A Rotation-Translation-Decoupled Solution for Robust and Efficient Visual-Inertial Initialization

Yijia He *
Chinese Academy of Sciences

Bo Xu *
Wuhan University

Zhanpeng Ouyang
ShanghaiTech University

Hongdong Li
Australian National University

Abstract

We propose a novel visual-inertial odometry (VIO) initialization method, which decouples rotation and translation estimation, and achieves higher efficiency and better robustness. Existing loosely-coupled VIO-initialization methods suffer from poor stability of visual structure-from-motion (SfM), whereas those tightly-coupled methods often ignore the gyroscope bias in the closed-form solution, resulting in limited accuracy. Moreover, the aforementioned two classes of methods are computationally expensive, because 3D point clouds need to be reconstructed simultaneously. In contrast, our new method fully combines inertial and visual measurements for both rotational and translational initialization. First, a rotation-only solution is designed for gyroscope bias estimation, which tightly couples the gyroscope and camera observations. Second, the initial velocity and gravity vector are solved with linear translation constraints in a globally optimal fashion and without reconstructing 3D point clouds. Extensive experiments have demonstrated that our method is $8 \sim 72$ *times faster (w.r.t. a* 10*-frame set) than the state-of-the-art methods, and also presents significantly higher robustness and accuracy. The source code is available at* https://github.com/boxuLibrary/drt-vio-init.

1. Introduction

Visual-inertial odometry (VIO) aims to estimate camera motion and recover 3D scene structure by fusing both image and IMU measurements. The low-cost and compactness of the camera module and IMU sensors make VIO widely used in virtual or augmented reality systems (VR/AR) and various autonomous navigation systems. Currently, most VIO systems track camera motion by minimizing nonlinear visual re-projection errors [14, 30], so the accuracy of the initial value will affect the convergence. In addition, the robustness and lower latency of the initialization are also very important for the downstream application, e.g. AR developers need accurate camera tracking within a few hundred milliseconds after launching VIO, regardless of the use case. For the sensor that has calibrated intrinsic and extrinsic parameters, the initial variables for VIO include the gravity vector, initial velocity, gyroscope and accelerometer biases.

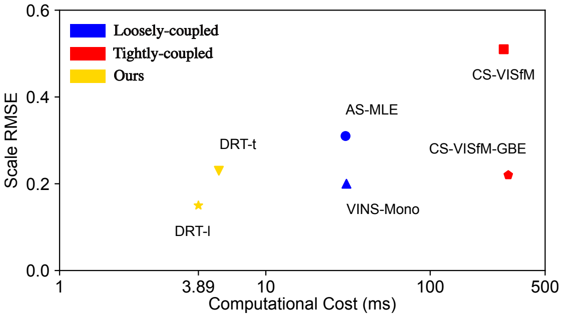

Figure 1. Comparison of computational cost and scale factor errors on EuRoC dataset. Different colors indicate different types of methods. Our proposed initialization method for decoupling rotation and translation (DRT) is accurate and computationally efficient.

Many VIO systems are initialized by setting the initial velocity to zero, then calculating the gravity vector and gyroscope bias with IMU measurements [14,20,36]. However, this method only works when the system is strictly static. For sensors in motion, loosely-coupled and tightly-coupled initialization methods are widely studied. As shown in Fig. 2, the loosely-coupled methods [5,28,30] combine the camera poses estimated by visual SfM and the IMU measurements to estimate the initial state variables. However, visual SfM is prone to inaccuracy or failure when co-viewed

[*] Equal contribution.

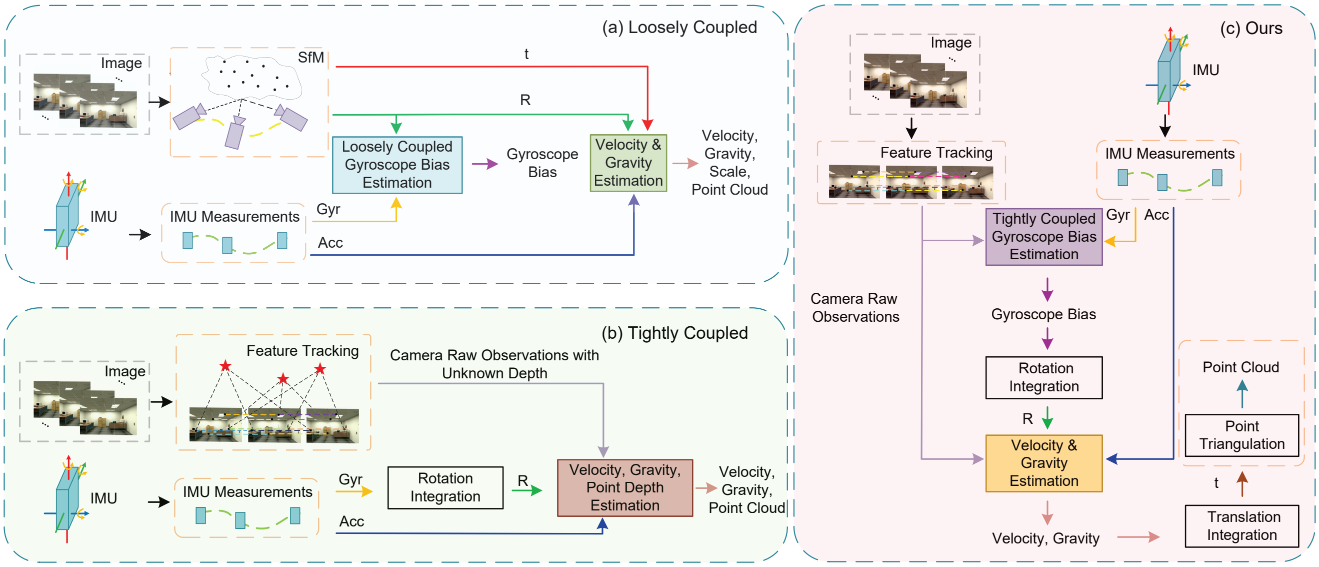

Figure 2. Comparison between our method and previous VIO initialization methods. Different colored arrows indicate different information flows for VI fusion. Our method takes full advantage of the complementary information between vision and IMU. In contrast, previous loosely-coupled methods do not incorporate IMU information into visual SfM, and previous tightly-coupled methods do not use visual observations to remove gyroscope bias, either of which affects the robustness and accuracy of VIO initialization.

frames are insufficient or the camera rotates rapidly. The motion information measured by IMU is not used to improve the robustness of visual SfM. The tightly-coupled methods [8, 9, 24, 25] firstly use gyroscope measurements and calibrated extrinsic parameters to estimate camera rotation, then use closed-form solution constructed with vision and accelerometer observations to solve for the initial velocity and gravity vector. However, this type of method has poor accuracy on systems equipped with inexpensive and noisy IMU (e.g. cell phones), because no visual observations are used to estimate the gyroscope bias. Moreover, the three-dimensional coordinates of point clouds are obtained with the closed-form solution, resulting in a large and time-consuming solution matrix. Both the above two kinds of methods under-utilize the complementary advantages between visual and inertial sensors, resulting in limited accuracy and robustness.

According to [17, 18, 26, 38], image observations could be directly used to optimize frame-to-frame rotation and camera poses could be efficiently solved with linear global translation constraints [3]. Inspired by this, we propose a novel rotation-translation-decoupled VIO initialization framework. Gyroscope measurements are directly integrated into the camera rotation estimation, which greatly improves the robustness of initialization, and the translation related initial variables are solved efficiently without estimating the 3D structure. As shown in Fig. 1, our method achieves the lowest scale error and is significantly faster than previous methods. The scale factor error is one of the metrics for evaluating the initialization. Our main contributions are

- We propose a rotation-only solution to directly optimize gyroscope bias using image observations, which can obtain camera rotation more efficiently and more robustly compared to vision-only methods.
- We propose a globally optimal solution for estimating the initial velocity and gravity vector based on linear translation constraints. Its linearity and independence of scene structure significantly benefit computational efficiency.
- Our proposed initialization framework outperforms the state-of-the-art in both accuracy and robustness on public datasets while being $8 \sim 72$ times faster in calculation time for a 10-frame set. We published our code to facilitate communication.

2. Related Work

Visual-inertial odometry has been widely studied in terms of reducing time consumption or improving accuracy [7, 21, 22, 37]. A robust and accurate initialization method is indispensable for VIO. Many influential VIO systems [4, 14, 28, 30] have their own designed initialization methods.

Martinelli [25] proposed an impressive tightly-coupled closed-form solution that uses tracked visual features and accelerometer measurements to jointly estimate initial state variables and features depth. But the gyroscope is assumed to be unbiased. [16] and [8] demonstrated that the gyroscope bias will significantly affect the accuracy of the closed-form solution, and proposed a method to iteratively optimize the gyroscope bias. Recently, [11] reduces the computational complexity by using the projection matrix to eliminate the variables of the solution matrix, and the gyroscope bias is obtained with a global bundle adjustment optimization. However, these methods of estimating gyroscope bias by minimizing the nonlinear loss function are sensitive to the results of closed-form solutions and are computationally time-consuming.

With the development of visual odometry or SfM [10, 13, 27], loosely-coupled methods for estimating VIO initial variables with high-precision camera trajectories as measurements were naturally proposed [28,31]. Recently, Campos et al. [5] pointed out that the previous method did not consider the IMU measurement uncertainty, and proposed to use the maximum a posteriori to optimize the initial variables. Zuñiga-Noël et al. [40] extended this method to a non-iterative efficient analytical solution.

Benefiting from deep learning-based monocular depth estimation [32,33], Zhou et al. [39] used the learned monocular depth as input to improve the robustness of VIO initialization in scenarios with small parallax and low motion excitation. However, this method is limited by the generalization ability of the learning-based model and the large computational cost of the convolutional network.

3. Notations and Preliminaries

In this section, notations are defined and IMU motion model is given. Let $\mathcal{F}_{c_i}$ and $\mathcal{F}_{b_i}$ denote the camera frame and IMU frame at time-index i. $\mathbf{T}_{b_i b_j}$ to be the Euclidean transformation that take 3D points from IMU frame at time-index j to the one at time-index i, which consisted of translation $\mathbf{p}_{b_i b_j}$ and rotation $\mathbf{R}_{b_i b_j}$. The calibrated extrinsic transformation from $\mathcal{F}_b$ to $\mathcal{F}_c$ is denoted by $\mathbf{T}_{cb}$. $\lfloor \cdot \rfloor_\times$ and $\|\cdot\|$ are skew-symmetric operator and Euclidean norm operator, respectively.

The IMU integration follows the standard approach on $\mathrm{SO}(3)$ manifold as proposed in [12].

$$\begin{aligned}
\mathbf{p}_{b_1 b_j} &= \mathbf{p}_{b_1 b_i} + \mathbf{v}^{b_1}_{b_1} \Delta t_{ij} - \frac{1}{2}\mathbf{g}^{b_1}\Delta t_{ij}^2 + \mathbf{R}_{b_1 b_i}\boldsymbol{\alpha}^{b_i}_{b_j} \\
\mathbf{v}^{b_1}_{b_j} &= \mathbf{v}^{b_1}_{b_i} - \mathbf{g}^{b_1}\Delta t_{ij} + \mathbf{R}_{b_1 b_i}\boldsymbol{\beta}^{b_i}_{b_j} \\
\mathbf{R}_{b_1 b_j} &= \mathbf{R}_{b_1 b_i}\boldsymbol{\gamma}^{b_i}_{b_j}
\end{aligned} \tag{1}$$

where $\mathbf{v}^{b_1}_{b_1}$ and $\mathbf{g}^{b1}$ represent the initial velocity and gravity vector in $\mathcal{F}_{b_1}$ which are need to be estimated. $\boldsymbol{\alpha}^{b_i}_{b_j}$, $\boldsymbol{\beta}^{b_i}_{b_j}$, $\boldsymbol{\gamma}^{b_i}_{b_j}$ are defined as the pre-integration of translation, velocity, and rotation, respectively. Δt_{ij} is the time interval from time i to time j.

$$\begin{aligned}
\boldsymbol{\alpha}^{b_i}_{b_j} &= \sum_{k=i}^{j-1}\left(\left(\sum_{f=i}^{k-1}\mathbf{R}_{b_i b_f}\mathbf{a}^m_f \Delta t\right)\Delta t + \frac{1}{2}\mathbf{R}_{b_i b_k}\mathbf{a}^m_k \Delta t^2\right) \\
\boldsymbol{\beta}^{b_i}_{b_j} &= \sum_{k=i}^{j-1}\mathbf{R}_{b_i b_k}\mathbf{a}^m_k \Delta t \\
\boldsymbol{\gamma}^{b_i}_{b_j} &= \prod_{k=i}^{j-1}\mathrm{Exp}\left(\boldsymbol{\omega}^m_k \Delta t\right)
\end{aligned} \tag{2}$$

where function $\mathrm{Exp}(\cdot) : \mathfrak{so}(3) \rightarrow SO(3)$ for Lie algebra to Lie group. $\boldsymbol{\omega}^m_k$ and $\boldsymbol{a}^m_k$ denote the gyroscope and accelerometer measurements at time k, respectively. Δt represents the time interval between adjacent IMU data.

Note that the pre-integration formula does not take into account the bias of the measurement. Considering the acceleration bias and the gravity vector are coupled together and cannot be distinguished in a small motion, ignoring the acceleration bias will not greatly affect the initialization results [8, 31]. In this work, the acceleration bias is assumed to be zero, and the effect of gyroscope bias on the $\boldsymbol{\gamma}^{b_i}_{b_j}$ can be represented by a first-order Taylor approximation [12].

$$\hat{\boldsymbol{\gamma}}^{b_i}_{b_j} = \boldsymbol{\gamma}^{b_i}_{b_j}\,\mathrm{Exp}\left(\mathbf{J}^{\gamma^{b_i}_{b_j}}_{\mathbf{b}_g}\mathbf{b}_g\right) \tag{3}$$

where $\mathbf{b}_g$ is gyroscope bias which to be estimated, $\hat{\boldsymbol{\gamma}}^{b_i}_{b_j}$ is the updated $\boldsymbol{\gamma}^{b_i}_{b_j}$, $\mathbf{J}^{\gamma^{b_i}_{b_j}}_{\mathbf{b}_g}$ is the Jacobian of the derivative of $\boldsymbol{\gamma}^{b_i}_{b_j}$ with respect to $\mathbf{b}_g$ and is a constant can be calculated [12].

4. Our Initialization Framework

An accurate and robust rotation estimation is crucial for improving the trajectory accuracy of the system since the rotation will affect the accumulation of translation vectors. In this section, we first introduce our method for robust estimation of gyroscope bias using at least two images. Then, we derive two linear solutions for the initial velocity and gravity vector after the rotation is obtained by gyroscope integration.

4.1. Gyroscope Bias Optimizer

The rotation between the two cameras can be directly iteratively optimized using the geometric constraints constructed by feature correspondences [17], but this method requires the initial value of the rotation to be close to the ground truth. We extend this method to visual-inertial systems to avoid the above problem and extend it to solving rotations between multiple views.

In this paragraph, we revisit the main idea of directly optimizing frame-to-frame rotation [17]. As shown in Fig. 3,

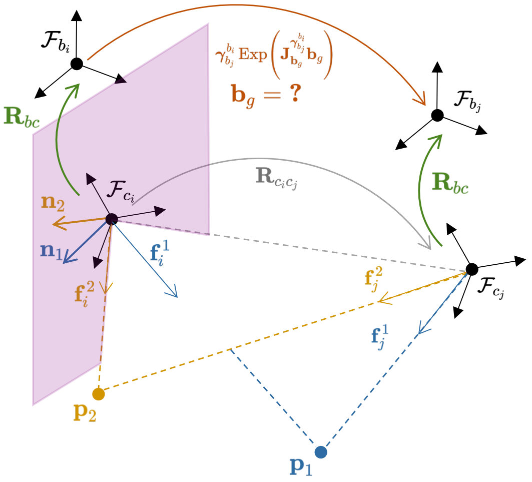

Figure 3. Geometric relationships of unit feature observation vectors and gyroscope bias. The normal vectors (yellow and blue) perpendicular to the corresponding epipolar plane (yellow and blue) should be coplanar (purple plane), which can form a constraint for solving the relative rotation (gray). The gyroscope bias optimizer converts the value to be solved into gyroscope bias (orange) by the known extrinsic rotation (green).

if a 3D point $\mathbf{p}_1$ can be observed by two cameras, the two camera centers $\mathcal{F}_{c_i}$ and $\mathcal{F}_{c_j}$ and the 3D point construct an epipolar plane. Define $\mathbf{f}_i^1$ and $\mathbf{f}_j^1$ represent the unit vectors pointing from $\mathcal{F}_{c_i}$ and $\mathcal{F}_{c_j}$ to $\mathbf{p}_1$, respectively. The normal vector of the epipolar plane can be calculated by cross product, $\mathbf{n}^k = \lfloor \mathbf{f}_i^k \rfloor_\times \mathbf{R}_{c_ic_j}\mathbf{f}_j^k$. The normal vectors of all epipolar planes will be perpendicular to the translation vector, which means these normal vectors need to be coplanar. Suppose we have n 3D points observed in two frames, stacking all normal vectors into a matrix $\mathbf{N} = \begin{bmatrix}\mathbf{n}^1 & \dots & \mathbf{n}^n\end{bmatrix}$, then coplanarity is algebraically equivalent to the minimum eigenvalue of the matrix $\mathbf{M} = \mathbf{N}\mathbf{N}^\top$ equal to zero. The final problem of calculating the relative rotation $\mathbf{R}_{c_ic_j}$ is parameterized as

$$\begin{aligned} \mathbf{R}_{c_ic_j}^* &= \underset{\mathbf{R}_{c_ic_j}}{\operatorname{argmin}} \lambda_{\mathbf{M}_{ij},\min} \\ \text{with } \mathbf{M}_{ij} &= \sum_{k=1}^{n} \left(\lfloor \mathbf{f}_i^k \rfloor_\times \mathbf{R}_{c_ic_j}\mathbf{f}_j^k\right)\left(\lfloor \mathbf{f}_i^k \rfloor_\times \mathbf{R}_{c_ic_j}\mathbf{f}_j^k\right)^\top \end{aligned} \tag{4}$$

where $\lambda_{\mathbf{M}_{ij},\min}$ is the smallest eigenvalue of $\mathbf{M}_{ij}$.

For the visual and inertial system with known extrinsic parameters $\mathbf{R}_{bc}, \mathbf{p}_{bc}$, the relative motion between cameras can be represented in the IMU body frame:

$$\begin{aligned} \mathbf{R}_{c_ic_j} &= \mathbf{R}_{bc}^\top \mathbf{R}_{b_ib_j} \mathbf{R}_{bc} \\ \mathbf{p}_{c_ic_j} &= \mathbf{R}_{bc}^\top \left(\mathbf{p}_{b_ib_j} + \mathbf{R}_{b_ib_j}\mathbf{p}_{bc} - \mathbf{p}_{bc}\right) \end{aligned} \tag{5}$$

where $\mathbf{R}_{b_ib_j}$ represents the rotation from $\mathcal{F}_{b_j}$ to $\mathcal{F}_{b_i}$, which can be obtained by integrating the gyroscope measurements between time i and time j using Eq. (3).

Combining Eq. (3), (4) and (5), $\mathbf{M}_{ij}$ can be represented as a new matrix related to gyroscope information

$$\begin{aligned} \mathbf{M}'_{ij} = \sum_{k=1}^{n} &\left(\lfloor \mathbf{f}_i^k \rfloor_\times \mathbf{R}_{bc}^\top \gamma_{b_j}^{b_i}\,\mathrm{Exp}\left(\mathbf{J}_{\mathbf{b}_g}^{\gamma_{b_j}^{b_i}}\mathbf{b}_g\right)\mathbf{R}_{bc}\mathbf{f}_j^k\right) \\ &\left(\lfloor \mathbf{f}_i^k \rfloor_\times \mathbf{R}_{bc}^\top \gamma_{b_j}^{b_i}\,\mathrm{Exp}\left(\mathbf{J}_{\mathbf{b}_g}^{\gamma_{b_j}^{b_i}}\mathbf{b}_g\right)\mathbf{R}_{bc}\mathbf{f}_j^k\right)^\top \end{aligned} \tag{6}$$

where only $\mathbf{b}_g$ needs to be estimated. Using the properties of the rotation matrix and vector cross product, Eq. (6) can be further simplified to the following form

$$\begin{aligned} \mathbf{M}'_{ij} = \sum_{k=1}^{n} &\left(\mathbf{R}\lfloor \mathbf{f}_i^{k\prime} \rfloor_\times \mathrm{Exp}\left(\mathbf{J}_{\mathbf{b}_g}^{\gamma_{b_j}^{b_i}}\mathbf{b}_g\right)\mathbf{f}_j^{k\prime}\right) \\ &\left(\mathbf{R}\lfloor \mathbf{f}_i^{k\prime} \rfloor_\times \mathrm{Exp}\left(\mathbf{J}_{\mathbf{b}_g}^{\gamma_{b_j}^{b_i}}\mathbf{b}_g\right)\mathbf{f}_j^{k\prime}\right)^\top \end{aligned} \tag{7}$$

where $\mathbf{R} = \mathbf{R}_{bc}^\top \gamma_{b_j}^{b_i}$, $\mathbf{f}_i^{k\prime} = \mathbf{R}^\top \mathbf{f}_i^k$, and $\mathbf{f}_j^{k\prime} = \mathbf{R}_{bc}\mathbf{f}_j^k$. Please refer to the supplement material Sec.1 for details.

Let $\mathcal{E}$ denote the set of keyframe pairs that observe enough common features. Since the gyroscope bias is slowly time-varying, it can be assumed to be a constant during VIO initialization, so any keyframe pair $(i, j) \in \mathcal{E}$ can be used to estimate the gyroscope bias. To fully utilize all visual observations, multiple keyframe pairs are combined to optimize the solution

$$\begin{aligned} \mathbf{b}_g^* &= \underset{\mathbf{b}_g}{\operatorname{argmin}} \lambda \\ \text{with } \lambda &= \sum_{(i,j)\in\mathcal{E}} \lambda_{\mathbf{M}'_{ij},min} \end{aligned} \tag{8}$$

Eq. (8) is the cornerstone and one of the main contributions of this paper. To solve Eq. (8), quaternions are used as minimal rotation parameterization, and the Levenberg-Marquardt strategy with automatic differentiation in ceres [1] is used to iteratively optimize the solution [17]. Since the gyroscope bias is small (usually less than $0.1\ rad/s$), we can set the initial value $\mathbf{b}_g = \mathbf{0}$ during iterative optimization.

In addition, after solving $\mathbf{b}_g$, we can calculate the rotation matrices between all cameras by integrating the bias-removed gyroscope measurements. Although there are multiple keyframes, only a three-dimensional variable $\mathbf{b}_g$ needs to be solved.

4.2. Velocity And Gravity Estimator

After the rotation is calculated, the initial velocity and gravity vector of the system can be solved efficiently without estimating 3D point clouds. In this section, tightly-coupled and loosely-coupled solvers based on linear translation constraints are presented separately.

4.2.1 Tightly-Coupled Solution

Assuming that the first frame of a multi-frame sequence is the world coordinate system, the position of each frame in the world coordinate system can be solved efficiently by directly using the linear global translation constraint [3]. Suppose there are three keyframes and the index of these keyframes is r, i, and l, respectively. The LiGT constraint can be expressed as :

$$\mathbf{B}\mathbf{p}_{c_1c_r} + \mathbf{C}\mathbf{p}_{c_1c_i} + \mathbf{D}\mathbf{p}_{c_1c_l} = 0, \quad 1 \le i \le n, i \ne l \tag{9}$$

where

$$\begin{aligned} \mathbf{B} &= \lfloor \mathbf{f}_i^k \rfloor_\times \mathbf{R}_{c_ic_l} \mathbf{f}_l^k \boldsymbol{a}_{lr}^\top \mathbf{R}_{c_rc_1} \\ \mathbf{C} &= \theta_{lr}^2 \lfloor \mathbf{f}_i^k \rfloor_\times \mathbf{R}_{c_ic_1} \\ \mathbf{D} &= -(\mathbf{B} + \mathbf{C}) \\ \boldsymbol{a}_{lr}^\top &= \left(\lfloor \mathbf{R}_{c_rc_l} \mathbf{f}_l^k \rfloor_\times \mathbf{f}_r^k \right)^\top \lfloor \mathbf{f}_r^k \rfloor_\times \\ \theta_{lr} &= \left\| \lfloor \mathbf{f}_r^k \rfloor_\times \mathbf{R}_{c_rc_l} \mathbf{f}_l^k \right\| \end{aligned} \tag{10}$$

When we have $m(m > 3)$ keyframes, there are multiple $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$. We can concatenate them and define as the coefficient matrix $\mathbf{L}$ which containing only visual observations and global rotations. Let $\mathbf{P} = \left(\mathbf{p}_{c_1c_r}^\top, ..., \mathbf{p}_{c_1c_n}^\top \right)^\top$, then Eq. (9) can be written as

$$\mathbf{L} \cdot \mathbf{P} = 0 \tag{11}$$

The positions of all cameras concerning the first keyframe can be solved by Eq. (11). Since the translation vectors of Eq. (9) are all about the camera coordinate system, we substitute Eq. (5) into Eq. (9) to transform the camera coordinate system into IMU coordinate system.

$$\mathbf{B}'\mathbf{u}_r + \mathbf{C}'\mathbf{u}_i + \mathbf{D}'\mathbf{u}_l = 0, \quad 1 \le i \le n, i \ne l \tag{12}$$

where

$$\begin{aligned} &\mathbf{B}' = \mathbf{B}\mathbf{R}_{bc}^\top, \quad \mathbf{C}' = \mathbf{C}\mathbf{R}_{bc}^\top, \quad \mathbf{D}' = \mathbf{D}\mathbf{R}_{bc}^\top \\ &\mathbf{u}_m = \left(\mathbf{p}_{b_1b_m} + \mathbf{R}_{b_1b_m}\mathbf{p}_{bc} - \mathbf{p}_{bc} \right), \quad m \in r, i, l \end{aligned} \tag{13}$$

All global translations in the above formulation can be replaced using IMU integration formulation in Eq. (1). Therefore, the system of linear equations (11) for solving the global position is transformed into the following equations for solving the initial velocity and gravity vector,

$$\begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \mathbf{v}_{b_1}^{b_1} \\ \mathbf{g}^{b_1} \end{bmatrix} = \mathbf{d} \tag{14}$$

where

$$\begin{aligned} \mathbf{A}_1 &= \mathbf{B}'\Delta t_{1r} + \mathbf{C}'\Delta t_{1i} + \mathbf{D}'\Delta t_{1l} \\ \mathbf{A}_2 &= \frac{1}{2} \left(\mathbf{B}'\Delta t_{1r}^2 + \mathbf{C}'\Delta t_{1i}^2 + \mathbf{D}'\Delta t_{1l}^2 \right) \\ \mathbf{d} &= -\mathbf{B}'\mathbf{s}_{1r} - \mathbf{C}'\mathbf{s}_{1i} - \mathbf{D}'\mathbf{s}_{1l} \\ \mathbf{s}_{1m} &= \boldsymbol{\alpha}_{b_m}^{b_1} + \mathbf{R}_{b_1b_m}\mathbf{p}_{bc} - \mathbf{p}_{bc}, \quad m \in r, i, l \end{aligned} \tag{15}$$

For detailed derivation, please refer to the supplement material Sec.2. Since the norm of the gravity vector is constant, the Lagrange multiplier method [6] is used to find the optimal solution for the constrained least squares problem.

4.2.2 Loosely-Coupled Solution

The loosely-coupled approach requires computing the camera translation first, which is then combined with the IMU measurements to compute the initial state variables. The camera pose can be calculated by Eq. (9), and if monocular camera is used, the position obtained from LiGT constraint is up to scale, then the metric scale factor s needs to be computed explicitly during initialization. Define $\mathcal{X}$ as a vector of initial state variables, $\mathbf{v}_{b_n}^{b_n}$ means the velocity of body in $\mathcal{F}_{b_n}$, $\mathbf{g}^{c_0}$ is the gravity in $\mathcal{F}_{c_0}$ and s is the metric scale.

$$\mathcal{X} = \left[\mathbf{v}_{b_0}^{b_0}, \mathbf{v}_{b_1}^{b_1}, \cdots, \mathbf{v}_{b_n}^{b_n}, s, \mathbf{g}^{c_0} \right] \in \mathbb{R}^{3(n+1)+1+3} \tag{16}$$

Assume that the body coordinate systems corresponding to two keyframes are $\mathcal{F}_{b_i}$ and $\mathcal{F}_{b_k}$, The following constraints [31] exist between IMU and visual measurements

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_{b_k}^{b_i} &= \mathbf{R}_{b_ic_0} \left(s \left(\mathbf{p}_{c_0b_k} - \mathbf{p}_{c_0b_i} \right) + \frac{1}{2}\mathbf{g}^{c_0}\Delta t_{ik}^2 \right) - \mathbf{v}_{b_i}^{b_i}\Delta t_{ik} \\ \hat{\boldsymbol{\beta}}_{b_k}^{b_i} &= \mathbf{R}_{b_ic_0} \left(\mathbf{R}_{c_0b_k}\mathbf{v}_{b_k}^{b_k} + \mathbf{g}^{c_0}\Delta t_{ik} \right) - \mathbf{v}_{b_i}^{b_i} \end{aligned} \tag{17}$$

The residual is the difference between the estimated and measured values, it can be parameterized as

$$\mathbf{r}\left(\mathcal{X}\right) = \begin{bmatrix} \boldsymbol{\alpha}_{b_k}^{b_i} - \hat{\boldsymbol{\alpha}}_{b_k}^{b_i} \\ \boldsymbol{\beta}_{b_k}^{b_i} - \hat{\boldsymbol{\beta}}_{b_k}^{b_i} \end{bmatrix} \tag{18}$$

In the presence of noise this system have no exact solution, so we still use the least squares solution:

$$\mathbf{H}' \begin{bmatrix} \mathbf{v}_{b_i}^{b_i} \\ \mathbf{v}_{b_k}^{b_k} \\ s \\ \mathbf{g}^{c_0} \end{bmatrix} = \mathbf{b}' \tag{19}$$

Finally, we stack $\mathbf{H}'$ generated by multiple adjacent frames into a coefficient matrix $\mathbf{H}$, and the same is true for $\mathbf{b}'$ and $\mathbf{b}$. Solving the least squares solution $\mathbf{H}\mathcal{X} = \mathbf{b}$, $\mathcal{X}$ can be obtained. Please refer to the supplement material Sec.2 for the specific forms of $\mathbf{H}'$ and $\mathbf{b}'$.

5. Experiments

In the section, simulation experiments are first used to verify the effectiveness of our method, and then the evaluation on real datasets demonstrates the accuracy, robustness and computational efficiency. Gyroscope bias error, velocity error, gravity direction error, and scale factor error are used to evaluate the performance of each algorithm. We perform a *Sim*(3) alignment [35] against the ground truth trajectory to get scale error. For the gyroscope bias error, let $\overline{bg}$ be the mean of all biases in the GT trajectory, and the percent of the relative error is computed with $\left| ||bg|| - ||\overline{bg}|| \right| / ||\overline{bg}||$. We divided all the datasets collected by continuous motion in different scenarios into several data segments. Each data segment used for initialization consists of 10 keyframes, where the keyframes are obtained by sampling image frames at a frequency of $4Hz$ as used in other works [5, 40]. In all quantitative experiments, only datasets with successful initialization, i.e., scale error less than 1 ($|s - 1| < 1$), were used for statistics, and the Root Mean Square Error (RMSE) is used for evaluation. All the experiments were conducted on a computer with Intel i7-9750H@ 2.6GHz CPU.

The loosely-coupled VI-initialization methods used for comparison include the VINS-Mono initialization (denoted as VINS-Mono) [30] and the analytical-solution [40] which is an improved work of the ORB-SLAM3 initialization [5] (denoted as AS-MLE). The code for the tightly-coupled initialization (denoted as CS-VISfM) [25] used for comparison is from the open-sourced SLAM OpenVINS [14]. We denote our method of solving the velocity and gravity vectors in a tightly-coupled or loosely-coupled manner as DRT-t and DRT-l, respectively. To verify the necessity of a gyroscope bias estimator (GBE), we evaluate the method combining our GBE with CS-VISfM (denoted as CS-VISfM-GBE) and the DRT-l method without GBE (denoted as DRT-l-wo-GBE) as ablation experiments.

5.1. Simulation Experiments

We simulated camera motion with $20Hz$ and IMU measurements with $200Hz$, forming an ellipse trajectory with sinusoidal vertical motion. The long-semi axis and short-semi axis of the ellipse trajectory are $4m$ and $3m$, respectively. The number of observed feature points in each frame is limited to 150, and the Gaussian noise with standard deviation $\delta_{pix} = 1$ pixel is added to the landmark observations. The simulated acceleration and gyroscope measurements are computed from the analytic derivation of the parametric trajectory and additionally corrupted by white noise and slowly time-varying bias terms [1]. To verify the convergence of the gyroscope bias estimator, we set different gyroscope biases from 0.02 rad/s to 0.18 rad/s during simulation.

Table 1. Initialization accuracy in gyroscope bias error (%), gravity direction error (°), velocity error (m/s) and scale error metrics with DRT-t method.

Metrics	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18
Bg	28.50	12.56	7.23	5.82	4.02	3.53	2.48	2.52	2.03
G.Dir	0.58	0.60	0.58	0.59	0.60	0.59	0.61	0.59	0.61
Vel	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09
Scale	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08

As shown in Tab. 1, the percentage of the gyro bias error decreases with the increase in the gyro bias magnitude, which is caused by the estimated absolute error of gyro bias all being about 0.01 rad/s. Meanwhile, the gravity direction errors obtained by DRT-t are about 0.6°, and the trajectory scale errors are 0.08, indicating the effectiveness of our initialization algorithm. The simulation results show that our method can not only converge on different gyroscope bias magnitude, but also accurately initialize the state variables.

5.2. Real Experiments

The popular EuRoC dataset [2] from a micro air vehicle (MAV) is used to verify the algorithms. This dataset contains 11 sequences of different motion patterns collected in two scenes. We sampled 1422 data segments with sufficient motion excitations to exhaustively evaluate the accuracy, robustness, and time-consuming of each algorithm. In the experiments, all algorithms use the same image processing operations, existing features are tracked by the KLT sparse optical flow algorithm [23], and new corner features are detected [34] to maintain 150 points for each image. The outliers are culled using RANSAC with a fundamental matrix model [15] for 1 pixel re-projection error. For the loosely-coupled algorithms, we adopt a general SfM framework to estimate camera motion, which first estimates the initial camera pose with the 5-point algorithm [29] and the PnP solver [19], and then uses bundle adjustment to optimize all poses and point clouds. The max running time of BA is set to 0.2s to fulfill real-time commands [30].

5.2.1 Accuracy evaluation

To verify the accuracy and robustness of our gyroscope bias estimation algorithm, our method is compared with two loosely-coupled methods, VINS-Mono and AS-MLE. Fig. 4 shows that our method significantly outperforms previous methods in almost all sequences. Specifically, the loosely-coupled methods have no results on the V103 and V203 sequences because they are successfully initialized on too few data segments (less than 5) to be statistically significant. This also illustrates the robustness of our gyroscope bias estimation method.

[1] We used the following IMU parameters: Gyroscope and accelerometer continuous-time noise density: $\sigma^g = 1.5e{-}4 \left[rad/(s\sqrt{Hz})\right]$, $\sigma^a = 1.9e^{-4} \left[m/(s^2\sqrt{Hz})\right]$. Gyroscope and accelerometer bias continous-time noise density: $\sigma^{bg} = 1e^{-5} \left[rad/(s^2\sqrt{Hz})\right]$, $\sigma^b a = 1e^{-5} \left[m/(s^3\sqrt{Hz})\right]$.

Table 2. Exhaustive initialization results for 10KFs setting in low, medium, and high angular velocity datasets from EuRoC. For each metric, the best in **red**, the second best in blue.

	Scale RMSE				**Velocity RMSE** (m/s)				**G.Dir RMSE** (°)			
	Low	Medium	High	**Mean**	Low	Medium	High	**Mean**	Low	Medium	High	**Mean**
AS-MLE	0.28	0.35	0.25	0.31	0.16	0.21	0.23	0.18	1.70	3.15	4.16	2.38
CS-VISfM	0.53	0.50	0.41	0.51	0.23	0.24	0.30	0.24	6.10	5.93	6.07	6.03
CS-VISfM-GBE	0.23	0.23	0.07	0.22	0.13	0.13	**0.06**	0.12	**1.18**	1.23	**0.86**	**1.18**
VINS-Mono	0.19	0.23	0.16	0.20	0.11	0.13	0.16	0.12	1.38	1.80	1.60	1.53
DRT-t	0.25	0.22	**0.06**	0.23	0.13	0.13	**0.06**	0.13	1.22	1.26	0.95	1.22
DRT-l	**0.15**	**0.15**	0.07	**0.15**	**0.09**	**0.10**	0.07	**0.09**	1.20	**1.22**	0.97	1.19
DRT-l-wo-GBE	0.48	0.46	0.51	0.48	0.22	0.24	0.28	0.23	5.92	5.68	5.78	5.83

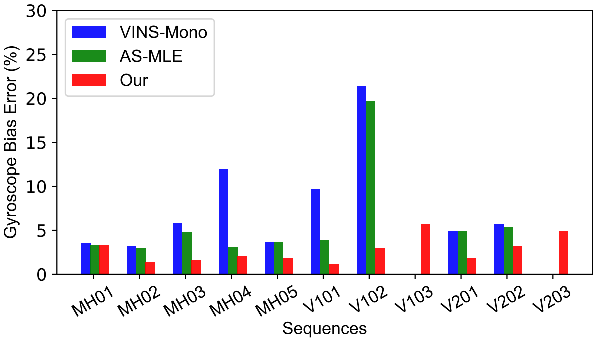

Figure 4. Gyroscope bias errors on EuRoC sequences. Loosely-coupled methods are difficult to initialize successfully on V103 and V203.

In the error statistics of other initial state variables such as scale factor and gravity vector, we classified the 1422 data segments according to the magnitude of angular velocity, including 638 low-speed data segments ($||\omega|| < 15°/s$), 327 high-speed data segments ($||\omega|| > 30°/s$), and 457 medium-speed data segments. Please refer to the supplementary material Sec.3 for the results of separate statistics for the 11 sequences. From Tab. 2, it can be seen that DRT-l significantly outperforms state-of-the-art initialization methods on almost all the motion scenarios, which verifies the effectiveness of our proposed framework. Specifically, comparing the two tightly coupled methods CS-VISfM-GBE and DRT-t, it can be found that the accuracy difference is marginal. In fact, the difference between the two methods is whether to introduce 3D point coordinates when constructing the constraint equation. Comparing VINS-Mono and DRT-t, it can be seen that the method of decoupling rotation and translation can estimate the gravity vector more accurately. However, in the tightly coupled method, the velocity and pose of each keyframe are calculated by integrating the accelerometer data from the initial moment, and the noise accumulated by the integration makes the accuracy of the scale factor and the velocity lower than that of the loosely coupled method. DRT-l combines the advantages of high rotation accuracy obtained by the decoupling method and the advantage of not requiring long-time integration of accelerometer data by the loosely coupled method, making it the best overall performance. It should be noted that compared with VINS-Mono, the LiGT constraint used to solve translation in DRT-l is not more accurate than the pose solved by SfM [3], so the core of the accuracy improvement is the higher rotation accuracy estimated by the decoupling method. The main contribution of the LiGT constraint is computational efficiency. Finally, by comparing the results of CS-VISfM against CS-VISfM-GBE and DRT-l-wo-GBE against DRT-l, we can find that their performance degrades significantly when the visual information is not used to remove the gyroscope bias. This validates the necessity of estimating the gyroscope bias and also illustrates the importance of accurate rotation estimation. For qualitative analysis, we visualize a dataset with the trajectory in Fig. 5. The successful initialization rates and accuracy of DRT outperform the other algorithms, which intuitively illustrates the superiority of our algorithm in different motion modes (e.g., rapid rotation).

5.2.2 Robustness evaluation

The robustness experiments are divided into two categories. One is the histogram distribution of the error, the more statistics on small errors, the better the robustness of the system. The other is the proportion of successful initialization on low-latency data segments. For the evaluation of lower latency initialization, the number of keyframes is reduced from 10 KFs to 5 KFs ($\approx 1s$), so as to reduce image observations and motion excitation. In Fig. 6, we plot the distribution of the percents for the scale error, velocity error, and gravity error metrics. It can be seen that no matter the 10KFs test results in the first row or the 5KFs test results in

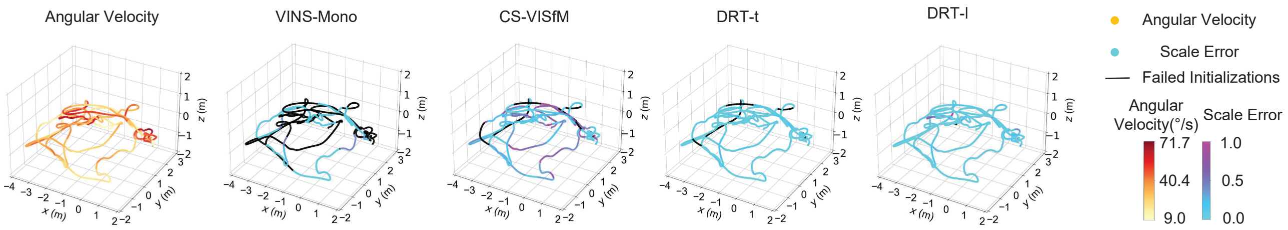

Figure 5. Angular velocity and scale error visualizations for the V202 dataset. **Left:** Trajectory colored by angular velocity magnitude. **Right:** Segments of poses colored by scale error magnitude for each initialization window in the dataset (lighter is better). Segments colored black indicate failed initializations for the respective methods.

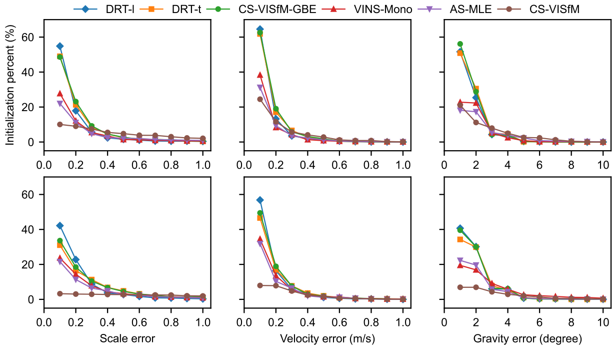

Figure 6. Distribution plots of successful percentages for primary error metrics. **First row**: Results with 10 keyframes. **Second row**: Results with 5 keyframes. For each plot, the X axis denotes the threshold for the error metric and the Y axis shows the fraction of initialization sequences with the respective error metric belonging to the threshold boundary on the X axis.

Table 3. Average initialization computation duration of EuRoC for 10KFs setting in milliseconds. The time consumption of SfM, gyroscope bias optimizer, velocity and gravity estimator, and the point triangulation module are calculated.

Module	AS-MLE	CS-VISfM	VINS-Mono	DRT-t	DRT-l
SfM	30.30	-	30.35	-	-
Bg Est.	0.15	-	0.44	1.95	1.94
Vel&Grav Est.	0.08	279.95	0.14	2.81	1.53
Point Tri.	0.01	-	0.01	0.42	0.41
Total Cost	30.54	279.95	30.94	5.18	3.89

the second row, the number of initialization sequences with DRT-l is the largest in the small-errors range, which demonstrates the robustness of our rotation and translation decoupled method in different scenarios (e.g., fast motion and low latency). The results of CS-VISfM are significantly better than CS-VISfM-GBE, which shows that our rotation-only optimizer plays a corner-stone role in robustness and accuracy. Comparing CS-VISfM-GBE and DRT-t, their performance difference in robustness is still tiny as they are in the accuracy evaluation.

5.2.3 Running time evaluation

To show the time-consuming details of each algorithm, the time-consuming of each module is counted separately, such as SfM, gyroscope bias optimizer, velocity and gravity estimator, and point triangulation. Since our method does not need to compute point clouds, in order to be consistent with other methods, we triangulate all observations after initializing the IMU variables. The point cloud triangulation module of the loosely-coupled method refers to scaling all point clouds with the estimated scale factor.

The computation cost (in milliseconds) of different initialization methods for the 10KFs setting is shown in Tab. 3. We can observe that the initialization speed of DRT-l is the fastest, and it only takes 3.89 ms, which is 72 times faster than the tightly-coupled method CS-VISfM and 8 times faster than the loosely-coupled method VINS-Mono. CS-VISfM needs to solve the large dimensional matrix containing the initial variables and the position of the points. The loosely-coupled method uses the SfM module to estimate the visual poses, which is the source of the most time-consuming in the initialization process.

6. Conclusion

This paper proposes a rotation and translation decoupled solution for visual-inertial initialization. A new formulation for optimizing gyroscope bias directly using visual observations and inertial information is derived, and a globally optimal solver for initial velocity and gravity vectors without estimating 3D point clouds is proposed. Extensive experiments demonstrate that our method is computationally efficient while achieving significant improvements in accuracy and robustness. However, our method ignores the effect of accelerometer bias, and modeling the accelerometer bias in translation constraints is our future work.

References

[1] Sameer Agarwal, Keir Mierle, and Others. Ceres solver. http://ceres-solver.org. 4

[2] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, 2016. 6

[3] Qi Cai, Lilian Zhang, Yuanxin Wu, Wenxian Yu, and Dewen Hu. A pose-only solution to visual reconstruction and navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 5, 7

[4] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 2021. 2

[5] Carlos Campos, José MM Montiel, and Juan D Tardós. Inertial-only optimization for visual-inertial initialization. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 51–57. IEEE, 2020. 1, 3, 6

[6] Andrea Censi. An icp variant using a point-to-line metric. In *2008 IEEE International Conference on Robotics and Automation*, pages 19–25. Ieee, 2008. 5

[7] Nikolaus Demmel, David Schubert, Christiane Sommer, Daniel Cremers, and Vladyslav Usenko. Square root marginalization for sliding-window bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13260–13268, 2021. 2

[8] Javier Domínguez-Conti, Jianfeng Yin, Yacine Alami, and Javier Civera. Visual-inertial slam initialization: A general linear formulation and a gravity-observing non-linear optimization. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 37–45. IEEE, 2018. 2, 3

[9] Tue-Cuong Dong-Si and Anastasios I Mourikis. Estimator initialization in vision-aided inertial navigation with unknown camera-imu calibration. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1064–1071. IEEE, 2012. 2

[10] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 3

[11] Georgios Evangelidis and Branislav Micusik. Revisiting visual-inertial structure-from-motion for odometry and slam initialization. *IEEE Robotics and Automation Letters*, 6(2):1415–1422, 2021. 3

[12] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. On-manifold preintegration for real-time visual–inertial odometry. *IEEE Transactions on Robotics*, 33(1):1–21, 2016. 3

[13] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 15–22. IEEE, 2014. 3

[14] Patrick Geneva, Kevin Eckenhoff, Woosik Lee, Yulin Yang, and Guoquan Huang. Openvins: A research platform for visual-inertial estimation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4666–4672. IEEE, 2020. 1, 2, 6

[15] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 6

[16] Jacques Kaiser, Agostino Martinelli, Flavio Fontana, and Davide Scaramuzza. Simultaneous state initialization and gyroscope bias calibration in visual inertial aided navigation. *IEEE Robotics and Automation Letters*, 2(1):18–25, 2016. 3

[17] Laurent Kneip and Simon Lynen. Direct optimization of frame-to-frame rotation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2352–2359, 2013. 2, 3, 4

[18] Seong Hun Lee and Javier Civera. Rotation-only bundle adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 424–433, 2021. 2

[19] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155–166, 2009. 6

[20] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015. 1

[21] Xin Li, Yijia He, Jinlong Lin, and Xiao Liu. Leveraging planar regularities for point line visual-inertial odometry. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5120–5127. IEEE, 2020. 2

[22] Haomin Liu, Mingyu Chen, Guofeng Zhang, Hujun Bao, and Yingze Bao. Ice-ba: Incremental, consistent and efficient bundle adjustment for visual-inertial slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1974–1982, 2018. 2

[23] Bruce D Lucas, Takeo Kanade, et al. *An iterative image registration technique with an application to stereo vision*, volume 81. Vancouver, 1981. 6

[24] Agostino Martinelli. Vision and imu data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination. *IEEE Transactions on Robotics*, 28(1):44–60, 2011. 2

[25] Agostino Martinelli. Closed-form solution of visual-inertial structure from motion. *International journal of computer vision*, 106(2):138–152, 2014. 2, 6

[26] Dominik Muhle, Lukas Koestler, Nikolaus Demmel, Florian Bernard, and Daniel Cremers. The probabilistic normal epipolar constraint for frame-to-frame rotation optimization under uncertain feature positions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1819–1828, 2022. 2

[27] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 3

[28] Raúl Mur-Artal and Juan D Tardós. Visual-inertial monocular slam with map reuse. *IEEE Robotics and Automation Letters*, 2(2):796–803, 2017. 1, 2, 3

[29] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):756–770, 2004. 6

[30] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 1, 2, 6

[31] Tong Qin and Shaojie Shen. Robust initialization of monocular visual-inertial estimation on aerial robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4225–4232. IEEE, 2017. 3, 5

[32] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 3

[33] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 3

[34] Jianbo Shi et al. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, pages 593–600. IEEE, 1994. 6

[35] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991. 6

[36] Vladyslav Usenko, Nikolaus Demmel, David Schubert, Jörg Stückler, and Daniel Cremers. Visual-inertial mapping with non-linear factor recovery. *IEEE Robotics and Automation Letters*, 5(2):422–429, 2019. 1

[37] Kejian Wu, Ahmed M Ahmed, Georgios A Georgiou, and Stergios I Roumeliotis. A square root inverse filter for efficient vision-aided inertial navigation on mobile devices. In *Robotics: Science and Systems*, volume 2. Rome, Italy, 2015. 2

[38] Ji Zhao. An efficient solution to non-minimal case essential matrix estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1777–1792, 2020. 2

[39] Yunwen Zhou, Abhishek Kar, Eric Turner, Adarsh Kowdle, Chao X Guo, Ryan C DuToit, and Konstantine Tsotsos. Learned monocular depth priors in visual-inertial initialization. *arXiv preprint arXiv:2204.09171*, 2022. 3

[40] David Zuñiga-Noël, Francisco-Angel Moreno, and Javier Gonzalez-Jimenez. An analytical solution to the imu initialization problem for visual-inertial systems. *IEEE Robotics and Automation Letters*, 6(3):6116–6122, 2021. 3, 6