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

Stereo-NEC: Enhancing Stereo Visual-Inertial SLAM Initialization with Normal Epipolar Constraints

Weihan Wang[a,b], Chieh Chou[b], Ganesh Sevagamoorthy[b], Kevin Chen[b], Zheng Chen[b], Ziyue Feng[b], Youjie Xia[b], Feiyang Cai[c], Yi Xu[b], Philippos Mordohai[a]

***Abstract*— We propose an accurate and robust initialization approach for stereo visual-inertial SLAM systems. Unlike the current state-of-the-art method, which heavily relies on the accuracy of a pure visual SLAM system to estimate inertial variables without updating camera poses, potentially compromising accuracy and robustness, our approach offers a different solution. We realize the crucial impact of precise gyroscope bias estimation on rotation accuracy. This, in turn, affects trajectory accuracy due to the accumulation of translation errors. To address this, we first independently estimate the gyroscope bias and use it to formulate a maximum a posteriori problem for further refinement. After this refinement, we proceed to update the rotation estimation by performing IMU integration with gyroscope bias removed from gyroscope measurements. We then leverage robust and accurate rotation estimates to enhance translation estimation via 3-DoF bundle adjustment. Moreover, we introduce a novel approach for determining the success of the initialization by evaluating the residual of the normal epipolar constraint. Extensive evaluations on the EuRoC dataset illustrate that our method excels in accuracy and robustness. It outperforms ORB-SLAM3, the current leading stereo visual-inertial initialization method, in terms of absolute trajectory error and relative rotation error, while maintaining competitive computational speed. Notably, even with 5 keyframes for initialization, our method consistently surpasses the state-of-the-art approach using 10 keyframes in rotation accuracy. The open source code is available at `https://github.com/ApdowJN/Stereo-NEC.git`.**

I. INTRODUCTION

The fusion of cameras and Inertial Measurement Units (IMUs) in Visual-Inertial Simultaneous Localization and Mapping (VI-SLAM) presents a cost-effective, low-power solution for robot perception and AR/VR applications. Cameras offer a rich environment representation, while IMUs measure acceleration and angular velocity, ensuring robustness in fast-motion and texture-less scenes. This synergy makes them ideal complements. Compared to monocular VI-SLAM systems, stereo VI-SLAM systems offer the advantages of a baseline with known scale and the capability to reconstruct 3D geometry even without camera motion.

Initialization in VI-SLAM systems is critical because it impacts their accuracy and robustness. VI-SLAM depends on reliable and precise initial estimates for scale, the gravity direction, initial velocity, acceleration, and gyroscope biases. However, accomplishing this task is challenging, demanding a swift and accurate recovery of observable parameters from visual and inertial data without prior knowledge.

Compared with the extensive research on monocular systems, there are relatively few VIO solutions designed for stereo systems [1]–[7].This is due to the increased computational demands of processing multiple images and stereo matching. Similar to monocular VI-SLAM initialization methods [8]–[17], methods for stereo VI-SLAM initialization are also categorized into two types: joint approaches [12], [18] and disjoint approaches [6], [7], [19], [20]. Joint approaches handle both visual and inertial parameters together by fusing visual observation and IMU integration. However, they tend to overlook the gyroscope bias in the closed-form solution, which results in limited accuracy, while they are computationally expensive. On the other hand, disjoint approaches first independently solve the Structure-from-Motion (SfM) problem and then derive inertial parameters based on camera poses from a pure visual SLAM system. Therefore, the accuracy of these methods relies heavily on the performance of pure visual SLAM. Previous monocular VI-SLAM approaches have been extended to stereo VI-SLAM using a similar disjoint initialization strategy. For instance, VINS-Fusion [6], an extension of VINS-Mono [16], follows this approach with a slight difference. VINS-Fusion jointly estimates velocity, gravity vector, and scale through visual-inertial bundle adjustment, rather than treating them separately. Huang et al. [19] extended their prior method [15] to stereo VI-SLAM by introducing an additional scale estimate. Similarly, ORB-SLAM3 [7] applied the same idea [17] to their stereo VI-SLAM system.

The accuracy of pure visual SLAM greatly impacts the performance of disjoint methods. However, even in state-of-the-art stereo VI-SLAM systems like ORB-SLAM3, accurate camera trajectory estimation is assumed in scenarios with adequate baseline between consecutive frames and mild rotation. Nevertheless, in challenging situations such as pure or intense rotation, ORB-SLAM3's initialization may result in reduced accuracy and robustness.

To overcome these limitations and enhance initialization accuracy and robustness in challenging scenarios, we propose Stereo-NEC, which leverages insights from our previous work [9] that takes into account the significant impact of gyroscope bias estimation on rotation accuracy and considers the connection between inertial parameters and visual

[a]Stevens Institute of Technology, Hoboken, NJ, USA, 07030, {wwang103,pmordoha}@stevens.edu
[b]OPPO US Research Center, Palo Alto, CA, USA, 94303, {weihan.wang, chieh.chou, ganesh.sevagamoorthy, kevin.chen, zheng.chen, ziyue.feng, youjie.xia, yi.xu}@oppo.com
[c]Stony Brook University, Stony Brook, NY 11794, feiyang.cai@stonybrook.edu
This research has been supported in part by the National Science Foundation under award 2024653.

observations. Our approach begins by obtaining accurate rotation estimates, which rely on precise gyroscope bias estimation. This, in turn, plays a crucial role in improving trajectory accuracy by reducing accumulation of translation errors. To achieve accurate gyroscope bias estimation, we extend the concept of monocular normal epipolar constraints (MNEC) [8], [21], [22]. MNEC incorporates normal vectors with relative rotation estimation and visual observations, while rotation is approximated using a first-order Taylor series that accounts for the influence of gyroscope bias. Once we have accurate rotation estimates after removing the gyroscope bias, we leverage them to enhance translation estimation by 3-DoF Bundle adjustment.

The main contributions of the proposed initialization method are:

- Proposing a new method that utilizes stereo normal epipolar constraints to estimate initial gyroscope bias and uses the latter to initialize a maximum a posteriori (MAP) problem for further refinement.
- Enhancing initialization accuracy and robustness by estimating the rotation separately through IMU rotation integration and then utilizing precise and reliable rotation estimates to enhance translation estimation via 3-DoF bundle adjustment.
- Introducing a novel approach to assess initialization success by evaluating the residual of the normal epipolar constraint.

II. PRELIMINARIES

A. Notation

In this paper, we adopt the following notation: the world frame, body frame, and camera frame are represented by $(\cdot)^{\mathrm{w}}$, $(\cdot)^{\mathrm{b}}$, and $(\cdot)^{\mathrm{c}}$, respectively. For stereo cameras, $\mathrm{c_L}$ and $\mathrm{c_R}$ represent the left and right camera, respectively. We denote rotation matrices with $\mathbf{R}$, velocity vectors with $\mathbf{v}$, translation vectors with $\mathbf{t}$, and the gravity vector with $\mathbf{g} = (0, 0, G)^\top$ and G is the magnitude of gravity. $\mathbf{R}_i^j$ denotes relative rotation from frame i to frame j, and $\mathbf{t}_i^j$ denotes relative translation. $\mathrm{b_k}$ is the body frame while taking the k-th image, and $\mathrm{c_k}$ is the left camera frame while taking the k-th image. Acceleration bias and gyroscope bias in the local body frame are represented by $\mathbf{b}_{\mathrm{a}}$ and $\mathbf{b}_{\mathrm{g}}$ respectively. $\boldsymbol{\alpha}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}}$, $\boldsymbol{\beta}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}}$, $\boldsymbol{\gamma}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}}$ represent preintegration of translation, velocity, and rotation from $\mathrm{b_k}$ to $\mathrm{b_{k+1}}$. $\Delta \mathrm{t_{k,k+1}}$ denotes the interval from time k to time k+1. λ represents an eigenvalue of a matrix, and $\lambda_{\min}$ specifically denotes the smallest eigenvalue. $\mathbf{n}$ refers to a normal vector. N represents the number of keyframes used for initialization.

B. Monocular Normal Epipolar Constraint

The monocular normal epipolar constraint (MNEC) [23] encodes the geometric relationship between bearing vectors and the normals of the corresponding epipolar planes defined for two poses of a single mobile camera. As shown in Fig. 1, when considering feature correspondences in two consecutive times, pairs of unit bearing vectors (f_i and f_i') originate from the optical centers ($\mathrm{O_k}$ and $\mathrm{O_{k+1}}$) at time instances k and k+1, respectively, and point towards the 3D points ($\mathbf{p}_i$). Each pair of bearing vectors define an epipolar plane along with its associated normal vector $\mathbf{n}_i = f_i \times \mathbf{R}_{\mathrm{c_{k+1}}}^{\mathrm{c_k}} f_i'$. The intersection of these epipolar planes forms a line, which is the baseline $\mathbf{t}_{\mathrm{c_{k+1}}}^{\mathrm{c_k}}$ between two frames (marked by the dashed line). The normal vectors collectively form the corresponding epipolar normal plane, which is perpendicular to the baseline $\mathbf{t}_{\mathrm{c_{k+1}}}^{\mathrm{c_k}}$. Assuming that n 3D points are observed, we stack the n normal vectors of epipolar planes into a matrix $\mathbf{N} = [\mathbf{n}_1 ... \mathbf{n}_n]$. The requirement for coplanarity is mathematically equivalent to the minimum eigenvalue of the matrix $\mathbf{M} = \mathbf{N}\mathbf{N}^\top$ being zero.

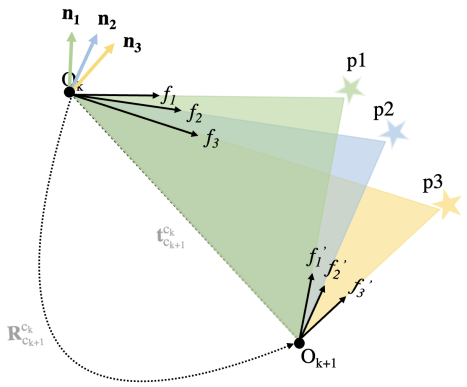

Fig. 1: 3D points ($\mathbf{p}_i$) are represented with different colored five-pointed stars. For each pair of bearing vectors, an epipolar plane is formed (highlighted in green, blue, yellow), and their corresponding normal vectors ($\mathbf{n}_i$) are shown in corresponding colors. The normal vectors lie in a same plane. The baseline $\mathbf{t}_{\mathrm{c_{k+1}}}^{\mathrm{c_k}}$ intersects all corresponding epipolar planes.

The residual of the MNEC is given by:

$$e_i = \left| \mathbf{n}_i^\top \mathbf{t}_{\mathrm{c_{k+1}}}^{\mathrm{c_k}} \right|.$$

There are two applications of MNEC:

1) Rotation: Kneip and Lynen [21] aim to determine the relative rotation $\mathbf{R}_{\mathrm{c_{k+1}}}^{\mathrm{c_k}}$ that minimizes the smallest eigenvalue $\lambda_{\min}(\mathbf{M}_{k,k+1})$:

$$\begin{aligned} {\mathbf{R}_{\mathrm{c_{k+1}}}^{\mathrm{c_k}}}^* &= \arg\min_{\mathbf{R}_{\mathrm{c_{k+1}}}^{\mathrm{c_k}}} \lambda_{\min}(\mathbf{M}_{\mathrm{k,k+1}}), \\ \mathbf{M}_{\mathrm{k,k+1}} &= \sum_{i=1}^{n} (f_i \times \mathbf{R}_{\mathrm{c_{k+1}}}^{\mathrm{c_k}} f_i')(f_i \times \mathbf{R}_{\mathrm{c_{k+1}}}^{\mathrm{c_k}} f_i')^\top \end{aligned} \tag{1}$$

The matrix $\mathbf{M}_{k,k+1}$ possesses specific properties: it is real, symmetric, and positive semi-definite. Furthermore, due to the coplanarity constraint on the normal vectors, its rank is 2. Solving Eq. (1) is achieved using the Levenberg-Marquardt algorithm.

2) Gyroscope Bias: Inspired by Kneip and Lynen's work, He et al. [22] employ the MNEC to directly optimize gyroscope bias, incorporating image observations and the camera-IMU extrinsic calibration matrix $[\mathbf{R}_{\mathrm{b}}^{\mathrm{c}} | \mathbf{t}_{\mathrm{b}}^{\mathrm{c}}]$. This modifies the

objective function from Eq. (1) to Eq. (2):

$$\begin{aligned}
\mathbf{b}_{\mathrm{g}}^{*} &= \arg\min_{\mathbf{b}_{\mathrm{g}}} \lambda_{\min}(\mathbf{M}_{k,k+1}), \\
\mathbf{n}_i &= f_i \times \mathbf{R}_{\mathrm{b}}^{\mathrm{c}} \hat{\boldsymbol{\gamma}}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}} \mathbf{R}_{\mathrm{c}}^{\mathrm{b}} f_i' \\
\mathbf{M}_{k,k+1} &= \sum_{i=1}^{n} (f_i \times \mathbf{R}_{\mathrm{b}}^{\mathrm{c}} \hat{\boldsymbol{\gamma}}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}} \mathbf{R}_{\mathrm{c}}^{\mathrm{b}} f_i')(f_i \times \mathbf{R}_{\mathrm{b}}^{\mathrm{c}} \hat{\boldsymbol{\gamma}}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}} \mathbf{R}_{\mathrm{c}}^{\mathrm{b}} f_i')^{\top} \\
\hat{\boldsymbol{\gamma}}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}} &= \boldsymbol{\gamma}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}} \mathrm{Exp}(\mathbf{J}_{\mathbf{b}_{\mathrm{g}}}^{\boldsymbol{\gamma}} \mathbf{b}_{\mathrm{g}})
\end{aligned} \tag{2}$$

where, considering the influence of gyroscope bias over the duration between the k-th keyframe and the (k+1)-th keyframe, $\hat{\boldsymbol{\gamma}}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}}$ is estimated using a first-order Taylor approximation of $\boldsymbol{\gamma}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}}$ and $\mathbf{J}_{\mathbf{b}_{\mathrm{g}}}^{\boldsymbol{\gamma}}$ represents how the preintegration changes due to a small difference in gyroscope bias estimation.

III. Proposed Approach

Our method is motivated by the realization that precise gyroscope bias estimation significantly influences rotation accuracy, ultimately affecting trajectory accuracy due to the accumulation of translation errors. Building upon our previous work [9], which employs an Error-state Kalman Filter (ESKF) to estimate gyroscope bias and corrects rotation estimation, we reduce the reliance on pure visual SLAM for rotation estimation. The underlying idea of our method is rooted in realizing the importance of gyroscope bias. We start by independently estimating gyroscope bias, which then is employed to formulate a MAP problem for further refinement. Following this refinement, we proceed to update the rotation estimation. With a robust and accurate rotation estimation in place, we leverage it to assist translation estimation. Our method consists of five steps aimed at deriving precise initial values for keyframes' poses and velocities, gravity direction, and IMU biases:

- **Step 0. Pure Visual SLAM**: Obtain the initial keyframe poses from stereo visual-only SLAM.
- **Step 1. Eigenvalue-based Gyroscope Bias Estimator**: Derive the initial gyroscope bias by formulating an eigenvalue minimization problem using both visual, and gyroscope measurements.
- **Step 2. Gyroscope Bias Refinement, Acceleration Bias, Velocity and Gravity Estimator**: Refine gyroscope bias and estimate keyframes' velocities, gravity direction and acceleration bias by solving an inertial-only MAP estimation problem.
- **Step 3. Rotation-Translation-Decoupled Optimization**: Update the camera rotation estimate by integrating gyroscope measurements with the gyroscope bias removed, and optimize the camera translation using 3-DoF BA.
- **Step 4. Joint Visual-Inertial Bundle Adjustment**: Utilize the solution derived from the previous steps as the initial estimate via visual-inertial MAP to obtain optimal estimates for keyframes' poses, keyframes' velocities, 3D points, gravity direction, and IMU biases.

In the following subsections, we provide a more detailed explanation of our initialization steps.

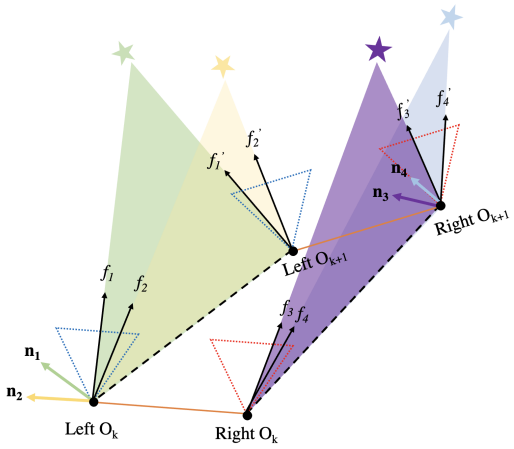

Fig. 2: An illustration depicting the geometry of the stereo normal epipolar constraint is shown. The blue and red dashed triangles represent the left and right cameras, respectively. Left O_k and Left O_{k+1} represent the two left optical centers at time k and k+1 respectively. Similarly, Right O_k and Right O_{k+1} correspond to the two right optical centers. The orange line represents the baseline of the stereo cameras. The temporal epipolar planes of the left camera are colored in green and yellow, while the temporal epipolar planes of the right camera are colored in blue and purple. Each corresponding normal vector ($\mathbf{n}_i$) of each temporal epipolar plane is depicted in corresponding colors. Only normal vectors from the same camera are coplanar.

A. Eigenvalue-based Gyroscope Bias Estimator

As shown in Fig. 2, the stereo normal epipolar constraint extends the monocular normal epipolar constraint, thereby enabling the incorporation of additional visual observations to enhance the robustness and accuracy of gyroscope bias estimation. To estimate the initial gyroscope bias while leveraging stereo observation information, we initialize the gyroscope bias by minimizing the smallest eigenvalue $\lambda_{\min}$ in Eq. (3):

$$\begin{aligned}
\mathbf{b}_{\mathrm{g}}^{*} &= \arg\min_{\mathbf{b}_{\mathrm{g}}} \lambda_{\min}, \\
\lambda_{\min} &= \sum_{(\mathrm{k},\mathrm{k+1})\in\mathcal{E}} (\lambda_{\min}({}^{\mathrm{L}}\mathbf{M}_{\mathrm{k,k+1}}) + \lambda_{\min}({}^{\mathrm{R}}\mathbf{M}_{\mathrm{k,k+1}})) \\
{}^{\mathrm{L}}\mathbf{M}_{\mathrm{k,k+1}} &= \sum_{i=1}^{n_{\mathrm{L}}} (f_{\mathrm{L}_i} \times \mathbf{R}_{\mathrm{b}}^{\mathrm{c_L}} \hat{\boldsymbol{\gamma}}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}} \mathbf{R}_{\mathrm{c_L}}^{\mathrm{b}} f_{\mathrm{L}_i}')(f_{\mathrm{L}_i} \times \mathbf{R}_{\mathrm{b}}^{\mathrm{c_L}} \hat{\boldsymbol{\gamma}}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}} \mathbf{R}_{\mathrm{c_L}}^{\mathrm{b}} f_{\mathrm{L}_i}')^{\top} \\
{}^{\mathrm{R}}\mathbf{M}_{\mathrm{k,k+1}} &= \sum_{i=1}^{n_{\mathrm{R}}} (f_{\mathrm{R}_i} \times \mathbf{R}_{\mathrm{b}}^{\mathrm{c_R}} \hat{\boldsymbol{\gamma}}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}} \mathbf{R}_{\mathrm{c_R}}^{\mathrm{b}} f_{\mathrm{R}_i}')(f_{\mathrm{R}_i} \times \mathbf{R}_{\mathrm{b}}^{\mathrm{c_R}} \hat{\boldsymbol{\gamma}}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}} \mathbf{R}_{\mathrm{c_R}}^{\mathrm{b}} f_{\mathrm{R}_i}')^{\top} \\
\hat{\boldsymbol{\gamma}}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}} &= \boldsymbol{\gamma}_{\mathrm{b_{k+1}}}^{\mathrm{b_k}} \mathrm{Exp}(\mathbf{J}_{\mathbf{b}_{\mathrm{g}}}^{\boldsymbol{\gamma}} \mathbf{b}_{\mathrm{g}})
\end{aligned} \tag{3}$$

where $(f_{\mathrm{L}_i}, f_{\mathrm{L}_i}')$ denotes a pair of bearing vectors in the left camera and $(f_{\mathrm{R}_i}, f_{\mathrm{R}_i}')$ denotes a pair of bearing vectors in the right camera. n_{L} and n_{R} denote the number of features observed by the left and right camera. $\mathbf{R}_{\mathrm{c_L}}^{\mathrm{b}}$ and $\mathbf{R}_{\mathrm{c_R}}^{\mathrm{b}}$ represent the rotation matrix of the left camera-IMU extrinsic calibration and the rotation matrix of the right camera-IMU extrinsic calibration, respectively. ${}^{\mathrm{L}}\mathbf{M}_{\mathrm{k,k+1}}$ and ${}^{\mathrm{R}}\mathbf{M}_{\mathrm{k,k+1}}$ are formed by using the normals from the left and right cameras. $\mathcal{E}$ signifies a set of keyframe pairs at two consecutive times.

B. Gyroscope Bias Refinement, Acceleration Bias, Velocity and Gravity Estimator

This step aims to attain optimal estimates of keyframes' velocities, gravity direction, and IMU biases using posterior estimation with the initial gyroscope bias from Step 1. The estimates from this step are as follows:

$$\mathcal{X} = \left[\mathbf{v}^{\mathrm{w}}_{\mathrm{b}_0:\mathrm{b}_{N-1}},\ \mathbf{R}^{\mathrm{w}}_{\mathrm{g}},\ \mathbf{b}_{\mathrm{g}},\ \mathbf{b}_{\mathrm{a}}\right]^{\top}$$

and the posterior distribution is:

$$p(\mathcal{X}|\mathcal{L}_{0:\mathrm{k}\text{-}1}) \propto p(\mathcal{L}_{0:\mathrm{k}\text{-}1}|\mathcal{X})p(\mathcal{X})$$

where $\mathcal{L}_{0:\mathrm{k}}$ refers to the preintegration of inertial measurements between consecutive keyframes, spanning from the first keyframe to the (k-1)-th keyframe, $p(\mathcal{L}_{0:\mathrm{k}\text{-}1}|\mathcal{X})$ represents the likelihood of the inertial measurements given IMU states, and $p(\mathcal{X})$ is the prior of the IMU states. The resulting optimal estimates $\mathcal{X}^*$ can be obtained by maximizing the posterior distribution, which is equivalent to minimizing its negative logarithm, leading to the following conversion:

$$\begin{aligned}\mathcal{X}^* &= \arg\max_{\mathcal{X}} p(\mathcal{X}|\mathcal{L}_{0:\mathrm{k}\text{-}1}) = \arg\min_{\mathcal{X}} \Big(-\log(p(\mathcal{X})) \\ &\quad - \sum_{k=0}^{N-2} \log\left(p(\mathcal{L}_{\mathrm{k},\mathrm{k}+1}|\mathbf{R}^{\mathrm{w}}_{\mathrm{g}}, \mathbf{b}_{\mathrm{g}}, \mathbf{b}_{\mathrm{a}}, \mathbf{v}^{\mathrm{w}}_{\mathrm{b}_{\mathrm{k}}:\mathrm{b}_{\mathrm{k}+1}})\right)\Big) \\ &= \arg\min_{\mathcal{X}} \left(\|\mathbf{r}_p\|^2_{\Sigma_p} + \sum_{k=0}^{N-2} \|\mathbf{r}_{\mathcal{L}_{\mathrm{k},\mathrm{k}+1}}\|^2_{\Sigma_{\mathcal{L}_{\mathrm{k},\mathrm{k}+1}}}\right)\end{aligned} \tag{4}$$

$\mathbf{R}^{\mathrm{w}}_{\mathrm{g}}$ denotes the rotation aligning gravity with the world's z-axis($\mathbf{g}^{\mathrm{w}} = \mathbf{R}^{\mathrm{w}}_{\mathrm{g}}\mathbf{g}$). $\mathbf{r}_p$ and $\mathbf{r}_{\mathcal{L}_{\mathrm{k},\mathrm{k}+1}}$ are the residuals of IMU biases prior and IMU measurements in two consecutive keyframes, and Σ_p and $\Sigma_{\mathcal{L}_{\mathrm{k},\mathrm{k}+1}}$ are the corresponding covariances.

$$\begin{aligned}\mathbf{r}_{\mathcal{L}_{\mathrm{k},\mathrm{k}+1}} &= \left[\delta\boldsymbol{\alpha}^{\mathrm{b}_{\mathrm{k}}}_{\mathrm{b}_{\mathrm{k}+1}},\ \delta\boldsymbol{\beta}^{\mathrm{b}_{\mathrm{k}}}_{\mathrm{b}_{\mathrm{k}+1}},\ \delta\boldsymbol{\gamma}^{\mathrm{b}_{\mathrm{k}}}_{\mathrm{b}_{\mathrm{k}+1}}\right]^{\top} \\ &= \begin{bmatrix} \mathbf{R}^{\mathrm{w}\ \top}_{\mathrm{b}_{\mathrm{k}}}(\mathbf{t}^{\mathrm{w}}_{\mathrm{b}_{\mathrm{k}+1}} - \mathbf{t}^{\mathrm{w}}_{\mathrm{b}_{\mathrm{k}}} - \frac{1}{2}\mathbf{g}^{\mathrm{w}}\Delta\mathrm{t}_{\mathrm{k},\mathrm{k}+1} - \mathbf{v}^{\mathrm{w}}_{\mathrm{b}_{\mathrm{k}}}\Delta\mathrm{t}_{\mathrm{k},\mathrm{k}+1}) - \boldsymbol{\alpha}^{\mathrm{b}_{\mathrm{k}}}_{\mathrm{b}_{\mathrm{k}+1}} \\ \mathbf{R}^{\mathrm{w}\ \top}_{\mathrm{b}_{\mathrm{k}}}(\mathbf{v}^{\mathrm{w}}_{\mathrm{b}_{\mathrm{k}+1}} - \mathbf{g}^{\mathrm{w}}\Delta\mathrm{t}_{\mathrm{k},\mathrm{k}+1} - \mathbf{v}^{\mathrm{w}}_{\mathrm{b}_{\mathrm{k}}}) - \boldsymbol{\beta}^{\mathrm{b}_{\mathrm{k}}}_{\mathrm{b}_{\mathrm{k}+1}} \\ \mathrm{Log}(\boldsymbol{\gamma}^{\mathrm{b}_{\mathrm{k}}}_{\mathrm{b}_{\mathrm{k}+1}}\mathbf{R}^{\mathrm{w}\ \top}_{\mathrm{b}_{\mathrm{k}}}\mathbf{R}^{\mathrm{w}}_{\mathrm{b}_{\mathrm{k}+1}}) \end{bmatrix}\end{aligned}$$

C. Rotation-Translation-Decoupled Optimization

As shown in Fig. 3, we separate the optimization of rotation and translation. This decision stems from our observation that rotation acquired through IMU rotation integration, after removing the gyroscope bias, is more accurate than the rotation derived from pure visual SLAM (refer to the RRE column labeled 'W/O VI-BA' in Table I). Consequently, after attaining optimal gyroscope bias from Step 2, we update each rotation $\mathbf{R}^{\mathrm{w}}_{\mathrm{b}_{\mathrm{k}}}$ within the sliding window via IMU rotation integration and optimize each translation $\mathbf{t}^{\mathrm{c}_{\mathrm{k}}}_{\mathrm{w}}$ with 3-DoF BA:

$$\begin{aligned}\mathbf{R}^{\mathrm{w}}_{\mathrm{c}_{\mathrm{k}}} &= \mathbf{R}^{\mathrm{w}}_{\mathrm{b}_{\mathrm{k}}}\mathbf{R}^{\mathrm{b}}_{\mathrm{c}_{\mathrm{L}}} \\ \mathbf{t}^{\mathrm{c}_{\mathrm{k}}\,*}_{\mathrm{w}} &= \arg\min_{\mathbf{t}^{\mathrm{c}_{\mathrm{k}}}_{\mathrm{w}}} \sum_{i\in\mathcal{M}} \rho\left(\|\mathbf{x}^i - \pi_{(\cdot)}(\mathbf{R}^{\mathrm{c}_{\mathrm{k}}}_{\mathrm{w}}\mathbf{X}^i + \mathbf{t}^{\mathrm{c}_{\mathrm{k}}}_{\mathrm{w}})\|^2_{\Sigma}\right)\end{aligned} \tag{5}$$

$\mathbf{X}^i$ denotes a 3D point in the world frame, obtained through triangulation. and $\mathbf{x}^i$ represents its corresponding 2D feature. ρ is the robust Huber cost function, $\pi_{(\cdot)}$ are reprojection functions (monocular π_m and rectified stereo π_s), and Σ corresponds to the covariance related to the scale level of the keypoints in the pyramid [24].

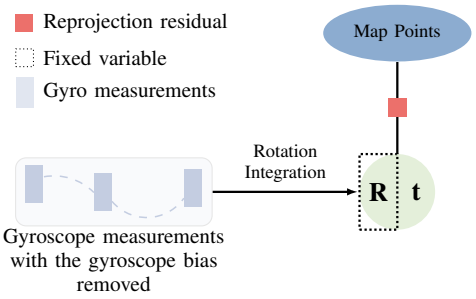

Fig. 3: Step 3: Rotation update via IMU integration, followed by translation optimization using 3-DoF bundle adjustment.

D. Joint Visual-Inertial Bundle Adjustment

After optimizing the rotation $\mathbf{R}^{\mathrm{c}_{\mathrm{k}}\,*}_{\mathrm{w}}$ and translation $\mathbf{t}^{\mathrm{c}_{\mathrm{k}}\,*}_{\mathrm{w}}$ in Step 3, we evaluate the success of the initialization process. Our approach involves computing the average residual of the normal epipolar constraint:

$$\begin{aligned}\bar{e} &= \frac{1}{N}\sum_{\mathrm{k}=0}^{N-1} \bar{e}_{\mathrm{k},\mathrm{k}+1} \\ \bar{e}_{\mathrm{k},\mathrm{k}+1} &= \frac{1}{n}\sum_{i=1}^{n}\left(\left|\mathbf{n}_i^{\top}\mathbf{t}^{\mathrm{c}_{\mathrm{k}}\,*}_{\mathrm{w}}\right|\right) \\ \mathbf{n}_i &= f_i \times \mathbf{R}^{\mathrm{c}_{\mathrm{L}}}_{\mathrm{b}}\hat{\boldsymbol{\gamma}}^{\mathrm{b}_{\mathrm{k}}}_{\mathrm{b}_{\mathrm{k}+1}}\mathbf{R}^{\mathrm{b}}_{\mathrm{c}_{\mathrm{L}}}f'_i\end{aligned}$$

where $\bar{e}_{\mathrm{k},\mathrm{k}+1}$ represents the average residual of the normal epipolar constraint from time k to time k+1. (f_i, f'_i) is a pair of bearing vectors visible to both left and right cameras. N represents the number of keyframes in the sliding window during initialization and n denotes the number of pairs of covisible bearing vectors at two consecutive keyframes. If $\bar{e}$ falls below a certain threshold, we consider the initialization successful and proceed with the application of VI-BA. This choice is made because the previous steps offer not only precise initial estimates to serve as seeds for joint VI-BA, accelerating its convergence, but also expedite the runtime of VI-BA.

IV. Experimental Results

A. Experimental Setup and Implementation

The EuRoC dataset [25] provides precise rotation and translation data for 11 MAV sequences, spanning various flight conditions. It includes synchronized visual-inertial sensor units with global shutter cameras and a MEMS IMU for angular rate and acceleration data. Camera intrinsic and camera-IMU extrinsic parameters are also available. All experiments are conducted on an Intel i9-10920X desktop with 64 GB of RAM. To ensure a fair comparison between the initialization method of ORB-SLAM3 [7] and Stereo-NEC, our method is integrated into ORB-SLAM3. In tackling Eq. (3), quaternions serve as the chosen minimal rotation parameterization. $\mathbf{M}_{\mathrm{k},\mathrm{k}+1}$ is a 3×3 matrix. Eq. (4)

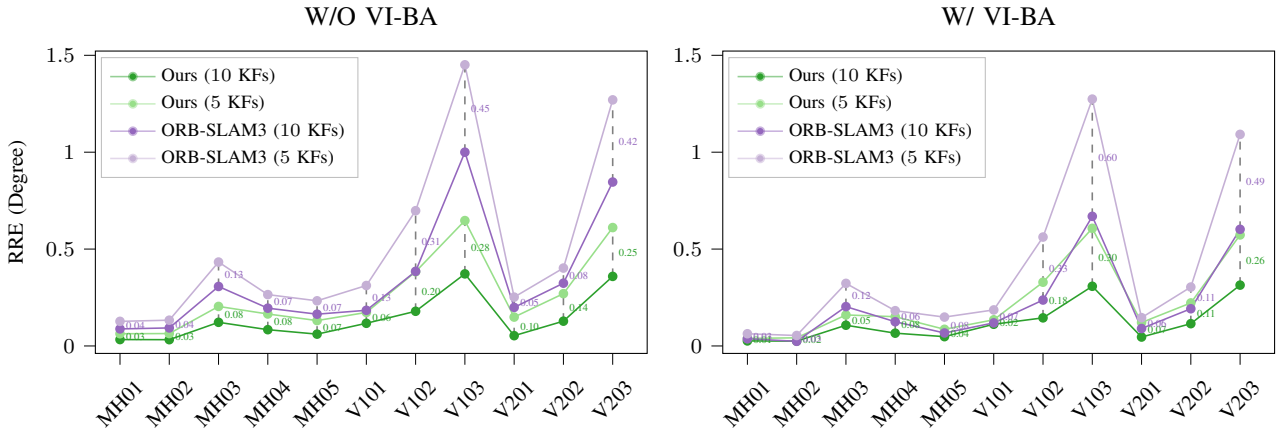

Fig. 4: Relative Rotation Error (RRE) for different methods with different number of keyframes. **Left:** Results with 5 and 10 keyframes without VI-BA for initialization. **Right:** Results with 5 and 10 keyframes with VI-BA applied for initialization.

and Eq. (5) can be solved iteratively using the Levenberg-Marquardt algorithm with analytic derivatives. $\bar{e}$ is set to 10^{-4} and the number of keyframes in the sliding window for initialization is 10. In all our evaluation results, the absolute trajectory error (ATE) is measured in meters and does not involve scale alignment.

B. Accuracy Evaluation

To measure accuracy on different trajectories, we perform an exhaustive initialization test. During this test, we launch an initialization with 10 keyframes every 2.5 seconds for each sequence, leading to the evaluation of 464 different initialization segments. Table I presents a comparison of the absolute trajectory error (ATE) and relative rotation error (RRE) for ORB-SLAM3 and Stereo-NEC, with and without visual-inertial bundle adjustment (VI-BA). Across various sequences, our method consistently outperforms ORB-SLAM3 in terms of both ATE and RRE with the exception of RRE in MH_02_easy, for which the two methods are very closely matched. On average, our method achieves 1.3 times better accuracy in ATE without utilizing VI-BA, compared to ORB-SLAM3. Furthermore, there is an impressive reduction of RRE by a factor of 2.5. When VI-BA is employed, the performance of our method is further improved: ATE drops to 0.014 meters, and RRE reduces to 0.119 degrees. In contrast, under the same conditions, ORB-SLAM3 registers an ATE of 0.018 meters and an RRE of 0.215 degrees.

The extensive results demonstrate that our approach, which involves integrating bias-removed gyroscope measurements for camera rotation updates, stands in contrast to ORB-SLAM3's reliance solely on camera rotation estimation from pure visual SLAM. This contrast leads to substantial improvements in rotation estimation. Moreover, the combination of precise rotation estimates and our translation-only optimization further enhances translation estimation.

C. Robustness Evaluation

We conduct two separate evaluations, each with and without VI-BA, to assess the methods' robustness under various conditions and initialization settings.

Reduced number of keyframes: We investigate how a more robust method would demonstrate a small decrease in RRE when initialized with fewer keyframes. The results are depicted in Fig. 4 which shows the RRE when using 5 or 10 keyframes for initialization.

When transitioning from 10 to 5 keyframes for initialization without utilizing VI-BA, our method experiences an average increase of 0.119 degrees in RRE. ORB-SLAM3 shows a slightly higher increase of 0.163 degrees under the same conditions. On the other hand, when our method employs VI-BA, the RRE increases by a modest average of 0.105 degrees, while ORB-SLAM3 exhibits a larger increase of 0.179 degrees. Notably, in scenes with large rotation (V1_03 difficult and V2_03 difficult), regardless of VI-BA usage, ORB-SLAM3 demonstrates RRE changes 1.7-2.0 times higher than those of our method. Moreover, across all sequences, our method achieves lower RRE, despite utilizing fewer keyframes (5 keyframes) for initialization, compared to ORB-SLAM3's RRE when using 10 keyframes. This observation serves as strong evidence for the precision and robustness of our approach.

Challenging conditions: We test the methods under challenging conditions, such as motion blur and illumination changes, using segments from V2_03_difficult. To do this, we categorize the data segments into three groups based on the 10 keyframes' average angular velocity magnitude $\bar{\omega}$: low-speed ($5°/s \leq \| \bar{\omega} \| < 15°/s$), medium-speed ($15°/s \leq \| \bar{\omega} \| < 30°/s$), and high-speed ($\| \bar{\omega} \| \geq 30°/s$) data segments. The results are presented in Table III.

On average, our method achieves an ATE of 0.035 meters and an RRE of 0.312 degrees without VI-BA, whereas ORB-SLAM3 achieves an ATE of 0.044 meters and an RRE of 0.706 degrees. When utilizing VI-BA, the ATE of our method decreases to 0.028 meters and the RRE to 0.268 degrees, while the ATE of ORB-SLAM3 remains at 0.035 meters and the RRE at 0.446 degrees. Specifically, in high-speed scenarios, ORB-SLAM3 demonstrates 2.4-3.1 times higher RRE and 2.0-2.2 times higher ATE, regardless of whether VI-BA is used or not. These findings highlight the improved performance and robustness of our method compared to

TABLE I: Comparison of the accuracy of Stereo-NEC and ORB-SLAM3, both using 10 keyframes for initialization, with (W/) and without (W/O) VI-BA, in terms of ATE (meters) and RRE (degrees). Each value in the table corresponds to the average of RMSE results from different methods, which launch initialization every 2.5 seconds on the EuRoC dataset.

Seq. Name	ATE (m)				RRE (deg)			
	W/O VI-BA		W/ VI-BA		W/O VI-BA		W/ VI-BA	
	Ours	ORB-SLAM3	Ours	ORB-SLAM3	Ours	ORB-SLAM3	Ours	ORB-SLAM3
MH_01_easy	**0.005**	0.008	**0.005**	0.006	**0.033**	0.088	**0.026**	0.037
MH_02_easy	**0.005**	0.006	**0.004**	**0.004**	**0.032**	0.093	0.025	**0.024**
MH_03_medium	**0.030**	0.032	**0.025**	0.027	**0.122**	0.307	**0.107**	0.203
MH_04_difficult	**0.029**	0.031	**0.018**	0.026	**0.084**	0.195	**0.066**	0.125
MH_05_difficult	**0.020**	0.027	**0.014**	0.016	**0.061**	0.164	**0.048**	0.067
V1_01_easy	**0.007**	0.008	**0.006**	**0.006**	**0.117**	0.184	**0.112**	0.119
V1_02_medium	**0.018**	0.020	**0.012**	0.015	**0.179**	0.385	**0.145**	0.237
V1_03_difficult	**0.040**	0.054	**0.026**	0.043	**0.372**	1.000	**0.308**	0.669
V2_01_easy	**0.004**	0.007	**0.003**	0.004	**0.053**	0.199	**0.046**	0.090
V2_02_medium	**0.013**	0.021	**0.009**	0.014	**0.128**	0.324	**0.115**	0.192
V2_03_difficult	**0.036**	0.047	**0.028**	0.039	**0.359**	0.846	**0.314**	0.602
Avg	**0.019**	0.024	**0.014**	0.018	**0.140**	0.344	**0.119**	0.215

TABLE II: Comparison of average initialization computation time for 10 keyframes setting in milliseconds (ms) on EuRoC. The results involve 10 keyframes, and the best results for each sequence are highlighted in bold.

Seq. Name	Ours & ORB-SLAM3	Ours				ORB-SLAM3			
	Pure Visual SLAM	Bias, Vel & Grav Est	Rot-Int & Trans-Opt	VI-BA	Total Cost	Bias, Vel & Grav Est	Rot-Int & Trans-Opt	VI-BA	Total Cost
MH_01_easy	514.84	291.58	15.55	**57.06**	879.03	**1.97**	-	68.23	**585.04**
MH_02_easy	474.32	311.86	17.53	**58.31**	862.02	**1.95**	-	68.56	**544.83**
MH_03_medium	458.50	292.21	16.37	**56.74**	823.82	**2.10**	-	62.83	**523.43**
MH_04_difficult	466.76	292.73	14.57	**52.33**	826.39	**2.22**	-	66.56	**535.54**
MH_05_difficult	472.50	308.98	17.15	**52.49**	851.12	**2.06**	-	64.21	**538.77**
V1_01_easy	563.99	304.82	22.79	**74.32**	965.92	**1.80**	-	77.42	**643.21**
V1_02_medium	495.96	298.41	16.43	**53.21**	864.01	**1.79**	-	60.86	**558.61**
V1_03_difficult	532.10	252.32	14.69	**49.81**	848.92	**2.05**	-	60.30	**594.45**
V2_01_easy	553.45	331.41	20.75	**73.60**	979.21	**1.90**	-	80.78	**636.13**
V2_02_medium	556.98	300.84	14.52	**52.20**	924.54	**2.06**	-	62.10	**621.14**
V2_03_difficult	484.68	264.52	12.51	**40.54**	802.25	**2.11**	-	57.70	**544.49**
Avg	506.73	295.43	16.62	**56.42**	875.20	**2.0**	-	66.32	**575.05**

TABLE III: Exhaustive initialization results for 10 keyframes with Low, Medium, and High Angular Velocity from V2_03_difficult sequence.

Seq. Name	ATE (m)				RRE (degree)			
	W/O VI-BA		W/ VI-BA		W/O VI-BA		W/ VI-BA	
	Ours	ORB-SLAM3	Ours	ORB-SLAM3	Ours	ORB-SLAM3	Ours	ORB-SLAM3
Low	**0.023**	0.028	0.017	**0.015**	**0.151**	0.378	0.133	**0.085**
Medium	**0.059**	0.059	**0.050**	0.051	**0.465**	0.737	**0.362**	0.513
High	**0.023**	0.045	**0.017**	0.038	**0.319**	1.003	**0.308**	0.741
Avg	**0.035**	0.044	**0.028**	0.035	**0.312**	0.706	**0.268**	0.446

ORB-SLAM3 in high-speed scenarios.

D. Computation Speed Evaluation

In Table II, we present the runtime comparison for each initialization module separately, including pure visual SLAM, IMU bias, velocity and gravity estimation (Bias, Vel & Grav Est), rotation integration, and translation optimization (Rot-Int & Trans-Opt) and VI-BA.

The results reveal that our method is around 10 ms faster on average than ORB-SLAM3 in VI-BA because it provides accurate initial estimates which aid faster convergence. However, our method takes 300.15 milliseconds longer on average for initialization compared to ORB-SLAM3. This is due to two additional steps in our method: 1) We first estimate the initial gyroscope bias before estimating keyframes' velocities, gravity direction, and acceleration bias, while ORB-SLAM3's Inertial-only step simultaneously estimates velocities, gravity direction, and IMU biases. 2) After obtaining the gyroscope bias, we refine the camera rotation estimation by integrating gyroscope measurements with the gyroscope bias removed, and we update the camera translation using a 3-DoF bundle adjustment, while ORB-SLAM3 does not update the camera pose from pure visual SLAM.

These two additional steps are indispensable, especially when dealing with a larger gyroscope bias IMU. Enhancing pose estimation has been demonstrated by updating the camera rotation through the integration of gyroscope bias-removed measurements and the camera translation via 3-DoF bundle adjustment with updated rotation estimation, as shown in IV-B. The additional 300 ms are only required during initialization and can be considered negligible for trajectories that last even a few seconds.

V. CONCLUSIONS

Our proposed method, Stereo-NEC, addresses limitations in the current state-of-the-art approach, ORB-SLAM3, which heavily relies on the accuracy of pure visual SLAM to estimate inertial variables without initial gyroscope bias estimation in inertial-only optimization. We independently estimate the gyroscope bias, then use it to refine other parameters through a maximum a posteriori problem. After this, we update rotation estimation via IMU integration with the gyroscope bias removed, enhancing translation estimation through 3-DoF bundle adjustment with updated rotation estimation. We also introduce a novel approach to determine initialization success by evaluating the residual of the normal epipolar constraint. As a result, our method improves both accuracy and robustness compared to ORB-SLAM3, while maintaining competitive computation speed.

REFERENCES

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

[2] T. Manderson, F. Shkurti, and G. Dudek, "Texture-aware slam using stereo imagery and inertial information," in *13th Conference on Computer and Robot Vision (CRV)*, 2016, pp. 456–463.

[3] M. K. Paul, K. Wu, J. A. Hesch, E. D. Nerurkar, and S. I. Roumeliotis, "A comparative analysis of tightly-coupled monocular, binocular, and stereo vins," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 165–172.

[4] V. Usenko, J. Engel, J. Stückler, and D. Cremers, "Direct visual-inertial odometry with stereo cameras," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1885–1892.

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

[6] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," *arXiv preprint arXiv:1901.03642*, 2019.

[7] C. Campos, R. Elvira, J. J. Gómez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[8] L. Kneip, A. Martinelli, S. Weiss, D. Scaramuzza, and R. Siegwart, "Closed-form solution for absolute scale velocity determination combining inertial measurements and a single feature correspondence," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 4546–4553.

[9] W. Wang, J. Li, Y. Ming, and P. Mordohai, "EDI: ESKF-based Disjoint Initialization for Visual-Inertial SLAM Systems," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[10] A. Martinelli, "Closed-form solution of visual-inertial structure from motion," *International Journal of Computer Vision (IJCV)*, vol. 106, no. 2, pp. 138–152, 2014.

[11] J. Kaiser, A. Martinelli, F. Fontana, and D. Scaramuzza, "Simultaneous state initialization and gyroscope bias calibration in visual inertial aided navigation," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 18–25, 2017.

[12] T. Dong-Si and A. I. Mourikis, "Estimator initialization in vision-aided inertial navigation with unknown camera-IMU calibration," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 1064–1071.

[13] C. Campos, J. M. M. Montiel, and J. D. Tardós, "Fast and robust initialization for visual-inertial slam," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 1288–1294.

[14] R. Mur-Artal and J. D. Tardós, "Visual-Inertial monocular SLAM with map reuse," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, 2017.

[15] W. Huang and H. Liu, "Online initialization and automatic camera-imu extrinsic calibration for monocular visual-inertial slam," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 5182–5189.

[16] 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.

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

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

[19] W. Huang, H. Liu, and W. Wan, "An online initialization and self-calibration method for stereo visual-inertial odometry," *IEEE Transactions on Robotics*, vol. 36, no. 4, pp. 1153–1170, 2020.

[20] P. Chen, W. Guan, and P. Lu, "ESVIO: Event-based stereo visual inertial odometry," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3661–3668, 2023.

[21] L. Kneip and S. Lynen, "Direct optimization of frame-to-frame rotation," in *IEEE International Conference on Computer Vision*, 2013, pp. 2352–2359.

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

[23] L. Kneip, R. Siegwart, and M. Pollefeys, "Finding the exact rotation between two images independently of the translation," in *European Conference on Computer Vision*, 2012, pp. 696–709.

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

[25] 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*, 2016.