg}^\top\ \mathbf{n}_{dwa}^\top]^\top$ denotes the discretized IMU noises; $\mathbf{n}_{d*} \sim \mathcal{N}(\mathbf{0}, \frac{\sigma_*^2\mathbf{I}_3}{\delta t_k})$ and the covariance for $\mathbf{n}_{dI}$ can be written block-diagonal matrix as:

$$\mathbf{Q}_{dI} = \mathbf{diag}\{\frac{\sigma_g^2}{\delta t_k}\mathbf{I}_3, \frac{\sigma_a^2}{\delta t_k}\mathbf{I}_3, \frac{\sigma_{wg}^2}{\delta t_k}\mathbf{I}_3, \frac{\sigma_{wa}^2}{\delta t_k}\mathbf{I}_3\} \tag{90}$$

Appendix B
Camera Measurement Jacobians

The camera intrinsic Jacobians $\mathbf{H}_{Cin}$, $\frac{\partial\tilde{\mathbf{z}}_C}{\partial\tilde{\mathbf{z}}_n}$ and $\frac{\partial\tilde{\mathbf{z}}_n}{\partial{}^C\tilde{\mathbf{p}}_f}$ for $\mathbf{H}_{\mathbf{p}_f}$ [Eq. (55)] can be found in our companion technique report [44]. The Jacobians of ${}^C\mathbf{p}_f$ regarding to $\mathbf{x}_I$ are written as:

$$\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_I} = \begin{bmatrix}\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_n} & \frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_b} & \frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_{in}}\end{bmatrix} \tag{91}$$

$$\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_n} = {}^C_I\hat{\mathbf{R}}{}^I_G\hat{\mathbf{R}}\begin{bmatrix}\lfloor{}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_I\rfloor{}^G_I\hat{\mathbf{R}} & -\mathbf{I}_3 & \mathbf{0}_3\end{bmatrix} \tag{92}$$

$$\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_b} = \mathbf{0}_{3\times 6},\quad \frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_{in}} = \mathbf{0}_{3\times 24} \tag{93}$$

The Jacobians of ${}^C\mathbf{p}_f$ regarding to the IMU-camera spatial-temporal calibration state $\mathbf{x}_{IC}$ are written as:

$$\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_{IC}} = \begin{bmatrix}\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\delta\boldsymbol{\theta}_{IC}} & \frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial{}^C\tilde{\mathbf{p}}_I} & \frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{t}_d} & \frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{t}_r}\end{bmatrix} \tag{94}$$

$$\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\delta\boldsymbol{\theta}_{IC}} = \lfloor{}^C_I\hat{\mathbf{R}}{}^I_G\hat{\mathbf{R}}\left({}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_I\right)\rfloor \tag{95}$$

$$\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial{}^C\tilde{\mathbf{p}}_I} = \mathbf{I}_3,\quad \frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{t}_r} = \frac{m}{M}\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{t}_d} \tag{96}$$

$$\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{t}_d} = {}^C_I\hat{\mathbf{R}}{}^I_G\hat{\mathbf{R}}\left(\lfloor\left({}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_I\right)\rfloor{}^G_I\hat{\mathbf{R}}{}^I\hat{\boldsymbol{\omega}} - {}^G\hat{\mathbf{v}}_I\right) \tag{97}$$

Note that when computing the Jacobians for t_d and t_r, we are using the following linearization:

$${}^G_{I(t)}\mathbf{R} \simeq {}^G_{I(\hat{t})}\hat{\mathbf{R}}\exp(\delta\boldsymbol{\theta}_I)\exp({}^I\hat{\boldsymbol{\omega}}\tilde{t}_d + \frac{m}{M}{}^I\hat{\boldsymbol{\omega}}\tilde{t}_r) \tag{98}$$

$${}^G\mathbf{p}_{I(t)} \simeq {}^G\hat{\mathbf{p}}_{I(\hat{t})} + {}^G\tilde{\mathbf{p}}_I + {}^G\hat{\mathbf{v}}_I\tilde{t}_d + \frac{m}{M}{}^G\hat{\mathbf{v}}_I\tilde{t}_r \tag{99}$$

The Jacobians of ${}^C\mathbf{p}_f$ regarding to $\mathbf{x}_f$ is written as:

$$\frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\tilde{\mathbf{x}}_f} = \frac{\partial{}^C\tilde{\mathbf{p}}_f}{\partial\delta{}^G\tilde{\mathbf{p}}_f} = {}^C_I\hat{\mathbf{R}}{}^I_G\hat{\mathbf{R}} \tag{100}$$

Appendix C
Observability Matrix

The $\mathbf{M}_n$ is computed as:

$$
\begin{aligned}
\mathbf{M}_n &= \mathbf{H}_{\mathbf{p}_f}{}^C_I\hat{\mathbf{R}}{}^{I_k}_G\hat{\mathbf{R}}\begin{bmatrix}\boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_2 & \boldsymbol{\Gamma}_3\end{bmatrix} \qquad (101)\\
\boldsymbol{\Gamma}_1 &= \lfloor{}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_{I_1} - {}^G\hat{\mathbf{v}}_{I_1}\delta t_k + \frac{1}{2}{}^G\mathbf{g}\delta t_k^2\rfloor{}^G_{I_1}\hat{\mathbf{R}}\\
\boldsymbol{\Gamma}_2 &= -\mathbf{I}_3\ ,\boldsymbol{\Gamma}_3 = -\mathbf{I}_3\delta t_k
\end{aligned}
$$

The $\mathbf{M}_b$ is computed as:

$$
\begin{aligned}
\mathbf{M}_b &= \mathbf{H}_{\mathbf{p}_f}{}^C_I\hat{\mathbf{R}}{}^{I_k}_G\hat{\mathbf{R}}\begin{bmatrix}\boldsymbol{\Gamma}_4 & \boldsymbol{\Gamma}_5\end{bmatrix} \qquad (102)\\
\boldsymbol{\Gamma}_4 &= -\left(\lfloor{}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_{I_k}\rfloor{}^G_{I_k}\hat{\mathbf{R}}\mathbf{J}_r\left(\Delta\hat{\boldsymbol{\theta}}_k\right)\delta t_k + {}^G_{I_k}\hat{\mathbf{R}}\boldsymbol{\Xi}_4\right){}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\\
\boldsymbol{\Gamma}_5 &= \left(\lfloor{}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_{I_k}\rfloor{}^G_{I_k}\hat{\mathbf{R}}\mathbf{J}_r\left(\Delta\hat{\boldsymbol{\theta}}_k\right){}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g\delta t_k\right.\\
&\qquad \left. + {}^G_{I_k}\hat{\mathbf{R}}\left(\boldsymbol{\Xi}_4{}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g + \boldsymbol{\Xi}_2\right)\right){}^I_a\hat{\mathbf{R}}\hat{\mathbf{D}}_a
\end{aligned}
$$

The $\mathbf{M}_{in}$ can be computed as:

$$
\begin{aligned}
\mathbf{M}_{in} &= \mathbf{H}_{\mathbf{p}_f}{}^C_I\hat{\mathbf{R}}{}^{I_k}_G\hat{\mathbf{R}}\begin{bmatrix}\boldsymbol{\Gamma}_6 & \boldsymbol{\Gamma}_7 & \boldsymbol{\Gamma}_8 & \boldsymbol{\Gamma}_9\end{bmatrix} \qquad (103)\\
\boldsymbol{\Gamma}_6 &= \left(\lfloor{}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_{I_k}\rfloor{}^G_{I_k}\hat{\mathbf{R}}\mathbf{J}_r\left(\Delta\hat{\boldsymbol{\theta}}_k\right)\delta t_k + {}^G_{I_k}\hat{\mathbf{R}}\boldsymbol{\Xi}_4\right)\mathbf{H}_{Dw}\\
\boldsymbol{\Gamma}_7 &= -\left(\lfloor{}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_{I_k}\rfloor{}^G_{I_k}\hat{\mathbf{R}}\mathbf{J}_r\left(\Delta\hat{\boldsymbol{\theta}}_k\right){}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g\delta t_k\right.\\
&\qquad \left. + {}^G_{I_k}\hat{\mathbf{R}}\left(\boldsymbol{\Xi}_4{}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g + \boldsymbol{\Xi}_2\right)\right){}^I_a\hat{\mathbf{R}}\mathbf{H}_{Da}\\
\boldsymbol{\Gamma}_8 &= -\left(\lfloor{}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_{I_k}\rfloor{}^G_{I_k}\hat{\mathbf{R}}\mathbf{J}_r\left(\Delta\hat{\boldsymbol{\theta}}_k\right){}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g\delta t_k\right.\\
&\qquad \left. + {}^G_{I_k}\hat{\mathbf{R}}(\boldsymbol{\Xi}_4{}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\hat{\mathbf{T}}_g + \boldsymbol{\Xi}_2)\right)\lfloor{}^{I_k}\hat{\mathbf{a}}\rfloor\\
\boldsymbol{\Gamma}_9 &= -\left(\lfloor{}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_{I_k}\rfloor{}^G_{I_k}\hat{\mathbf{R}}\mathbf{J}_r\left(\Delta\hat{\boldsymbol{\theta}}_k\right)\delta t_k + {}^G_{I_k}\hat{\mathbf{R}}\boldsymbol{\Xi}_4\right){}^I_w\hat{\mathbf{R}}\hat{\mathbf{D}}_w\mathbf{H}_{Tg}
\end{aligned}
$$

The $\mathbf{M}_{IC}$ can be computed as:

$$
\begin{aligned}
\mathbf{M}_{IC} &= \mathbf{H}_{\mathbf{p}_f}{}^C_I\hat{\mathbf{R}}{}^{I_k}_G\hat{\mathbf{R}}\begin{bmatrix}\boldsymbol{\Gamma}_{10} & \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} & \boldsymbol{\Gamma}_{13}\end{bmatrix} \qquad (104)\\
\boldsymbol{\Gamma}_{10} &= \lfloor\left({}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_{I_k}\right)\rfloor{}^G_{I_k}\hat{\mathbf{R}}{}^I_C\hat{\mathbf{R}}\\
\boldsymbol{\Gamma}_{11} &= {}^G_{I_k}\hat{\mathbf{R}}{}^I_C\hat{\mathbf{R}},\ \boldsymbol{\Gamma}_{13} = \frac{m}{M}\boldsymbol{\Gamma}_{12}\\
\boldsymbol{\Gamma}_{12} &= \lfloor\left({}^G\hat{\mathbf{p}}_f - {}^G\hat{\mathbf{p}}_{I_k}\right)\rfloor{}^G_{I_k}\hat{\mathbf{R}}{}^{I_k}\hat{\boldsymbol{\omega}} - {}^G\hat{\mathbf{v}}_{I_k}
\end{aligned}
$$

The $\mathbf{M}_{Cin}$ and $\mathbf{M}_f$ can be written as:

$$\mathbf{M}_{Cin} = \mathbf{H}_{Cin},\ \mathbf{M}_f = \mathbf{H}_{\mathbf{p}_f}{}^C_I\hat{\mathbf{R}}{}^{I_k}_G\hat{\mathbf{R}} \tag{105}$$

Appendix D
Proof of Lemma 1

For Eq. (58), we first verify $\mathcal{O}\mathbf{N} = \mathbf{0}$ as:

$$\left(\boldsymbol{\Gamma}_1 {}^{I_1}_{G}\hat{\mathbf{R}} - \boldsymbol{\Gamma}_2\lfloor {}^G\hat{\mathbf{p}}_{I_1}\rfloor - \boldsymbol{\Gamma}_3\lfloor {}^G\hat{\mathbf{v}}_{I_1}\rfloor - \lfloor {}^G\hat{\mathbf{p}}_f\rfloor\right){}^G\mathbf{g} = 0$$

Hence, $\mathcal{O}$ has at least 4 unobservable directions (4-DoF).

In the following, we will try to show that there are only 4 unobservable directions under general situations. With abusing of notion, we rewrite the observability matrix by segmenting the columns as:

$$\begin{aligned}\mathcal{O} &\triangleq \left[\begin{array}{c|c|c|c} \mathcal{O}_I & \mathcal{O}_{in} & \mathcal{O}_{IC} & \mathcal{O}_f \end{array}\right] \\ &\triangleq \left[\begin{array}{ccc|c|c|c} \mathbf{M}_{n,1} & \mathbf{M}_{b,1} & \mathbf{M}_{f,1} & \mathbf{M}_{in,1} & \mathbf{M}_{IC,1} & \mathbf{M}_{Cin,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{M}_{n,k} & \mathbf{M}_{b,k} & \mathbf{M}_{f,k} & \mathbf{M}_{in,k} & \mathbf{M}_{IC,k} & \mathbf{M}_{Cin,k} \end{array}\right]\end{aligned} \tag{106}$$

$\mathcal{O}_I$ corresponds to the IMU navigation state, IMU bias state and feature state. $\mathcal{O}_I$ is equivalent to the standard VINS observability matrix in [37] and it has a 4-DoF null space.

$\mathcal{O}_{in}$ corresponds to the IMU intrinsic parameters. By checking the Eq. (103), it is clearly that $\mathcal{O}_{in}$ will be affected by time-varying ${}^w\boldsymbol{\omega}(t)$ (in $\mathbf{H}_{Dw}$), ${}^a\mathbf{a}(t)$ (in $\mathbf{H}_{Da}$) and ${}^I\mathbf{a}(t)$ (in $\lfloor {}^I\mathbf{a}\rfloor$ and $\mathbf{H}_{Tg}$). Under fully-excited motions, $\mathcal{O}_{in}$ can be of full column rank.

$\mathcal{O}_{IC}$ corresponds to the IMU-camera spatial and temporal calibration parameters. By checking the Eq. (104), we can see that the $\mathcal{O}_{IC}$ is affected by the time-varying IMU pose $\{{}^I_G\mathbf{R}(t), {}^G\mathbf{p}_I(t)\}$ and the IMU kinematics $\{{}^I\boldsymbol{\omega}(t), {}^I\mathbf{v}(t)\}$. In addition, $\boldsymbol{\Gamma}_{13}$ in $\mathbf{M}_{IC}$ are also affected by the point feature measurement through $\frac{m}{M}$, of which m will change under general measurement assumptions. Hence, $\mathcal{O}_{IC}$ can be of full column rank with random motions.

$\mathcal{O}_{Cin}$ corresponds to the camera intrinsic parameters. It is clear that $\mathcal{O}_{Cin}$ is only affected the environmental structure and is of full column rank as long as $\{u_n, v_n\}$ varies in different image tracks.

Since $\mathcal{O}_{in}$, $\mathcal{O}_{IC}$ and $\mathcal{O}_{Cin}$ are affected by different system parameters, and under general motion conditions, $[\mathcal{O}_{in} \;\; \mathcal{O}_{IC} \;\; \mathcal{O}_{Cin}]$ is also of full column rank. Therefore, the column rank of $\mathcal{O}$ is determined by $\mathcal{O}_I$. Since $\mathcal{O}_I$ has a 4-DoF null space, the $\mathcal{O}$ also has 4-DoF. We also verify this conclusion through simulation in Fig. 3.

References

[1] G. Huang, "Visual-inertial navigation: A concise review," in *Proc. International Conference on Robotics and Automation*, Montreal, Canada, May 2019.

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

[3] W. Lee, P. Geneva, Y. Yang, and G. Huang, "Tightly-coupled gnss-aided visual-inertial localization," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 9484–9491.

[4] M. Zhang, X. Zuo, Y. Chen, Y. Liu, and M. Li, "Pose estimation for ground robots: On manifold representation, integration, reparameterization, and optimization," *IEEE Transactions on Robotics*, vol. 37, no. 4, pp. 1081–1099, 2021.

[5] W. Lee, K. Eckenhoff, Y. Yang, P. Geneva, and G. Huang, "Visual-inertial-wheel odometry with online calibration," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[6] M. Li and A. Mourikis, "High-precision, consistent EKF-based visual-inertial odometry," *International Journal of Robotics Research*, vol. 32, no. 6, pp. 690–711, 2013.

[7] C. Guo, D. Kottas, R. DuToit, A. Ahmed, R. Li, and S. Roumeliotis, "Efficient visual-inertial navigation using a rolling-shutter camera with inaccurate timestamps," in *Proc. of the Robotics: Science and Systems Conference*, Berkeley, CA, Jul. 13–17, 2014.

[8] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Rome, Italy, Apr. 10–14, 2007, pp. 3565–3572.

[9] K. J. Wu, A. M. Ahmed, G. A. Georgiou, and S. I. Roumeliotis, "A square root inverse filter for efficient vision-aided inertial navigation on mobile devices," in *Robotics: Science and Systems Conference (RSS)*, 2015.

[10] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A research platform for visual-inertial estimation," in *Proc. of the IEEE International Conference on Robotics and Automation*, Paris, France, 2020. [Online]. Available: https://github.com/rpng/open_vins

[11] K. Eckenhoff, P. Geneva, and G. Huang, "Mimc-vins: A versatile and resilient multi-imu multi-camera visual-inertial navigation system," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1360–1380, 2021.

[12] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 1–21, Feb. 2017.

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

[14] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, pp. 1–17, 2021.

[15] M. Li, H. Yu, X. Zheng, and A. I. Mourikis, "High-fidelity sensor modeling and self-calibration in vision-aided inertial navigation," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 409–416.

[16] Y. Yang, P. Geneva, K. Eckenhoff, and G. Huang, "Degenerate motion analysis for aided INS with online spatial and temporal calibration," *IEEE Robotics and Automation Letters (RA-L)*, vol. 4, no. 2, pp. 2070–2077, 2019.

[17] Y. Yang, P. Geneva, X. Zuo, and G. Huang, "Online imu intrinsic calibration: Is it necessary?" in *Proc. of Robotics: Science and Systems (RSS)*, Corvallis, Or, 2020.

[18] X. Lang, J. Lv, J. Huang, Y. Ma, Y. Liu, and X. Zuo, "Ctrl-vio: Continuous-time visual-inertial odometry for rolling shutter cameras," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 537–11 544, 2022.

[19] F. M. Mirzaei and S. I. Roumeliotis, "A Kalman filter-based algorithm for IMU-camera calibration: Observability analysis and performance evaluation," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1143–1156, Oct. 2008.

[20] J. Kelly and G. S. Sukhatme, "Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration," *International Journal of Robotics Research*, vol. 30, no. 1, pp. 56–79, Jan. 2011.

[21] C. Guo and S. Roumeliotis, "IMU-RGBD camera 3D pose estimation and extrinsic calibration: Observability analysis and consistency improvement," in *Proc. of the IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, May 6–10, 2013.

[22] S.-H. Tsao and S.-S. Jan, "Observability analysis and performance evaluation of ekf-based visual-inertial odometry with online intrinsic camera parameter calibration," *IEEE Sensors Journal*, vol. 19, no. 7, pp. 2695–2703, 2019.

[23] K. Hausman, J. Preiss, G. S. Sukhatme, and S. Weiss, "Observability-aware trajectory optimization for self-calibration with application to uavs," *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1770–1777, 2017.

[24] T. Schneider, M. Li, C. Cadena, J. Nieto, and R. Siegwart, "Observability-aware self-calibration of visual and inertial sensors for ego-motion estimation," *IEEE Sensors Journal*, vol. 19, no. 10, pp. 3846–3860, May 2019.

[25] M. Li and A. I. Mourikis, "Online temporal calibration for Camera-IMU systems: Theory and algorithms," vol. 33, no. 7, pp. 947–964, Jun. 2014.

[26] D. Schubert, N. Demmel, L. von Stumberg, V. Usenko, and D. Cremers, "Rolling-shutter modelling for visual-inertial odometry," in *International Conference on Intelligent Robots and Systems (IROS)*, November 2019.

[27] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 1280–1286.

[28] E. S. Jones and S. Soatto, "Visual-inertial navigation, mapping and localization: A scalable real-time causal approach," *International Journal of Robotics Research*, vol. 30, no. 4, pp. 407–430, Apr. 2011.

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

[30] Y. Xiao, X. Ruan, J. Chai, X. Zhang, and X. Zhu, "Online imu self-calibration for visual-inertial systems," *Sensors*, vol. 19, no. 7, 2019.

[31] J. H. Jung, S. Heo, and C. G. Park, "Observability analysis of imu intrinsic parameters in stereo visual–inertial odometry," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7530–7541, 2020.

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

[33] J. Huai, Y. Lin, Y. Zhuang, C. K. Toth, and D. Chen, "Observability analysis and keyframe-based filtering for visual inertial odometry with full self-calibration," *IEEE Transactions on Robotics*, vol. 38, no. 5, pp. 3219–3237, 2022.

[34] G. Huang, A. I. Mourikis, and S. I. Roumeliotis, "Observability-based rules for designing consistent EKF SLAM estimators," *International Journal of Robotics Research*, vol. 29, no. 5, pp. 502–528, Apr. 2010.

[35] A. Martinelli, "State estimation based on the concept of continuous symmetry and observability analysis: The case of calibration," *IEEE Transactions on Robotics*, vol. 27, no. 2, pp. 239–255, 2011.

[36] J. Hernandez, K. Tsotsos, and S. Soatto, "Observability, identifiability and sensitivity of vision-aided inertial navigation," in *Proc. of the IEEE International Conference on Robotics and Automation*, Seattle, WA, May 26–30, 2015, pp. 2319–2325.

[37] J. Hesch, D. Kottas, S. Bowman, and S. Roumeliotis, "Consistency analysis and improvement of vision-aided inertial navigation," *IEEE Transactions on Robotics*, vol. 30, no. 1, pp. 158–176, 2013.

[38] Y. Yang, C. Chen, W. Lee, and G. Huang, "Decoupled right invariant error states for consistent visual-inertial navigation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1627–1634, 2022.

[39] A. B. Chatfield, *Fundamentals of High Accuracy Inertial Navigation*. American Institute of Aeronautics and Astronautics, 1997.

[40] M. Brossard, A. Barrau, P. Chauchat, and S. Bonnabel, "Associating uncertainty to extended poses for on lie group imu preintegration with rotating earth," *IEEE Transactions on Robotics*, vol. 38, no. 2, pp. 998–1015, 2022.

[41] J. Nikolic, M. Burri, I. Gilitschenski, J. Nieto, and R. Siegwart, "Non-parametric extrinsic and intrinsic calibration of visual-inertial sensor systems," *IEEE Sensors Journal*, vol. 16, no. 13, pp. 5433–5443, 2016.

[42] N. Trawny and S. I. Roumeliotis, "Indirect Kalman filter for 3D attitude estimation," University of Minnesota, Dept. of Comp. Sci. & Eng., Tech. Rep., Mar. 2005. [Online]. Available: http://mars.cs.umn.edu/tr/reports/Trawny05b.pdf

[43] Y. Yang, B. P. W. Babu, C. Chen, G. Huang, and L. Ren, "Analytic combined imu integration for visual-inertial navigation," in *Proc. of the IEEE International Conference on Robotics and Automation*, Paris, France, 2020.

[44] Y. Yang, P. Geneva, X. Zuo, and G. Huang, "Technical report: Online self-calibration for visual-inertial navigation systems: models, analysis and degeneracy," University of Delaware, Tech. Rep. RPNG-2022-CALIB, 2022. [Online]. Available: https://yangyulin.net/papers/2022_tr_fullcalib.pdf

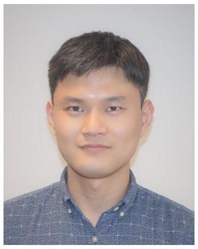
[45] K. J. Wu, C. X. Guo, G. Georgiou, and S. I. Roumeliotis, "VINS on wheels," in *Proc. of the IEEE International Conference on Robotics and Automation*, May 2017, pp. 5155–5162.

[46] Y. Yang and G. Huang, "Observability analysis of aided ins with heterogeneous features of points, lines and planes," *IEEE Transactions on Robotics*, vol. 35, no. 6, pp. 399–1418, Dec. 2019.

[47] P. S. Maybeck, *Stochastic Models, Estimation, and Control*. London: Academic Press, 1979, vol. 1.

[48] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stückler, and D. Cremers, "The tum vi benchmark for evaluating visual-inertial odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1680–1687.

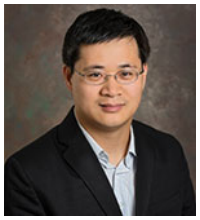
[49] P. Geneva, K. Eckenhoff, and G. Huang, "A linear-complexity EKF for visual-inertial navigation with loop closures," in *Proc. International Conference on Robotics and Automation*, Montreal, Canada, May 2019.



Yulin Yang received the B.Eng. degree in Mechanical Engineering from Shandong University, China, in 2009, M.Sc. in Mechanical Engineering from Xi'an Jiaotong University, China in 2012 and M.Sc. in Mathematics from University of Delaware, USA in 2020. From 2012 to 2015, he was a Research & Development Engineer at Siemens in Shanghai, China. He is currently working toward a Ph.D. in department of Mechanical Engineering at University of Delaware. He is the recipient of the University Doctoral Fellowship Award (2020). His research topics focus on visual inertial navigation, SLAM and nonlinear estimation.

Patrick Geneva received the B.Eng. degree in Mechanical Engineering from the University of Delaware with a minors in Computer Science & Mathematics. He is interested in the areas of robotics and state estimation with his primary research being on simultaneous localization and mapping (SLAM) and visual-inertial navigation systems (VINS) with a focus on efficient and resource constrained applications. He is the recipient of the University Doctoral Fellowship Award (2021), NASA Delaware Space Grant (DESG) Graduate Fellowship (2019 and 2022), and Mary and George Nowinski Award for Excellence in Undergraduate Research Award (2017).

Xingxing Zuo received the B.Eng. degree in mechanical engineering from the University of Electronic Science and Technology of China (UESTC) in 2016, and the Ph.D. degree in Control Science and Engineering from Zhejiang University in 2021. He is currently a postdoc researcher at the Technical University of Munich. His research interests include computer vision, state estimation, sensor fusion, deep learning, localization and mapping for autonomous robots in complex environments. He was the finalist for the Best Paper Award in Robot Vision in the 2021 IEEE International Conference on Robotics and Automation (ICRA).



Guoquan Huang (Senior Member, IEEE) received his BS in automation (electrical engineering) from the University of Science and Technology Beijing, China, in 2002, and MS and PhD in computer science from the University of Minnesota–Twin Cities, in 2009 and 2012, respectively. He currently is an Associate Professor of Mechanical Engineering (ME) and Computer and Information Sciences (CIS) at the University of Delaware (UD), where he is leading the Robot Perception and Navigation Group (RPNG). From 2012 to 2014, he was a Postdoctoral Associate with MIT CSAIL (Marine Robotics). His research interests focus on state estimation and spatial perception for autonomous vehicles and mobile devices, including probabilistic sensing, estimation, localization, mapping, perception, and navigation. He serves as an Associate Editor for the IEEE Transactions on Robotics (T-RO), IEEE Robotics and Automation Letters (RA-L), and IET Cyber-Systems and Robotics (CSR). Dr. Huang has received the 2015 UD Research Award (UDRF), 2015 NASA DE Space Research Seed Award, 2016 NSF CRII Award, 2018 SATEC Robotics Delegation (ASME), the Google Daydream/ARVR/AI Faculty Research Award (2018, 2019, 2022), and the IROS 2019 FPV Drone Racing VIO Competition Winner. He was the recipient of the 2023 GNC Best Overseas Paper, ICRA 2022 Outstanding Navigation Paper Award, and the Finalists of the ICRA 2021 Best Paper Award (Robot Vision) and RSS 2009 Best Paper Award