Check for updates

Special Issue: RSS2023

The International Journal of Robotics Research
2024, Vol. 0(0) 1–29
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02783649241262452
journals.sagepub.com/home/ijr

Fast and robust learned single-view depth-aided monocular visual-inertial initialization

Nathaniel Merrill, Patrick Geneva*, Saimouli Katragadda, Chuchu Chen and Guoquan Huang

Abstract

In monocular visual-inertial navigation, it is desirable to initialize the system as quickly and robustly as possible. A state-of-the-art initialization method typically constructs a linear system to find a closed-form solution using the image features and inertial measurements and then refines the states with a nonlinear optimization. These methods generally require a few seconds of data, which however can be expedited (less than a second) by adding constraints from a robust but only up-to-scale monocular depth network in the nonlinear optimization. To further accelerate this process, in this work, we leverage the scale-less depth measurements instead in the linear initialization step that is performed prior to the nonlinear one, which only requires a single depth image for the first frame. Importantly, we show that the typical estimation of all *feature states independently in the closed-form solution can be modeled as estimating* only *the scale and bias parameters of the learned depth map. As such, our formulation enables building a smaller minimal problem than the state of the art, which can be seamlessly integrated into RANSAC for robust estimation. Experiments show that our method has state-of-the-art initialization performance in simulation as well as on popular real-world datasets (TUM-VI, and EuRoC MAV). For the TUM-VI dataset in simulation as well as real-world, we demonstrate the superior initialization performance with only a 0.3 s window of data, which is the smallest ever reported, and validate that our method can initialize more often, robustly, and accurately in different challenging scenarios.*

Keywords

Visual-inertial odometry, SLAM, initialization, depth learning, minimal problem, robust estimation

Received 10 January 2024; Revised 8 May 2024; Accepted 27 May 2024

1. Introduction

Visual-inertial odometry (VIO) facilitates real-time 3D motion tracking through the utilization of a camera and an inertial measurement unit (IMU) (Huang, 2019). The small size, low cost, efficiency, and complementary sensing characteristics have made VIO emerge as a foundational technology for AR/VR, robotics (Camurri et al., 2020; Chen et al., 2022; Wu et al., 2017), and autonomous applications (Bayard et al., 2019; Eisele et al., 2019; Özaslan et al., 2017).

Two typical classes of VIO estimator designs are nonlinear optimization-based approaches (Campos et al., 2021; Leutenegger et al., 2015; Qin et al., 2018; Usenko et al., 2019) and light-weight filter-based ones (e.g., an extended Kalman filter (EKF)) (Bloesch et al., 2017; Geneva et al., 2020; Hesch et al., 2014; Li and Mourikis, 2013; Mourikis and Roumeliotis, 2007). Both of these approaches rely on good initial conditions (e.g., velocity and gravity) in order to run successfully, and it is highly desirable to calculate the initial conditions as quickly as possible in order to decrease the time the user or end application has to wait to start, especially if the VIO system is reset and needs to re-initialize on the fly. The initial conditions can be recovered by making assumptions about the motion (e.g., static), but under dynamic scenarios it is better to solve a visual-inertial structure from motion (VI-SfM) problem in order to initialize without making risky assumptions (Dong-Si and Mourikis, 2012; Martinelli, 2014). However, even VI-SfM can fail, especially under low-excitation scenarios.

To tackle this initialization problem, a recent method by Zhou et al. (2022) proposed to leverage learned monocular depth to provide additional constraints to the VI-SfM and help in the low excitation case, where the monocular priors are applied to each keyframe in the final bundle adjustment

Robot Perception and Navigation Group, University of Delaware, Newark, DE, USA

*Denotes equal contribution

Corresponding author:
Nathaniel Merrill, Department of Computer and Information Sciences, University of Delaware, 130 Academy Street, Newark, DE 19716, USA.
Email: nmerrill@udel.edu

(BA) step. To initialize the visual-inertial bundle adjustment (VI-BA), this method utilizes a closed-form solution similar to Li and Mourikis (2014), which compared to the nonlinear VI-BA is far more unstable due to the larger number of linear approximations required. In this work, we instead propose a simple yet effective method to utilize learned monocular depth priors in the *closed-form* linear initialization instead of the VI-BA refinement step, leveraging the single-image depth learned over millions of diverse examples as known prior information to reduce the number of parameters that need to be estimated in the fragile linear system. Specifically, the primary contributions of our work include:

- We propose a new formulation for closed-form visual-inertial linear initialization which leverages affine-invariant (scale-less) single-image depth to reduce the number of feature parameters to just a scale and bias.
- Our novel formulation allows for seamless integration of the minimal linear system into a robust RANSAC outlier rejection algorithm, which can be used to reject both bad depth priors as well as outlier feature tracks that may be present, whereas the typical linear system is less suitable for RANSAC.
- Extensive simulations show the proposed affine-invariant depth-aided linear system is able to provide an improved initial guess and result in lower orientation and velocity errors for short initialization windows after nonlinear refinement. Perturbation studies quantify the impact noise magnitudes and assumed bias on the recovered states.
- We validate our method on two public real-world datasets, and show that our method can improve the performance under the challenging scenario of 0.5 s of data with five keyframes. We additionally show superior initialization performance for the new and *even more* challenging scenario of a 0.3 s initialization window, and extensive ablation studies show that our method has superior performance in the presence of outliers and a reduced number of available feature tracks.

It is important to note that this work significantly extends our previous conference paper (Merrill et al., 2023) by including a minimal case analysis, detailed simulations, and sensitivity studies, as well as much more thorough real-world results. More specifically, a more comprehensive list of baselines are included for each experiment, a new dataset (EuRoC Machine Hall) is added, the VIO tracking accuracy is evaluated on each dataset, the robustness to low number of feature tracks is evaluated on every dataset, the linear system results are evaluated on every dataset, and the relative pose error (RPE) is evaluated for the VIO tracking accuracy on each dataset. Additionally, timing on an embedded device (Jetson Orin) is also provided.

The paper is organized as follows: Section 2 provides a review of related works, Section 3 provides background on the typical visual-inertial initialization problem, the proposed method is detailed in Section 4, simulation investigations are performed in Section 6, and then an extensive evaluation on real-world datasets is performed in Section 7 against the state-of-the-art baselines. Finally, we offer some discussion of the limitations of our method in Section 8 before concluding the paper in Section 9.

2. Related works

Many works have investigated different methods for performing visual-inertial initialization, and can be generally divided into two different categories: (1) loosely-coupled algorithms and (2) closed-form solutions. Loosely-coupled algorithms split the problem into first recovering an up-to-scale camera-only SfM trajectory result and then recover the scale given the inertial measurements, while closed-form solutions directly formulate a linear system involving both visual and inertial measurements.

2.1. Loosely-coupled algorithms

The works by Mur-Artal and Tardós (2017b) and Qin and Shen (2017); Qin et al. (2018) use a loosely-coupled approach. Mur-Artal and Tardós (2017b) leverage ORB-SLAM (Campos et al., 2021; Mur-Artal and Tardós, 2017a) SfM results and formulate a small linear system involving the up-to-scale poses and inertial preintegration to directly recover scale and gravity—which are then refined along with the accelerometer bias in a secondary step. A later work by Campos et al. (2020) additionally uses the up-to-scale SfM poses, but instead directly optimizes up-to-scale velocities, gravity direction, biases, and scale. Since an initial guess of scale is required for nonlinear optimization, they run the initialization multiple times at different initial scales and select the one which gives the smallest cost.

Qin and Shen (2017) and Qin et al. (2018) leverage a simplified SfM pipeline to obtain the up-to-scale trajectory, and then formulate a linear system that recovers scale, gravity, and velocity. A more recent work by Zuñiga-Noël et al. (2021) showed that up-to-scale SfM results could be leveraged in a quadratically-constrained least-squares problem, similar to closed-form solutions, which constrains the known magnitude of gravity to improve the accuracy. Another work by Concha et al. (2021) proposed a method that quickly initializes the 6 degrees of freedom (DoF) pose without motion parallax by decoupling the problem into the rotation, translation direction (5DoF) and magnitude of the translation (1DoF). While promising due to their robustification with RANSAC to handle outliers, they do not directly leverage inertial information in these low parallax scenarios. A key downside of loosely-coupled algorithms is that they are reliant on good SfM results, which require significant parallax and are typically computationally expensive to obtain.

2.2. Closed-form solutions

The earliest works on closed-form solutions are by Dong-Si and Mourikis (2011, 2012) and Martinelli (2011, 2014). In particular, Dong-Si and Mourikis (2012) propose the use of a quadratically-constrained least-squares problem which enforces the gravity magnitude, and showed improvements over methods which did not enforce this constraint. They focus on the recovery of an unknown IMU-to-camera rotation and translation, and directly recover the 3DoF feature positions in the first reference frame—where Martinelli (2014) recovers the depth of each feature for each bearing observation in every frame. A work by Li and Mourikis (2014) tries to address the lack of robustness by incorporating measurement noise by using estimated feature depths to simplify the feature reprojection cost into an approximate convex minimization problem. A key drawback is requiring knowledge of the average scene depth.

Another work by Kaiser et al. (2016) focuses on evaluating sensitivities to accelerometer and gyroscope biases, which is further extended by Campos et al. (2019) to include an observability and consensus test to remove poor initialization results near pure rotation and with limited acceleration motions. A recent work by Evangelidis and Micusik (2021) focuses on reducing the computational demands of Martinelli's (2014) linear system, and showed that the marginalization (projection) of the depth of each feature bearing and redundant 3DoF feature in a reference frame was possible and efficient.

2.3. Learning-aided initialization

Recently, a handful of works have emerged which investigate the use of learning-based methods to aid traditional SfM and visual-inertial initialization problems. Liu et al. (2022) utilizes a large MiDaS (Ranftl et al., 2022) depth estimation model to replace the traditional 5-point algorithm (Nistér, 2004) with a PnP alignment to the learned depth cloud. Another work by Hruby et al. (2022) employs model learning to select a starting problem solution which could numerically be continued without requiring significant samples within a RANSAC formulation. Both of these methods, while outside of the visual-inertial field, utilize learning in the linear initialization stage—similar in spirit to our approach. Linear initialization, whether in visual or visual-inertial systems, has always been a highly-unstable process, and can gain large benefits from learned prior information.

The work closest to ours is that by Zhou et al. (2022). This work is the first to leverage learned affine-invariant depth priors to better constrain the VI-BA—which is performed after solving a closed-form solution by Li and Mourikis (2014). This prior work showed that the inclusion of affine-invariant depth constraints in their VI-BA improved the problem conditioning, robustness, and accuracy under low-excitation scenarios. As compared to this, we look to leverage the affine-invariant depth directly within the *linear* initialization stage. As opposed to recovering each feature state independently, our linear system is simplified to only recovering the scale and bias of the predicted depth map. This additionally enables the application of RANSAC to further robustify the problem to outliers.

3. Monocular visual-inertial linear initialization

We consider a sensor platform consisting of a monocular camera and an inertial measurement unit (IMU). During the initialization time period N images at $[t_0, \ldots, t_N]$ are recorded along with IMU readings. The minimal state we wish to recover is (Dong-Si and Mourikis, 2011, 2012):

$$\mathbf{x} = \begin{bmatrix} {}^{I_0}\mathbf{p}_{f_1}^\top & \cdots & {}^{I_0}\mathbf{p}_{f_M}^\top & {}^{I_0}\mathbf{v}_{I_0}^\top & {}^{I_0}\mathbf{g}^\top \end{bmatrix}^\top \tag{1}$$

where $\{I_0\}$ denote the first IMU frame, ${}^{I_0}\mathbf{p}_{f_i}$ is the 3DoF feature position with respect to $\{I_0\}$, and ${}^{I_0}\mathbf{v}_{I_0}$, ${}^{I_0}\mathbf{g}$ are the velocity of the platform and local gravity expressed in the $\{I_0\}$ frame, respectively.

3.1. Inertial measurement model

A canonical three-axis IMU provides linear acceleration, ${}^I\mathbf{a}_m$, and angular velocity, ${}^I\boldsymbol{\omega}_m$, measurements expressed in the local IMU frame $\{I\}$:

$$\mathbf{a}_m(t) = \mathbf{a}(t) + {}^I_G\mathbf{R}(t)\,{}^G\mathbf{g} + \mathbf{b}_a(t) + \mathbf{n}_a(t) \tag{2}$$

$$\boldsymbol{\omega}_m(t) = \boldsymbol{\omega}(t) + \mathbf{b}_g(t) + \mathbf{n}_g(t) \tag{3}$$

where ${}^G\mathbf{g} \simeq [0,0,9.81]^\top$ is the gravitational acceleration expressed in the global frame $\{G\}$, and $\mathbf{n}_g$, $\mathbf{n}_a$ are zero-mean white Gaussian noises. ${}^I_G\mathbf{R}$ denotes the rotation matrix that transforms a position expressed in the global frame to one in the local frame. We assume that the biases $\mathbf{b}_a$ and $\mathbf{b}_g$ are known with reasonable accuracy. The continuous time IMU kinematics which evolve the state from time t_k to t_{k+1} are (Chatfield, 1997; Trawny and Roumeliotis, 2005):

$${}^{I_{k+1}}_{G}\mathbf{R} = {}^{I_{k+1}}_{I_k}\Delta\mathbf{R}\,{}^{I_k}_{G}\mathbf{R} \tag{4}$$

$${}^G\mathbf{p}_{I_{k+1}} = {}^G\mathbf{p}_{I_k} + {}^G\mathbf{v}_{I_k}\Delta T - \frac{1}{2}{}^G\mathbf{g}\Delta T^2 + {}^{I_k}_{G}\mathbf{R}^\top {}^{I_k}\boldsymbol{\alpha}_{I_{k+1}} \tag{5}$$

$${}^G\mathbf{v}_{I_{k+1}} = {}^G\mathbf{v}_{I_k} - {}^G\mathbf{g}\Delta T + {}^{I_k}_{G}\mathbf{R}^\top {}^{I_k}\boldsymbol{\beta}_{I_{k+1}} \tag{6}$$

where ${}^{I_k}\alpha_{k+1}$ and ${}^{I_k}\beta_{k+1}$ are the preintegration terms (Eckenhoff et al., 2019; Forster et al., 2015; Lupton and Sukkarieh, 2012):

$$ {}^{I_k}\boldsymbol{\alpha}_{I_{k+1}} = \int_{t_k}^{t_{k+1}} \int_{t_k}^{s} {}^{k}_{u}\Delta\mathbf{R}(\mathbf{a}_m(u) - \mathbf{b}_a(u) - \mathbf{n}_a(u))duds $$

$$ {}^{I_k}\boldsymbol{\beta}_{I_{k+1}} = \int_{t_k}^{t_{k+1}} {}^{k}_{u}\Delta\mathbf{R}(\mathbf{a}_m(u) - \mathbf{b}_a(u) - \mathbf{n}_a(u))du $$

We can transform an integration from t_0 to t_k in the global into the first IMU frame $\{I_0\}$:

$$ {}^{I_k}_{I_0}\mathbf{R} \triangleq {}^{I_k}_{I_0}\Delta\mathbf{R} \tag{7} $$

$$ {}^{I_0}\mathbf{p}_{I_k} \triangleq {}^{I_0}\mathbf{v}_{I_0}\Delta T_k - \frac{1}{2}{}^{I_0}\mathbf{g}\Delta T_k^2 + {}^{I_0}\boldsymbol{\alpha}_{I_k} \tag{8} $$

$$ {}^{I_0}\mathbf{v}_{I_k} \triangleq {}^{I_0}\mathbf{v}_{I_0} - {}^{I_0}\mathbf{g}\Delta T_k + {}^{I_0}\boldsymbol{\beta}_{I_k} \tag{9} $$

where $\Delta T_k = (t_k - t_0)$ is the time span for integration. These can be found by rotating the orientation and velocity with ${}^{I_0}_{G}\mathbf{R}$ and computing the relative position change ${}^{I_0}\mathbf{p}_{I_k} = {}^{I_0}_{G}\mathbf{R}(G\mathbf{p}_{I_k} - {}^{G}\mathbf{p}_{I_0})$, and defines the *relative* IMU integration in the fixed $\{I_0\}$ frame (Geneva and Huang, 2022).

3.2. Feature bearing observations

Assuming a calibrated perspective camera, the bearing measurement of the ith feature at timestep t_k can be related to the state by the following:

$$ \mathbf{z}_{i,k} := \boldsymbol{\Lambda}\left({}^{C_k}\mathbf{p}_{f_i}\right) + \mathbf{n}_k \tag{10} $$

$$ {}^{C_k}\mathbf{p}_{f_i} = {}^{C}_{I}\mathbf{R}\,{}^{I_k}_{I_0}\mathbf{R}\left(I_0\mathbf{p}_{f_i} - {}^{I_0}\mathbf{p}_{I_k}\right) + {}^{C}\mathbf{p}_{I} \tag{11} $$

where $\boldsymbol{\Lambda}([x\ y\ z]^\top) = [x/z\ y/z]^\top$ is the camera perspective projection model, $\mathbf{z}_{i,k} = [u_{i,k}, v_{i,k}]^\top$ is the normalized feature bearing measurement with white Gaussian noise $\mathbf{n}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k)$, and $\{{}^{C}_{I}\mathbf{R}, {}^{C}\mathbf{p}_{I}\}$ are the known camera-IMU transformation. Equation (10) can be re-written as the following linear constraint (Dong-Si and Mourikis, 2012):

$$ \begin{bmatrix} 1 & 0 & -u_{i,k} \\ 0 & 1 & -v_{i,k} \end{bmatrix} {}^{C_k}\mathbf{p}_{f_i} \triangleq \boldsymbol{\Gamma}_{i,k}\,{}^{C_k}\mathbf{p}_{f_i} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{12} $$

We can then substitute equations (8) and (11) to give:

$$ \mathbf{A}_{i,k}\,\mathbf{x} = \mathbf{b}_{i,k} \tag{13} $$

$$ \mathbf{A}_{i,k} = \boldsymbol{\Upsilon}_{i,k}\begin{bmatrix} \cdots & \mathbf{I}_3 & \cdots & -\Delta\mathbf{T}_k & \Delta\mathbf{T}_k^2 \end{bmatrix} \tag{14} $$

$$ \mathbf{b}_{i,k} = \boldsymbol{\Upsilon}_{i,k}\,{}^{I_0}\boldsymbol{\alpha}_{I_k} - \boldsymbol{\Gamma}_{i,k}\,{}^{C}\mathbf{p}_{I} \tag{15} $$

where $\Delta\mathbf{T}_k = \Delta T_k\mathbf{I}_3$ and $\boldsymbol{\Upsilon}_{i,k} = \boldsymbol{\Gamma}_{i,k}\,{}^{C}_{I}\mathbf{R}\,{}^{I_k}_{I_0}\mathbf{R}$. This can be "stacked" to recover a complete $\mathbf{A}\mathbf{x} = \mathbf{b}$, and given M features from N images, $\mathbf{A} \in \mathbb{R}^{2MN \times (3M+6)}$ and $\mathbf{b} \in \mathbb{R}^{2MN}$.

3.3. Constrained linear least-squares

We follow the method by Dong-Si and Mourikis (2012, 2011) and Geneva and Huang (2022) and formulate a constrained linear least-squares problem given the stacked observations (see equation (13)):

$$ \min \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 = \left\| \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ {}^{I_0}\mathbf{g} \end{bmatrix} - \mathbf{b} \right\|_2 \tag{16} $$

$$ \text{subject to} \quad \left\| {}^{I_0}\mathbf{g} \right\|_2 = g \tag{17} $$

The optimal solution can be derived using Lagrange multipliers (Dong-Si and Mourikis, 2011). The gravity constraint has been shown to have a noticeable impact on shorter trajectory lengths by Kaiser et al. (2016).

4. Learned depth-aided initialization

We now consider we are given a *single* affine-invariant (up-to scale and bias) depth map, $\mathbf{D}$, in the first frame of reference at time t_0. As compared to recovering the full feature states in equation (1), we instead formulate all features as a function of this depth map and the feature bearing in the first camera frame $\{C_0\}$. The minimal state we wish to recover is:

$$ \mathbf{x}' = \begin{bmatrix} a & b & {}^{I_0}\mathbf{v}_{I_0}^\top & {}^{I_0}\mathbf{g}^\top \end{bmatrix}^\top \tag{18} $$

where we have assumed that the affine-invariant depth map $\mathbf{D}$ is sufficiently accurate and can provide an estimate of the 3D structure in front of the camera up to a scale a and bias parameter b from just a single frame (Ranftl et al., 2022). An overview of the proposed method can be seen in Figure 1.

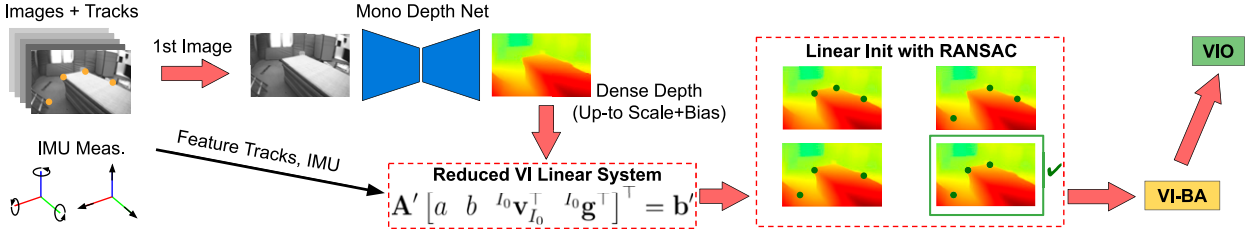

Figure 1. Overview of the proposed monocular-depth aided visual-inertial initialization method.

4.1. Depth-aided feature bearing model

We now modify the feature model in Section 3.2 to be a function of the affine-invariant depth map. We assume that for a single image the scale a and bias b are constant for the whole depth map. Specifically, for feature ${}^{I_0}\mathbf{p}_{f_i}$ we can express the metric depth scalar $z_i = \mathrm{Z}(u_{i,0}, v_{i,0})$ as a function of a, b, and $d_i = \mathrm{D}(u_{i,0}, v_{i,0})$:

$$\begin{aligned} {}^{I_0}\mathbf{p}_{f_i} &= {}^{I}_{C}\mathbf{R}\,{}^{C_0}\mathbf{p}_{f_i} + {}^{I}\mathbf{p}_{C} \\ &= z_i\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_i} + {}^{I}\mathbf{p}_{C} \\ &= (ad_i + b)\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_i} + {}^{I}\mathbf{p}_{C} \end{aligned} \tag{19}$$

where ${}^{I_0}\boldsymbol{\theta}_{C_0\to f_i} = {}^{I}_{C}\mathbf{R}[u_{i,0}\ v_{i,0}\ 1]^\top / \left\|[u_{i,0}\ v_{i,0}\ 1]^\top\right\|$ is the bearing vector of the feature rotated (but not translated) into the IMU frame, see Figure 2 for example frame of references. This treats the normalized 2D coordinates of the feature in the first camera frame $u_{i,0}$ and $v_{i,0}$ as a known quantity. Substituting equation (19) into equation (11) we can recover the following linear system:

$$\mathbf{A}'_{i,k}\mathbf{x}' = \mathbf{b}'_{i,k} \tag{20}$$

$$\mathbf{A}'_{i,k} = \boldsymbol{\Upsilon}_{i,k}\begin{bmatrix}\mathbf{B}_i & -\boldsymbol{\Delta}\mathbf{T}_k & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2\end{bmatrix} \tag{21}$$

$$\mathbf{b}'_{i,k} = \boldsymbol{\Upsilon}_{i,k}\,{}^{I_0}\boldsymbol{\alpha}_{I_k} - \boldsymbol{\Upsilon}_{i,k}\,{}^{I}\mathbf{p}_{C} - \boldsymbol{\Gamma}_{i,k}\,{}^{C}\mathbf{p}_{I} \tag{22}$$

$$\mathbf{B}_i = \begin{bmatrix} d_i\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_i} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_i}\end{bmatrix}. \tag{23}$$

Given M features from N images, $\mathbf{A}' \in \mathbb{R}^{2MN\times(2+6)}$ and $\mathbf{b}' \in \mathbb{R}^{2MN}$. One can see that the state size remains constant, no matter how many features are included in the problem. The structure of our system can be seen in Figure 3.

Remarks: As evident, this formulation of the linear initialization problem significantly relaxes the original one—reducing the need to estimate the 3D position of every feature to just estimating the scale and bias of the depth map predicted at t_0—which is shared between all features. Given a reasonable predicted affine-invariant depth $\mathbf{D}$ and a and b are well constrained, if the recovered scale parameter a is positive, all of the features will be in front of the camera as desired, and there will be no spurious feature positions (e.g., too close or too far due to high uncertainty).

It should be noted that the monocular depth network MiDaS (Ranftl et al., 2022) leveraged in this work actually produces affine-invariant *inverse* depth maps $\mathbf{D}_{\text{inv}}$, where $\mathrm{D}(u_i, v_i) = 1/\mathrm{D}_{\text{inv}}(u_i, v_i)$ (dropping the subscript for clarity), and the metric inverse depth is expressed as $\mathrm{Z}_{\text{inv}}(u_i, v_i) = a_{\text{inv}}\mathrm{D}_{\text{inv}}(u_i, v_i) + b_{\text{inv}}$. The use of affine-invariant *depth* instead of *inverse depth* is also reported by Liu et al. (2022), which utilizes the same class of depth networks as us. Due to the division, one may suspect that the scale and bias for depth, a and b, would be a nonlinear function of a_{inv} and b_{inv}, but in fact, it can be expressed linearly with the following relationship:

$$(a\,\mathrm{D}(u_i, v_i)) + b)(a_{\text{inv}}\,\mathrm{D}_{\text{inv}}(u_i, v_i) + b_{\text{inv}}) = 1. \tag{24}$$

Thus, estimating the scale and bias a and b in equation (18) instead of a_{inv} and b_{inv} is valid, and a_{inv}, b_{inv} can be recovered from a solution of a, b via stacking and solving

$$\begin{bmatrix}(a + b\mathrm{D}_{\text{inv}}(u_i, v_i)) & (a\mathrm{D}(u_i, v_i) + b)\end{bmatrix}\begin{bmatrix}a_{\text{inv}} \\ b_{\text{inv}}\end{bmatrix} = 1 \tag{25}$$

for all u_i, v_i, which is simply equation (24) rearranged. Similarly, equation (24) can be rearranged to recover a and b from estimates of a_{inv} and b_{inv} by just grouping different terms. Conversely a_{inv} and b_{inv} can be similarly recovered from a and b.

The fact that a, b and a_{inv}, b_{inv} can be related linearly also means that we can scale $\mathbf{D}_{\text{inv}}$ arbitrarily before using it in the linear system. To this end, for ensured numeric stability of $\mathbf{D}$, we scale $\mathbf{D}_{\text{inv}}$, which can have arbitrary value, into the range [1, 2] before computing $\mathbf{D}$ via:

$$\mathrm{D}_{\text{inv}}(u_i, v_i) = \frac{\mathrm{D}^0_{\text{inv}}(u_i, v_i) - \min(\mathbf{D}^0_{\text{inv}})}{\max(\mathbf{D}^0_{\text{inv}}) - \min(\mathbf{D}^0_{\text{inv}})} + 1 \tag{26}$$

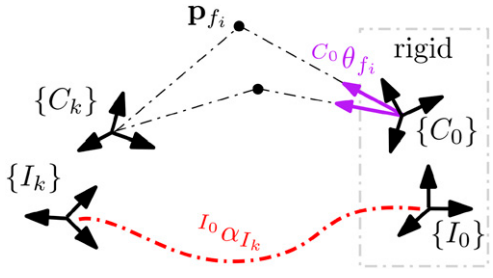

Figure 2. Frame of references used in the problem. Two features observed from both the $\{C_k\}$ and $\{C_0\}$ frame are shown. The transformation from the $\{I_k\}$ and $\{I_0\}$ is found through IMU integration. The bearing ${}^{C_0}\boldsymbol{\theta}_{f_i}$ is used along with the affine-invariant depth to recover the scale a and bias b.

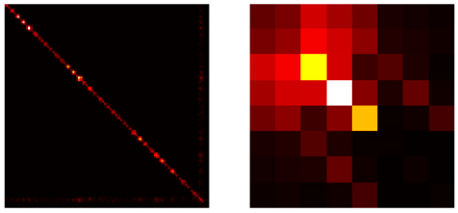

Figure 3. Structure of $\mathbf{A}^\top\mathbf{A}$ of Dong-Si and Mourikis (2012) (**DS 3D**) (left) and the proposed $\mathbf{A}'^\top\mathbf{A}'$ (right). The DS 3D system contains 35 features here (making it 111 × 111). While sparse, it is much larger than the proposed, which is 8 × 8 no matter how many features are included. The log condition number for DS 3D is 9.35 while the proposed is 8.15.

where $\mathbf{D}^0_{\text{inv}}$ is the raw affine-invariant inverse depth map from the monocular depth network. Note that the range [1,2] is chosen arbitrarily to avoid possible division by zero.

4.2. Outlier rejection in linear initialization

A key advantage of our proposed linear system formulation is its ability to be easily inserted into *small* minimal problems in a RANSAC loop to robustify it to outliers. In theory, each measurement in the minimal problem for equation (18) can be chosen from a different feature since each feature track constrains the same a and b states. However, in practice, we group the measurements by feature and view in order to (1) reject bad feature tracks and (2) reject bad depth network predictions. An overview of our RANSAC approach can be seen in Algorithm 1. A minimal set of features and poses are first randomly grouped and the constrained linear system, equation (16), is solved to recover the scale, bias, velocity, and gravity. These states are then used to compute the reprojection error for each measurement not used in the problem, and construct the inlier measurement set $\mathcal{S}$. The solution from the inlier set which gives the minimal error is selected as the best state estimate.

Algorithm 1 Linear Initialization with RANSAC

Require: Blocks $\mathbf{A}'_{ki}$, $\mathbf{b}'_{ki}$ of the complete linear system for $i \in \{1,\dots,M\}$, $k \in \{1,\dots,N\}$, minimal problem size $M_{\min}, N_{\min}$, maximum number of iterations K, thresholds $d_{\min}$, γ
Ensure: Robustified solution to linear system $\mathbf{x}'_{\text{best}}$

$e_{\text{best}} \leftarrow \infty$
for $i \in \{1,\dots,K\}$ **do**
  $\mathcal{S} \leftarrow$ Rand. sample $N_{\min}$ meas. from $M_{\min}$ feats.
  $\mathbf{A}'_s,\ \mathbf{b}'_s \leftarrow$ Stack blocks $i,k \in \mathcal{S}$
  $a, b, {}^{I_0}\mathbf{v}_{I_0}, {}^{I_0}\mathbf{g} \leftarrow \text{solve}(\mathbf{A}'_s,\ \mathbf{b}'_s)$
  for i,k not in $\mathcal{S}$ **do**
    $\mathbf{r} \leftarrow \mathbf{A}'_{ik}\begin{bmatrix} a & b & {}^{I_0}\mathbf{v}_{I_0}^\top & {}^{I_0}\mathbf{g}^\top \end{bmatrix}^\top - \mathbf{b}'_{ik}$
    if $||\mathbf{r}|| < \gamma$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup (i,k)$
    end if
  end for
  if $|\mathcal{S}| \geq d_{\min}$ **then**
    $\mathbf{A}'_{\text{inl}},\ \mathbf{b}'_{\text{inl}} \leftarrow$ Stack blocks $i,k \in \mathcal{S}$
    $a, b, {}^{I_0}\mathbf{v}_{I_0}, {}^{I_0}\mathbf{g} \leftarrow \text{solve}(\mathbf{A}'_{\text{inl}},\ \mathbf{b}'_{\text{inl}})$
    $\mathbf{r} \leftarrow \mathbf{A}'_{\text{inl}}\begin{bmatrix} a & b & {}^{I_0}\mathbf{v}_{I_0}^\top & {}^{I_0}\mathbf{g}^\top \end{bmatrix}^\top - \mathbf{b}'_{\text{inl}}$
    if $||\mathbf{r}|| < e_{\text{best}}$ **then**
      $e_{\text{best}} \leftarrow ||\mathbf{r}||$
      $\mathbf{x}'_{\text{best}} \leftarrow \begin{bmatrix} a & b & {}^{I_0}\mathbf{v}_{I_0}^\top & {}^{I_0}\mathbf{g}^\top \end{bmatrix}^\top$
    end if
  end if
end for

We emphasize that the RANSAC approach becomes feasible due to our relaxation of the original linear system from the inclusion of the affine-invariant depth map. While the hard minimal problem for our RANSAC algorithm is 3 views and 2 features (discussed in Section 5), we use 3 views and 4 features in the minimal problems in all experiments for slightly improved conditioning which we found to be more robust to a low number of available feature tracks.

4.3. Nonlinear refinement

We recover the 3D position of all features (inlier or not) via equation (19), and recover gravity aligned orientation by transforming the recovered gravity ${}^{I_0}\mathbf{g}$ into a gravity aligned frame ${}^{G}\mathbf{g} = [0,0,9.81]^\top$. The VI-BA problem which refines the state estimates, takes into account measurement uncertainties, and relinearizes the states to iteratively improve the accuracy. The state vector of this optimization process can be defined as:

$$\mathbf{x}_{\text{mle}} = \begin{bmatrix} \mathbf{x}_{I_0}^\top \dots \mathbf{x}_{I_N}^\top & {}^{G}\mathbf{p}_{f_1}^\top \dots {}^{G}\mathbf{p}_{f_M}^\top \end{bmatrix}^\top \tag{27}$$

$$\mathbf{x}_{I_k} = \begin{bmatrix} {}^{I_k}_{G}\overline{q}^\top & {}^{G}\mathbf{p}_{I_k}^\top & {}^{G}\mathbf{v}_{I_k}^\top & \mathbf{b}_{g,k}^\top & \mathbf{b}_{a,k}^\top \end{bmatrix}^\top \tag{28}$$

where each keyframe has its own bias estimate in order to model the bias's time-varying characteristics. Note also that we do *not* include the depth prior in the nonlinear optimization as Zhou et al. (2022) does, because it would require estimating the depth for *all* keyframe images (which could be computational and energy intensive even if possible in real time), rather than the *single* first one (which is *all* that is required in our solution). We empirically found that only including the depth prior in the first keyframe in the VI-BA optimization leads to the exact same result as optimizing without it, but perhaps could improve it if we had a scale prior as in Zhou et al. (2022). Thus, we omit the depth prior from the VI-BA and only use it in linear initialization, although including depth priors for all keyframes in the optimization helps as shown by Zhou et al. (2022). However, as shown later in Section 7, adding the depth prior in the VI-BA on top of our method does not always help the performance.

We solve the optimization problem with inertial $\mathbb{C}_I$, camera $\mathbb{C}_C$, and prior $\mathbb{C}_P$ cost terms:

$$\underset{\mathbf{x}_{\text{mle}}}{\text{argmin}}\ \mathbb{C}_I + \mathbb{C}_C + \mathbb{C}_P \tag{29}$$

With the following inertial cost function (Eckenhoff et al., 2019; Forster et al., 2015; Lupton and Sukkarieh, 2012):

$$\mathbb{C}_I \triangleq \sum_k \left\| \mathbf{x}_{I_{k+1}} \boxminus \mathbf{f}(\mathbf{x}_{I_k}, \mathbf{a}_{m_k}, \boldsymbol{\omega}_{m_k}) \right\|^2_{\mathbf{Q}_k} \tag{30}$$

where $\mathbf{Q}_k$ is the linearized measurement noise covariance. The camera re-projection cost is defined as (Geneva et al., 2020):

$$\mathbb{C}_C \triangleq \sum_{i,k} \left\| \mathbf{z}_{i,k} - \mathbf{h}(\mathbf{x}_{\text{mle}}) \right\|^2_{\mathbf{R}_i} \tag{31}$$

where $\mathbf{h}(\cdot)$ includes the camera's intrinsic distortion, projection, and camera-to-IMU extrinsic transformation, and $\mathbf{R}_i$ is the image pixel noise covariance.

In addition to constraining the unobservable initial global position and yaw rotation (Hesch et al., 2013; Zhang et al., 2018), we found that the gyroscope and especially accelerometer biases can nearly be unobservable and hard to initialize, and thus, we provide reasonable priors to these states to avoid numerical instabilities. The sensitivity to poor bias priors is investigation in Section 6.3. The prior cost is defined as:

$$\mathbb{C}_P \triangleq \left\| \mathbf{x}_{\text{mle}} \boxminus \breve{\mathbf{x}}_{\text{mle}} \right\|^2_{\boldsymbol{\Omega}_P^{-1}} \tag{32}$$

where $\breve{\mathbf{x}}_{\text{mle}}$ is the fixed state linearization point and $\boldsymbol{\Omega}_P$ is the prior information matrix—where large values are picked for unobservable state variables.

After the nonlinear refinement, the marginal covariance of the most recent IMU state in the VI-BA is recovered, and used to initialize the filter. In practice, we found that the covariance needs to be inflated a bit in order to properly initialize the filter.

5. Minimal case analysis

It is crucial to determine the minimal number of images and features required to estimate all the unknown parameters in equation (20). Provided the kth image, the general matrix form of equation (20) is given by:

$$\mathbf{D}\mathbf{K}\mathbf{x}' = \mathbf{b} \tag{33}$$

where we have partitioned the block diagonal matrix $\mathbf{D}$, dense matrix $\mathbf{K}$, and vector $\mathbf{b}$ as follows:

$$\mathbf{D} := \text{diag}(\Upsilon_{1,k}, \cdots, \Upsilon_{M,k}) \tag{34}$$

$$\mathbf{K} := \begin{bmatrix} \mathbf{A}'^{\top}_{1,k} & \cdots & \mathbf{A}'^{\top}_{M,k} \end{bmatrix}^{\top} \tag{35}$$

$$\mathbf{b} := \begin{bmatrix} \mathbf{b}'^{\top}_{1,k} & \cdots & \mathbf{b}'^{\top}_{M,k} \end{bmatrix}^{\top} \tag{36}$$

Without loss of generality, we assume that features can be observed in all images in order to simplify the minimal case analysis. As such, the number of measurements is 2 MN, where M is the number of features and N denotes the number of frames. The state size is 1 + 1 + 3+ 3 = 8, where include scalar a and b, 3DoF velocity ${}^{I_0}\mathbf{v}_{I_0}$, and 3DoF gravity ${}^{I_0}\mathbf{g}$.* Thus, the necessary condition is 2 $MN \geq 8$. We now identify the following cases for the number of available images:

- $N = 1$: The necessary condition is not met, regardless of the number of features.
- $N = 2$: The necessary condition will never be met regardless of the number of features.
- $N = 3$: The necessary condition is met when $M \geq 2$.
- $N \geq 4$: The necessary condition is met when $M \geq 1$.

Focusing on the two identified minimal cases we have: (i) two features seen in three images and (ii) one feature seen in four images. For both, the number of measurements is higher than the number of unknown variables, making the problem over-constrained, allowing for the computation of a distinct, singular solution. Our focus here is specifically on the rank of the $\mathbf{K}$ sub-matrix within the proposed affine-invariant depth-aided linear problem, see equation (35). For each scenario, we demonstrate that employing Gaussian elimination can streamline the matrix structure, revealing the rank, and facilitate a deeper analysis and understanding.

5.1. Two images (N = 2)

We begin by considering a scenario involving two images: the first image captured at t_0 with M features, and add an extra image taken at time t_k. Focusing on the $\mathbf{K}$ sub-matrix and defining the base frame I_0 as the first one, we can perform a column-wise Gaussian elimination:

$$\mathbf{K} = \begin{bmatrix} {}^{I_0}d_1\, {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_1} & -\boldsymbol{\Delta}\mathbf{T}_k & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\, {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_M} & -\boldsymbol{\Delta}\mathbf{T}_k & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 \end{bmatrix}$$

$$1/2\boldsymbol{\Delta}\mathbf{T}_k \underset{\sim}{*} C_3 + C_4$$

$$\begin{bmatrix} {}^{I_0}d_1\, {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_1} & -\frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 & \mathbf{0}_3 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\, {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0 \to f_M} & -\frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_k^2 & \mathbf{0}_3 \end{bmatrix}$$

We can conclude through inspection of the row rank that:[1]

$$\text{rank}(\mathbf{K}) \leq 8 - 3 \tag{37}$$

Thus this matrix is not full rank and the necessary condition will never meet regardless of the number of features.

5.2. Three images (N = 3)

The $\mathbf{K}$ matrix for the case of a base image at time t_0, and two extra images at t_1 and t_2 can be written as:

$$\mathbf{K} = \begin{bmatrix} {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_2 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_2^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_2 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_2^2 \end{bmatrix}$$

$$R_{3M+i} - R_i \underset{\sim}{\forall} i \in \{1,\ldots,3M\}$$

$$\begin{bmatrix} {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{\Delta}\mathbf{T}_2+\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}(\boldsymbol{\Delta}\mathbf{T}_2^2-\boldsymbol{\Delta}\mathbf{T}_1^2) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{\Delta}\mathbf{T}_2+\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}(\boldsymbol{\Delta}\mathbf{T}_2^2-\boldsymbol{\Delta}\mathbf{T}_1^2) \end{bmatrix}$$

$$R_i - R_{i+1}\ \forall i \in \underset{\sim}{\{3M+1,\ldots,6M\}}$$

$$\begin{bmatrix} {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{\Delta}\mathbf{T}_2+\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}(\boldsymbol{\Delta}\mathbf{T}_2^2-\boldsymbol{\Delta}\mathbf{T}_1^2) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

We can conclude through inspection of the row rank that:

$$\text{rank}(\mathbf{K}) = \min(3M+3, 8) \tag{38}$$

The necessary condition will be satisfied if $3N + 3 \geq 8 \Rightarrow M \geq 5/3$. The minimal number of features is 2.

5.3. Four images (N = 4)

The **K** matrix for the case of a base image at time t_0, and three extra images at t_1, t_2, and t_3 can be written as:

$$\mathbf{K} = \begin{bmatrix} {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_2 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_2^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_2 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_2^2 \\ {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_3 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_3^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_3 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_3^2 \end{bmatrix}$$

$$\sim \begin{bmatrix} {}^{I_0}d_1\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_1} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ {}^{I_0}d_M\,{}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & {}^{I_0}\boldsymbol{\theta}_{C_0\to f_M} & -\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}\boldsymbol{\Delta}\mathbf{T}_1^2 \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{\Delta}\mathbf{T}_2+\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}(\boldsymbol{\Delta}\mathbf{T}_2^2-\boldsymbol{\Delta}\mathbf{T}_1^2) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{\Delta}\mathbf{T}_3+\boldsymbol{\Delta}\mathbf{T}_1 & \frac{1}{2}(\boldsymbol{\Delta}\mathbf{T}_3^2-\boldsymbol{\Delta}\mathbf{T}_1^2) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where we have applied a Gaussian elimination similar to that in the previous section. We can conclude through inspection of the row rank that:

$$\text{rank}(\mathbf{K}) = \min(3M+6, 8) \tag{39}$$

The necessary condition will be satisfied if $3M + 6 \geq 8 \Rightarrow M \geq 2/3$. The minimal number of features is 1.

6. Simulation studies

We simulate a realistic 145 meter indoor monocular handheld trajectory based on the TUM-VI (Schubert et al., 2018) room 1 trajectory (see Figure 4). Table 1

captures all key sensor parameters, along with default algorithm parameters used throughout the simulation unless otherwise specified. The performance of the proposed method is compared against the baseline initialization method by Dong-Si and Mourikis (2012) (see Section 3), which has been implemented within the existing state-of-the-art method in OpenVINS's Geneva et al. (2020) and open sourced within the ov_init package Geneva and Huang (2022). For evaluation, we compare against two variants of this baseline: **DS 3D** and **DS 1D**, which implements the work of Dong-Si and Mourikis (2012) with 3D and 1D feature states, respectively. DS 3D is the current default initialization available in OpenVINS (Geneva et al., 2020). We denote the proposed method without the addition of RANSAC as **Ours w/o RANSAC**, and the proposed system aided with RANSAC as **Ours**. As mentioned in the previous section, the proposed RANSAC minimal problem uses 3 views and 4 features for robustness.

For the specific details on the visual-inertial simulator which generates realistic visual bearings and inertial measurements, we refer the reader to the original OpenVINS paper (Geneva et al., 2020). The continuous-time simulator was extended to support generation of a sparse depth-map, which is then normalized to a fixed affine-invariant range before being passed to the initialization module. We have chosen to directly perturb the metric sparse depth map, as compared to the affine-invariant depths, to ensure that a sufficient realistic magnitude is being added.

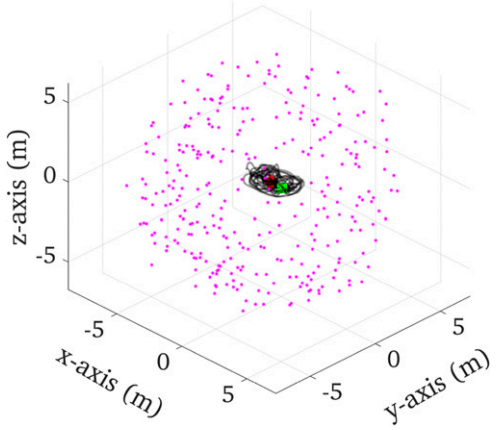

Figure 4. Simulation TUM-VI Room 1 trajectory and environmental features generated. We additionally simulate a ∼20 features near infinity (∼250 meter depths), not pictured.

Table 1. Simulation Parameters and Prior Standard Deviations for Measurement Perturbations.

Parameter	Value	Parameter	Value
Gyro. White Noise	2.054e-4	Gyro. Rand. Walk	1.111e-5
Accel. White Noise	2.076e-3	Accel. Rand. Walk	4.133e-4
Image Obs. Noise	1.0	Depth-map Noise	5 cm
Cam Freq. (Hz)	20	IMU Freq. (Hz)	400
Num. Poses	5	Tracked Feat	75

6.1. Effect of temporal initialization window

The first simulation study is on the effect of the window size on the accuracy of both the linear and nonlinear refined estimates. We fix the total number of keyframe (KF) poses to five, and change the length of time they are spread over. One would expect the depth of features to become more recoverable as the window length increases due to the additional translation and rotation observed, and thus the proposed method should have the largest benefit for the small window low-parallax cases.

Summarized in Table 2, we can first observe that both the baseline and proposed linear systems perform with similar levels of accuracy. Note that the scale error reported throughout the paper is the scale error of the estimated positions. We fit a Sim (3) between the estimated and ground truth trajectory, and with a Sim (3) with scale s, the scale error is calculated as $100(\max(s, 1/s) - 1)$. If we look at the metric feature error in Figure 5, it is clear that the proposed method is able to provide a better initial guess for the nonlinear refinement, further supporting why inclusion of affine-invariant depth provides benefits. Looking at the errors after nonlinear refinement (going back to Table 2), we can confirm that the proposed method provides the largest benefit at the shorter low-parallax window sizes, with it providing minimal improvements at the much longer 1 s window length. Note that we have not simulated any outliers and thus any benefit from robustness is absent (see Section. 7.3.3 for this impact).

6.2. Effect of measurement noise

Next we investigate the impact of measurement noise on the initialization accuracy. As seen in Figure 6, the pixel bearing noise has little effect on the linear system. As the depth measurement error is increased, even with 10 cm perturbations, the proposed method does not have any sufficient degradation when compared to the baseline, which does not leverage these measurements. However, in Figure 7, it is clear that all methods after MLE refinement suffer a similar amount to the feature bearing observation noise, with the baseline methods having a much more varied scale range as compared to the proposed method. On the other hand, our method after MLE refinement is similarly not affected much by the depth measurement noise. For 1.5px noise levels after MLE refinement, we have summarized the

Table 2. Average Errors Over 10 Runs of the Recovered Inertial State, After Solving of the Linear System (Top Half), and After a Following Nonlinear Refinement (Bottom Half). Feature Bearings and Depths Were Corrupted With 1 deg and 5 cm, Respectively.

	Win. (s)	Algorithm	Ori (deg)	Vel (m/s)	Scale error (%)
Linear System	0.3	DS 3D	13.05 ± 7.96	1.16 ± 0.45	658.08 ± 525.34
		DS 1D	13.17 ± 8.09	1.17 ± 0.46	664.15 ± 529.87
		Ours w/o RANSAC	13.28 ± 8.19	1.19 ± 0.45	673.86 ± 506.07
		Ours	13.50 ± 10.18	1.20 ± 0.50	658.88 ± 474.74
	0.5	DS 3D	11.60 ± 7.26	1.15 ± 0.47	494.02 ± 379.76
		DS 1D	11.99 ± 7.20	1.16 ± 0.48	507.59 ± 381.42
		Ours w/o RANSAC	11.96 ± 7.28	1.16 ± 0.48	502.76 ± 367.28
		Ours	12.12 ± 7.80	1.19 ± 0.54	483.76 ± 394.49
	1.0	DS 3D	6.86 ± 4.87	0.90 ± 0.55	309.04 ± 366.85
		DS 1D	7.57 ± 5.43	1.00 ± 0.59	328.82 ± 384.71
		Ours w/o RANSAC	6.79 ± 4.20	0.94 ± 0.49	318.94 ± 389.18
		Ours	7.17 ± 5.64	0.95 ± 0.70	245.73 ± 382.46
After MLE Optimization	0.3	DS 3D	8.00 ± 5.16	0.61 ± 0.35	232.62 ± 350.39
		DS 1D	8.03 ± 5.15	0.60 ± 0.36	224.70 ± 363.11
		Ours w/o RANSAC	7.24 ± 4.20	0.47 ± 0.30	95.19 ± 165.55
		Ours	7.36 ± 4.73	0.48 ± 0.30	95.24 ± 160.69
	0.5	DS 3D	3.85 ± 2.91	0.32 ± 0.26	50.98 ± 99.43
		DS 1D	3.87 ± 2.94	0.31 ± 0.26	50.92 ± 111.32
		Ours w/o RANSAC	3.68 ± 2.53	0.28 ± 0.21	28.06 ± 40.84
		Ours	3.71 ± 2.63	0.28 ± 0.21	27.93 ± 38.55
	1.0	DS 3D	1.34 ± 1.26	0.16 ± 0.17	12.30 ± 17.68
		DS 1D	1.31 ± 1.25	0.16 ± 0.16	11.36 ± 15.03
		Ours w/o RANSAC	1.33 ± 1.15	0.17 ± 0.16	11.65 ± 17.04
		Ours	1.39 ± 1.16	0.17 ± 0.16	11.81 ± 16.50

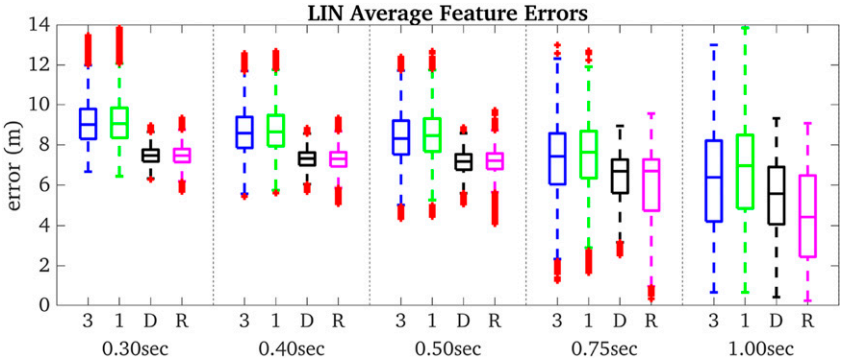

Figure 5. Metric feature errors after linear system recovery for varying window lengths over 10 runs. We define: DS 3D as **3** (blue), DS 1D as **1** (green), Ours w/o RANSAC as **D** (black), and Ours as **R** (magenta). Note that outliers outside the sample 3σ bound have been filtered for presentation clarity.

statistics in Table 3. The proposed approach has clear gains in the velocity recovery throughout all noise levels, while the depth map noise has little effect on the final state accuracy levels.

6.3. Sensitivities to perturbations

A key assumption of all methods is that the IMU biases with sufficient accuracy are known a priori. These biases are treated as true within the linear system, and with a 0.01 rad/s and 0.05 m/s^2 prior during nonlinear refinement for the gyroscope and accelerometer, respectively. We first investigate large unknown perturbations to the gyroscope bias for the linear system in Figure 8 and after MLE refinement in Figure 9. The linear system is not affected much by the perturbation in gyroscope bias—perhaps due to the fact that the orientation errors coming from the linear system are already very large, so perturbations in the gyroscope bias may not affect the results as much. On the other hand, the MLE result is noticeably affected by the gyroscope bias perturbation. The roll pitch errors after MLE refinement all quickly increase by large amounts for all methods along with the velocities. There is a clear scale accuracy

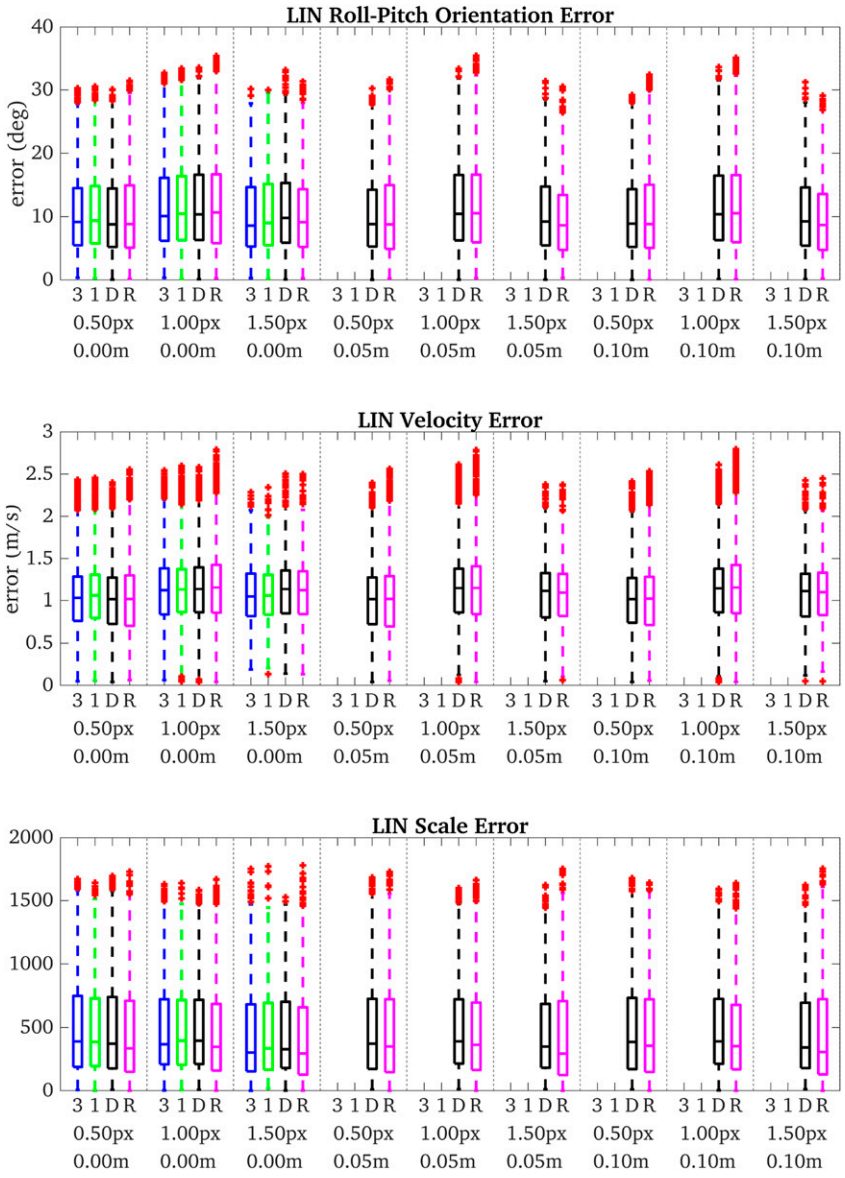

Figure 6. Orientation and velocity errors of the final pose of each linear system for different noise levels of feature bearings (px) and depths (m, on scaled depth pre-normalization). We define: DS 3D as **3** (blue), DS 1D as **1** (green), Ours w/o RANSAC as **D** (black), and Ours as **R** (magenta). Note that outliers outside the sample 3σ bound have been filtered for presentation clarity.

improvement after MLE refinement provided by the proposed method for the lower gyroscope bias perturbation levels. When looking at the accelerometer bias sensitivity in Figures 10 and 11, we can see a similar story. Shown in Figures 12 and 13 is the sensitivity to perturbations in the simulated gravity magnitude. Again, the linear system results are mostly not affected by the perturbation.

On the other hand, it can be observed that the proposed method after MLE refinement has comparable orientation and velocity error, but improved scale error over the baselines. We found it impressive that the initialization states did not have higher errors due to such large perturbation, which shows the robustness of all methods—with the proposed having a particularly good ability to recover accurate scale throughout.

7. Real-world experiments

To validate the proposed singe-image depth-aided monocular VIO initialization in the real world, we employ the two most popular public VI datasets: EuRoC MAV (Burri et al., 2016) and TUM-VI (Schubert et al., 2018). We choose an evaluation method similar to that of Zhou et al. (2022), where we divide each sequence into 10 s windows, run initialization for each of the entry points, and averaging the results from each successful run. This evaluation method has the advantage that it

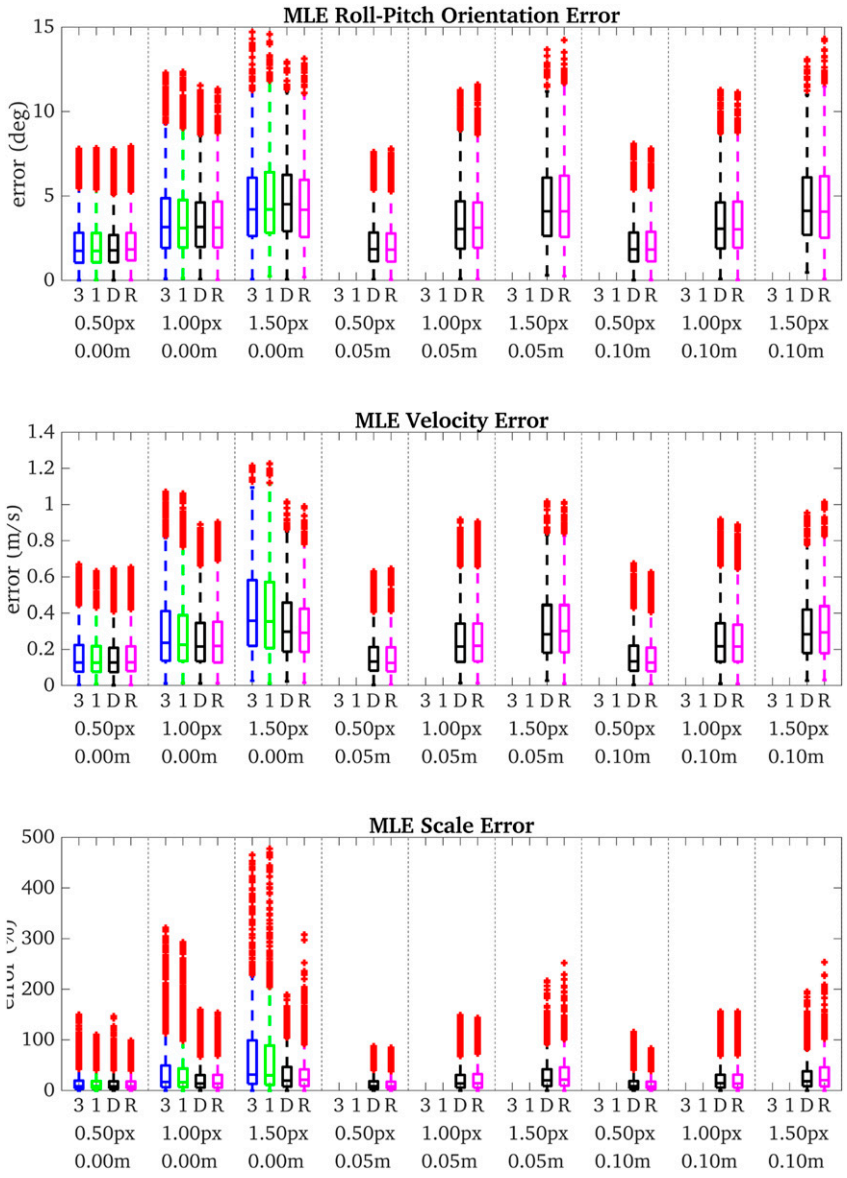

Figure 7. Orientation and velocity errors of the final pose after MLE refinement for different noise levels of feature bearings (px) and depths (m, on scaled depth pre-normalization). We define: DS 3D as **3** (blue), DS 1D as **1** (green), Ours w/o RANSAC as **D** (black), and Ours as **R** (magenta). Note that outliers outside the sample 3σ bound have been filtered for presentation clarity.

initializes in many places within each sequence and tests not only the accuracy of the initialization window poses, but also the accuracy of VIO using the initialization result. A run is considered successful if (1) the linear system successfully returned a result, (2) the MLE optimization has converged, and (3) the covariance could successfully be recovered without being rank-deficient.

In our experiments, we mainly consider the absolute trajectory error (ATE) (Zhang and Scaramuzza, 2018) metric for position and orientation. We additionally use all recovered poses to perform a Sim(3) alignment to the ground truth in order to report the scale error (defined in Section 6.1). For the ATE, trajectories are aligned to the ground truth using the first frame by solving for the optimal position and yaw transform between the estimate and ground truth (see Zhang and Scaramuzza (2018)). Since we do *not* use a scale-aware alignment such as Sim(3) to compute the ATE, scale accuracy will directly impact the position, and gravity errors will also affect the orientation ATE. For the VIO tracking accuracy, we also consider the relative pose error (RPE), where the trajectory is grouped into segments of different length and then the error of those segments is calculated. Since VIO is only run for a maximum of 10 s for each initialization, the RPE window lengths are shorter than typically reported due to the fact that VIO does not travel very far within this period. RPE is generally considered a more important metric than ATE since it investigates the accuracy more thoroughly at multiple different window lengths rather than just for the

Table 3. Estimation Accuracy After Nonlinear MLE Refinement Over 10 Runs With 1.5 Pixel Bearing Observation Noise.

Algorithm	Noise	Ori (deg)	Vel (m/s)	Scale error (%)
DS 3D	-	4.98 ± 3.31	0.43 ± 0.27	84.33 ± 127.91
DS 1D	-	5.03 ± 3.21	0.42 ± 0.27	84.00 ± 134.92
Ours w/o RANSAC	0.00 m	4.97 ± 2.87	0.35 ± 0.22	38.71 ± 52.84
	0.05 m	4.75 ± 2.97	0.35 ± 0.23	40.26 ± 63.36
	0.10 m	4.77 ± 2.81	0.33 ± 0.21	36.75 ± 53.66
Ours	0.00 m	4.73 ± 2.90	0.34 ± 0.22	43.06 ± 95.05
	0.05 m	4.87 ± 3.24	0.35 ± 0.23	42.26 ± 71.71
	0.10 m	4.79 ± 3.23	0.35 ± 0.23	41.05 ± 71.57

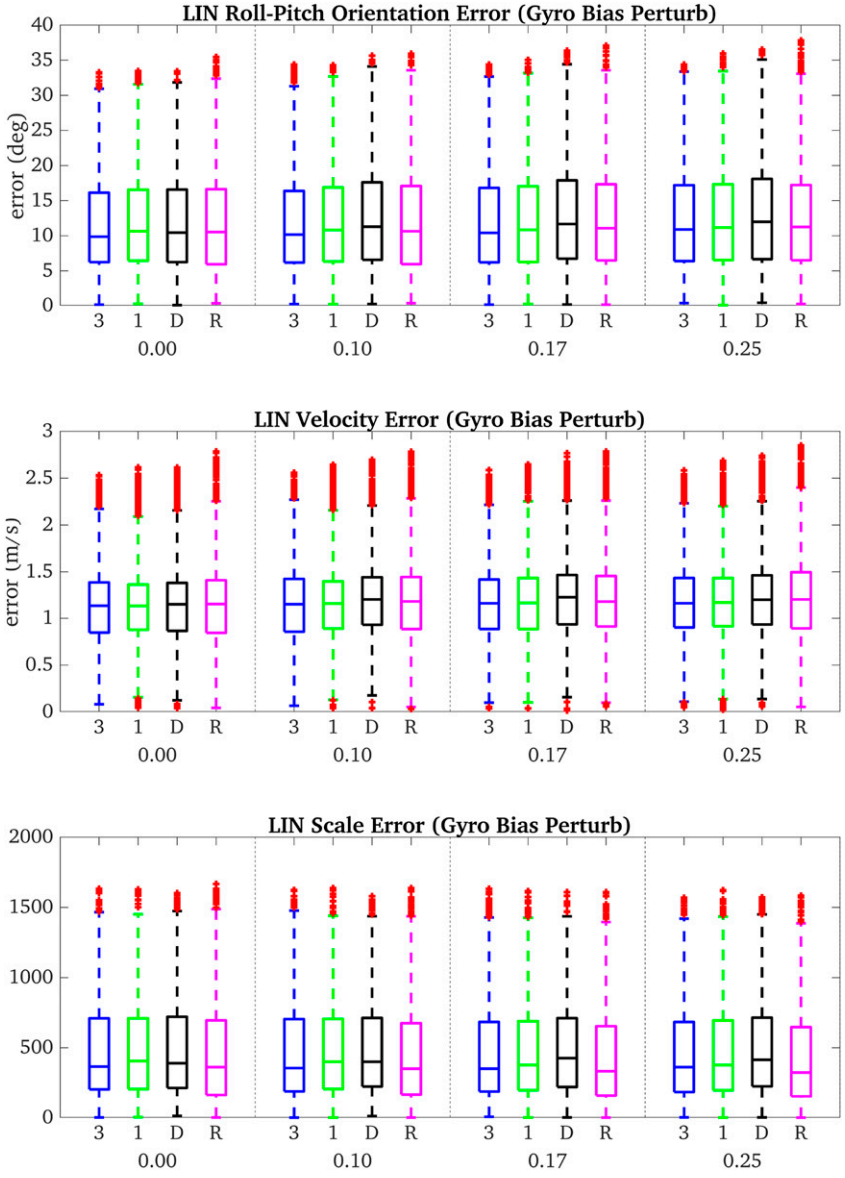

Figure 8. Monte-Carlo errors for orientation, velocity, and scale of the pose from the linear system for different constant **gyroscope bias** perturbations (in random direction). We define: DS 3D as **3** (blue), DS 1D as **1** (green), Ours w/o RANSAC as **D** (black), and Ours as **R** (magenta). Note that outliers outside the sample 3σ bound have been filtered for presentation clarity.

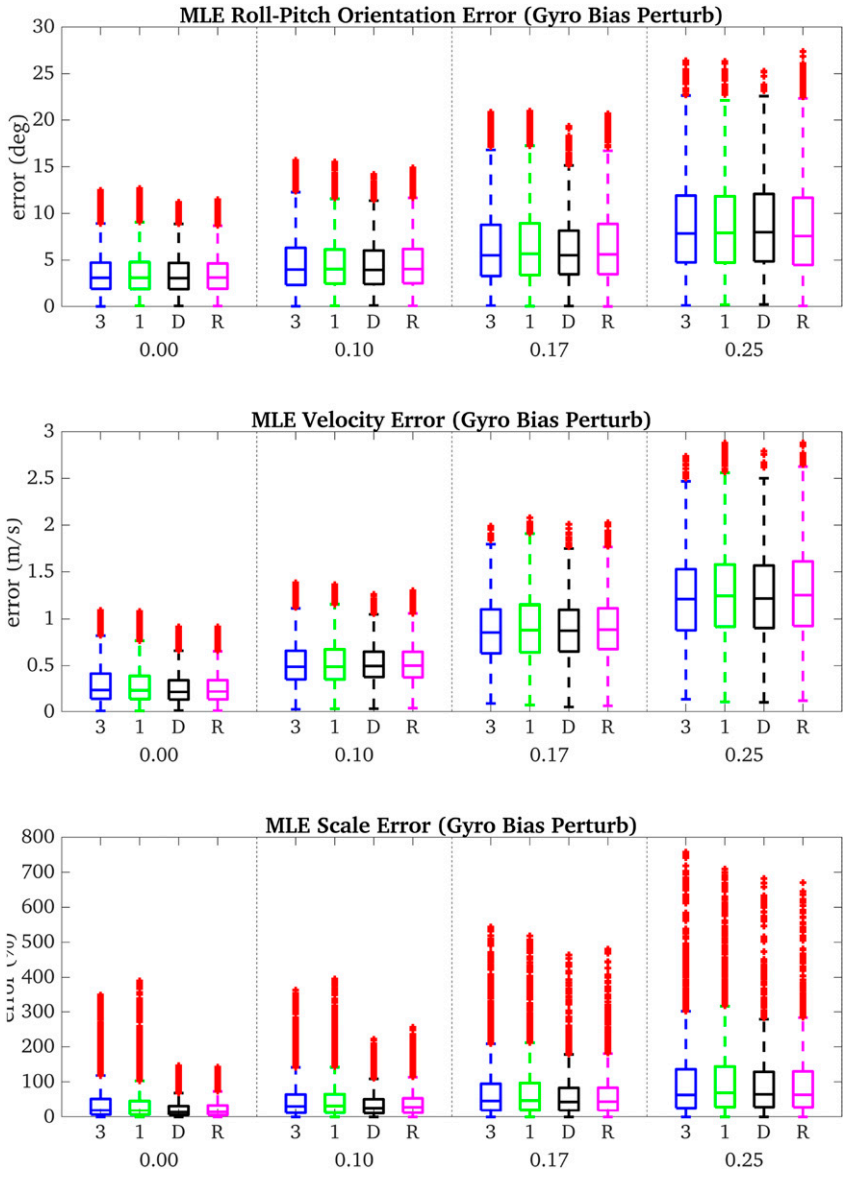

Figure 9. Monte-Carlo errors for orientation, velocity, and scale of the final pose after MLE refinement for different constant **gyroscope bias** perturbations (in random direction). We define: DS 3D as **3** (blue), DS 1D as **1** (green), Ours w/o RANSAC as **D** (black), and Ours as **R** (magenta). Note that outliers outside the sample 3σ bound have been filtered for presentation clarity.

whole trajectory at once, and might capture some insights that the ATE cannot.

7.1. Implementation details

Unless otherwise noted, 75 features on average are used during initialization. For the monocular depth network, we leverage an off-the-shelf pre-trained MiDaS network (the v2.1 small model) (Ranftl et al., 2022). This particular model is one of the most efficient available from the MiDaS model zoo, and is suitable to run on mobile devices. During all experiments, the network is run on the available GPU. Ceres solver (Agarwal et al., 2023) is used for all nonlinear optimizations. A separate thread is launched for initialization from the main tracking thread, but no extra thread is used to run the depth network asynchronously. While this could be done to improve initialization latency, we choose to simply run the network on-demand since it is *only* required to run once per initialization window (the first frame), unless the depth prior is used in the VI-BA, in which case it has to be run for each keyframe.

7.2. Baseline methods

For evaluation, we mainly consider two methods: (1) **DS 3D** and **DS 1D**, variants of Dong-Si and Mourikis (2012), and (2) **DS 3D + DP** and **DS 1D + DP** (DP standing for depth prior in the VI-BA), which is our re-implementation of Zhou et al. (2022) using the OpenVINS implementation of Dong-Si and Mourikis (2012) and the MiDaS v2.1 small network (Ranftl et al., 2022). Note that since we utilize MiDaS, which is completely affine-invariant

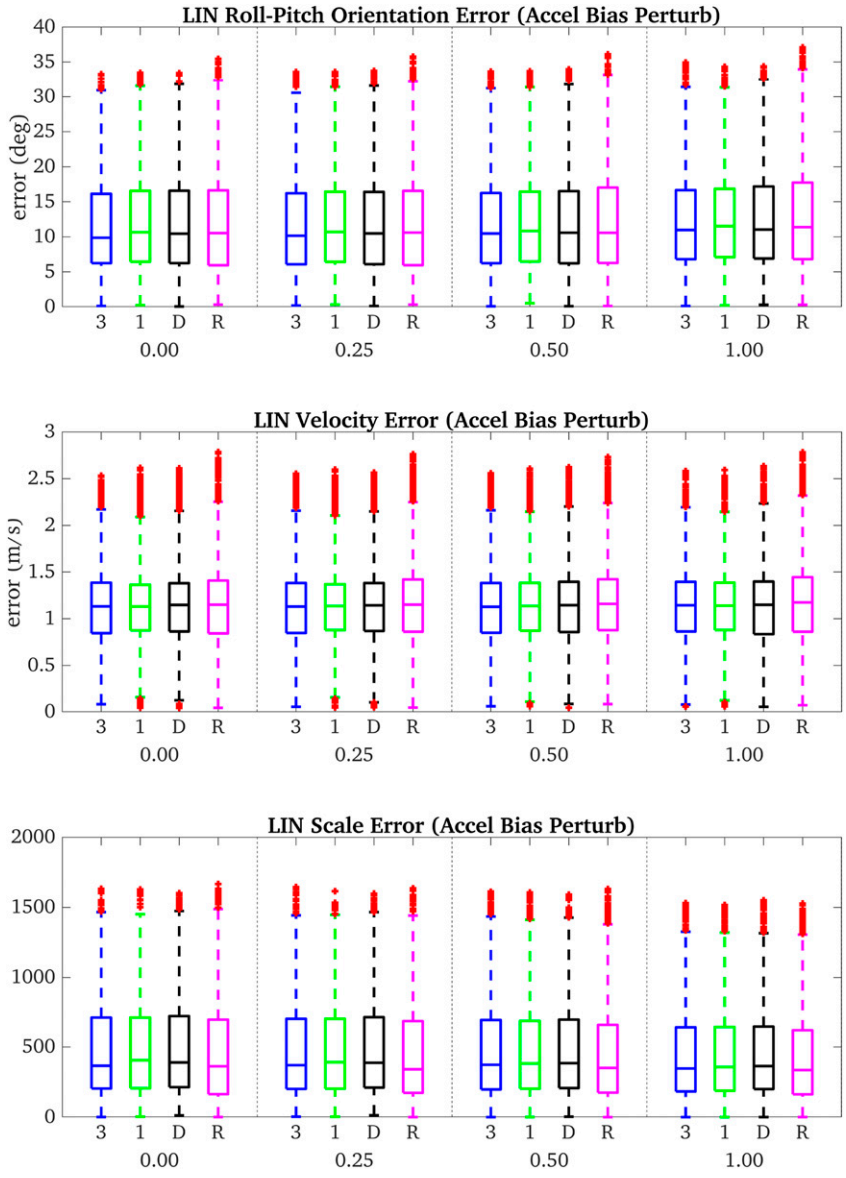

Figure 10. Monte-Carlo errors for orientation, velocity, and scale of the pose from the linear system for different constant **accelerometer bias** perturbations (in random direction). We define: DS 3D as **3**, DS 1D as **1**, Ours w/o RANSAC as **D**, and Ours as **R**. Note that outliers outside the sample 3σ bound have been filtered for presentation clarity.

(scale-less), as opposed to the custom depth network in Zhou et al. (2022), which is weakly-supervised with metric scale, we are unable to include the 1, 0 prior on the scale and bias (a and b) in the VI-BA. Including this prior could potentially improve the results, but it is unfortunately not applicable to MiDaS. Other than this difference, we strictly followed the formulation presented by Zhou et al. (2022) for this re-implementation. We also investigate the benefit of adding the additional depth prior to our method, which does require running the network for all keyframes rather than just the first one.

7.3. TUM-VI dataset

The first dataset we consider is the TUM-VI dataset (Schubert et al., 2018), where we only evaluate using the left fisheye image. While the MiDaS v2.1 small network was not explicitly trained on fisheye to our knowledge (some datasets used by MiDaS are proprietary), we observe that the network still produces reasonable depth maps when run on the raw fisheye images (which we prefer in order to maintain the full FoV). Some qualitative results of the raw MiDaS output can be seen in Figure 14. The results of the linear systems are reported in Table 4, where it can be seen that our method has less accurate pose accuracy coming from the linear system. However, as shown in simulation, our linear system typically produces more accurate feature positions, which unfortunately cannot be shown in the real world experiments due to a lack of ground-truth feature positions. All methods initialized 100% of the time here.

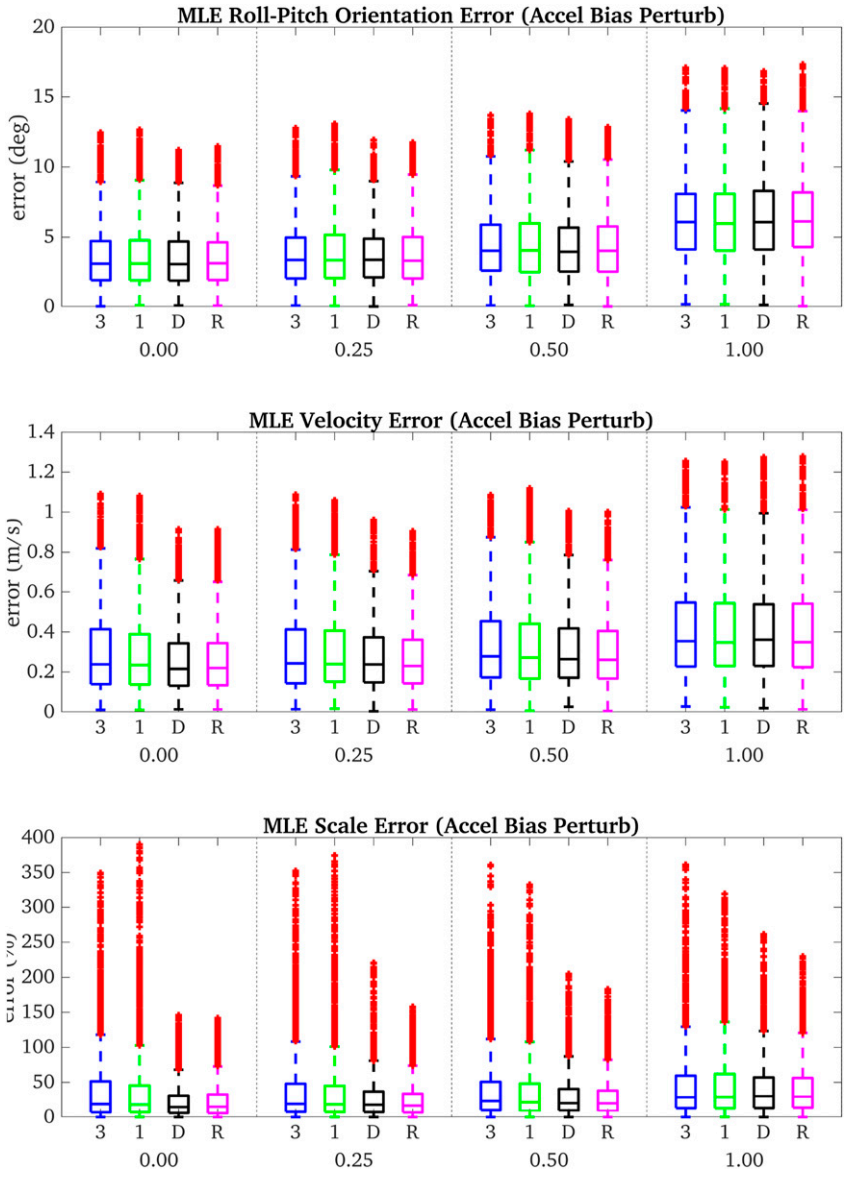

Figure 11. Monte-Carlo errors for orientation, velocity, and scale of the final pose after MLE refinement for different constant **accelerometer bias** perturbations (in random direction). We define: DS 3D as **3**, DS 1D as **1**, Ours w/o RANSAC as **D**, and Ours as **R**. Note that outliers outside the sample 3σ bound have been filtered for presentation clarity.

On the other hand, the results shown in Tables 5 and 6 show that the proposed method is able to achieve higher accuracy in the average for all metrics for the initialization window accuracy. One can also see that including the depth priors in the VI-BA, as in Zhou et al. (2022), improves over the baseline Dong-Si and Mourikis (2012) method as expected but is slightly less accurate than ours. Interestingly, adding the depth prior to our method (Ours + DP) does not improve over just using the depth in the linear system in this case.

In Table 7 and Figure 15, we report the VIO tracking accuracy using the initialization results. While the ATE results in Table 7 show that our method is not the best, the RPE results in Figure 15 show that our method has comparable RPE to the rest of the methods. Adding the additional depth prior to our method seems to improve the VIO performance on this dataset. All methods were successfully initialized for 100% (80/80) of the 10 s windows generated for this experiment.

To showcase the capability of our method to initialize with less information, we experiment with reducing the number of features being tracked during initialization. All experiments up until now have used 75 features, while here we experiment with 60, 45, 30, and 15 features—simulating a reduced number of available measurements due to low texture or other tracking failures. Table 8 reports the results. It is clear that our method is more robust to a low number of feature tracks available than the others, and that adding the depth prior to our method actually slightly hurts the performance.

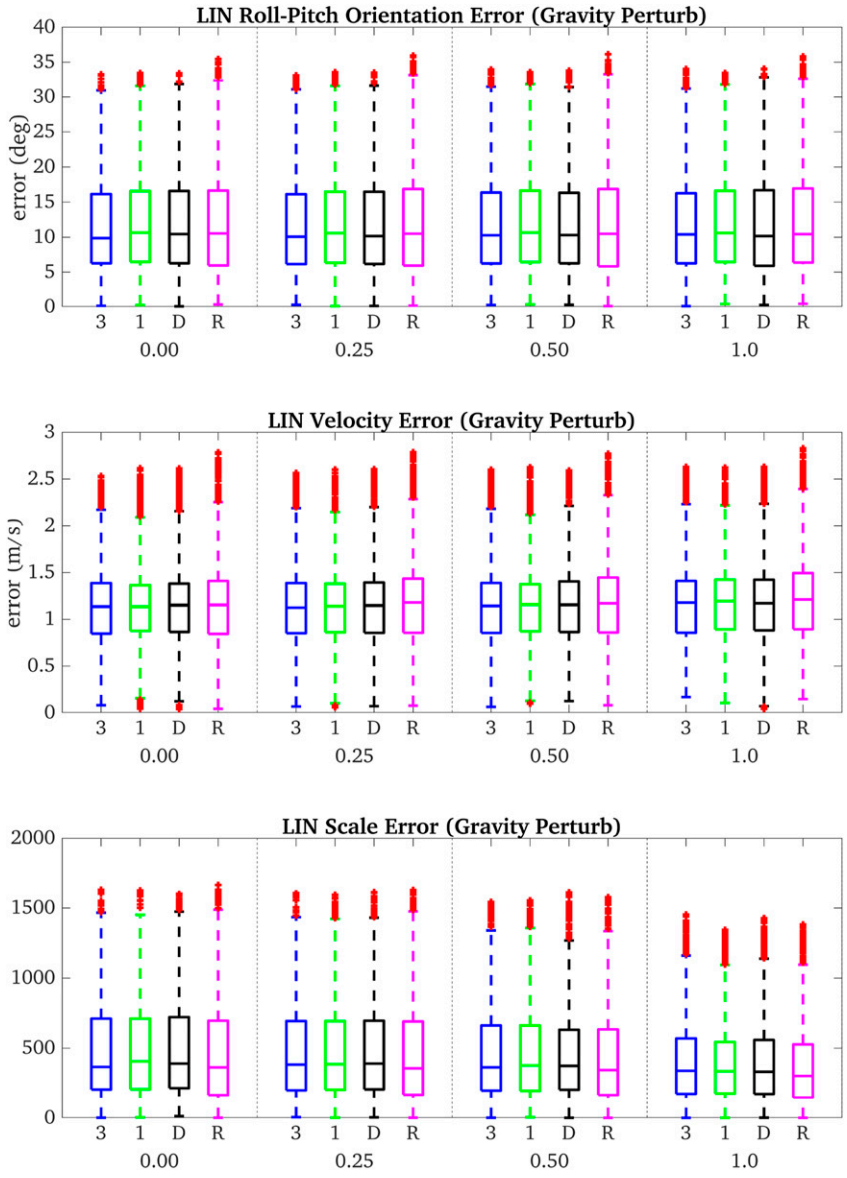

Figure 12. Monte-Carlo errors for orientation, velocity, and scale of the pose from the linear system for different constant **gravity magnitude** perturbations (random sign). We define: DS 3D as **3** (blue), DS 1D as **1** (green), Ours w/o RANSAC as **D** (black), and Ours as **R** (magenta). Note that outliers outside the sample 3σ bound have been filtered for presentation clarity.

7.3.1. Timing analysis. Here, we investigate the computational cost for the different initialization algorithms on the TUM-VI room1 dataset. Timings for a desktop device equipped with an Intel i5-6600K CPU and Nvidia RTX 2070 Super GPU are reported in Figure 16. In particular, we report the network inference time, building and solving the linear system, building and solving the optimization problem, and recovering the covariance. As expected, the proposed method is able to solve the linear system more efficiently due to the simplified linear model and the reduction of state size, but it should be noted that we do not take into account any sparsity when solving the linear system for any method. The depth network inference time is reasonably efficient given it only needs to be performed once for a 0.3–0.5 s window. The cost of building and solving the MLE problem is similar across methods, while the covariance recovery takes most of the time. We also timed the system on an embedded Jetson Orin device. The results are reported in Figure 17, where it can be seen that our method is overall more efficient than the baseline DS 3D.

7.3.2. Extreme low-parallax scenario. To further showcase the benefit of our method, we investigate a new and *even more* challenging scenario: initialization with 5 keyframes over a *0.3 s window*. To the best of our knowledge, this is the shortest initialization window *ever reported* for

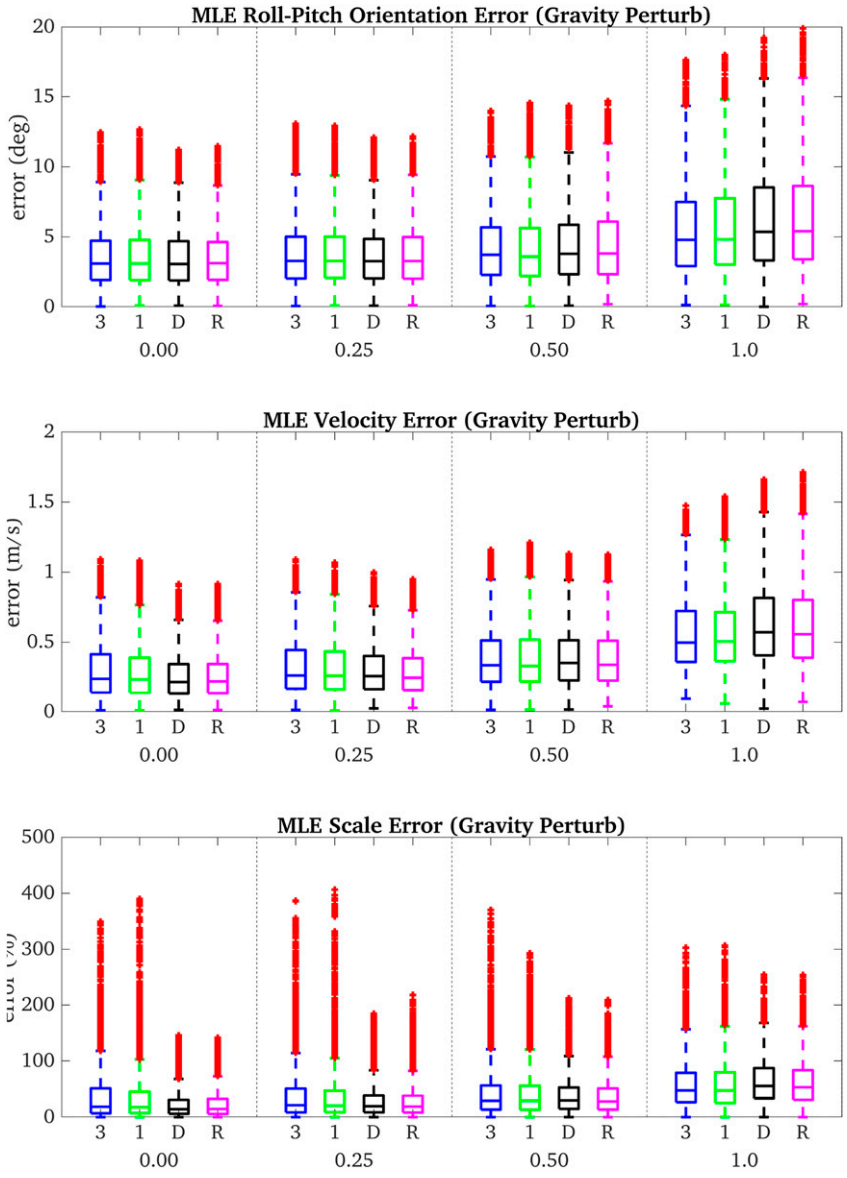

Figure 13. Monte-Carlo errors for orientation, velocity, and scale of the final pose after MLE refinement for different constant **gravity magnitude** perturbations (random sign). We define: DS 3D as **3** (blue), DS 1D as **1** (green), Ours w/o RANSAC as **D** (black), and Ours as **R** (magenta). Note that outliers outside the sample 3σ bound have been filtered for presentation clarity.

monocular VIO with unknown initial conditions. The linear system results are reported in Table 9. Again, our method is less accurate than the baselines from the linear system; however, DS 3D only initialized 78 out of 80 times here while the other methods were successful 100% of the time.

To evaluate the sensitivity of our method to different depth estimation networks, we additionally evaluate our method with the more accurate DepthAnything (Yang et al., 2024) ViT-Small network, which, while the most efficient of the DepthAnything networks, is far more computationally expensive than the MiDaS small network we employ. Denoted as "Ours (DA)," it can be seen in Table 9 that using the more-accurate DepthAnything network does not improve the result. This shows that our method is not highly-sensitive to the quality of the depth prediction.

Tables 10 and 11 report the ATE and scale error of the initialization window, respectively. The proposed method has overall superior orientation, position, and scale accuracy in the initialization window, and, again, utilizing the more-accurate DepthAnything network does not improve the performance. Table 12 and Figure 18 report the VIO tracking error. The VIO tracking accuracy for this extremely challenging scenario shows that the proposed method gains significant accuracy. Adding the depth prior to our method achieves a slight improvement in ATE (for the orientation) but actually slightly worse RPE. Not all

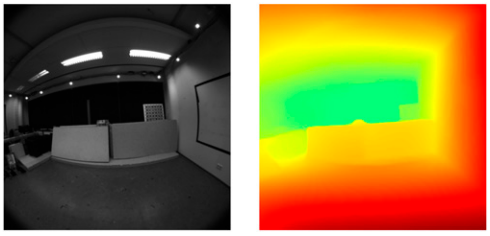

Figure 14. Qualitative result of the MiDaS (Ranftl et al., 2022) v2.1 small on the raw fisheye images of TUM-VI. We found that the network produces reasonable depth maps despite not being explicitly trained for this camera model; however, training the network with fisheye data could potentially improve performance.

methods successfully initialized in every run in this experiment. Both DS 3D methods (with and without depth prior in VI-BA) were successful 78 times, while all other methods (including the proposed methods) successfully initialized 80 times out of the 80 10 s windows over the datasets.

Table 13 reports the results of reducing the number of feature tracks. The proposed method can tolerate a severe reduction in the number of features available, while the proposed RANSAC method can still outperform the baselines and, as shown in the next experiment, remain robust to outliers. Adding the depth prior helped slightly in this case.

Table 4. Initialization Window ATE (deg/m) From the Linear System on TUM-VI (5 KFs, 0.5 s Window).

Algorithm	room1	room2	room3	room4	room5	room6	Average
DS 3D	4.910/0.064	2.615/0.086	2.649/0.126	1.702/0.077	6.587/0.140	2.796/0.096	**3.543/0.098**
DS 1D	5.851/0.079	3.032/0.105	3.008/0.136	2.975/0.090	6.985/0.143	2.891/0.102	4.124/0.109
Ours w/o RANSAC	14.835/0.273	5.319/0.237	5.689/0.256	4.326/0.132	9.830/0.236	3.573/0.137	7.262/0.212
Ours	13.397/0.242	5.169/0.223	5.342/0.242	3.865/0.110	8.287/0.182	3.426/0.123	6.581/0.187

Table 5. Initialization Window ATE (deg/m) on TUM-VI After VI-BA (5 KFs, 0.5 s Window).

Algorithm	room1	room2	room3	room4	room5	room6	Average
DS 3D	1.707/0.021	0.650/0.011	0.819/0.014	1.371/0.035	0.849/0.015	0.546/0.010	0.990/0.018
DS 1D	1.103/0.011	0.609/0.011	0.742/0.010	1.916/0.040	1.028/0.013	0.457/0.010	0.976/0.016
DS 3D + DP	0.899/0.009	0.648/0.011	0.819/0.014	1.372/0.035	0.852/0.015	0.493/0.012	0.847/0.016
DS 1D + DP	2.458/0.020	0.944/0.011	2.084/0.019	1.327/0.013	1.961/0.014	0.569/0.008	1.557/0.014
Ours w/o RANSAC	0.852/0.012	0.596/0.010	0.709/0.009	0.785/0.008	1.145/0.016	0.440/0.009	**0.754/0.011**
Ours	0.866/0.011	0.650/0.010	0.718/0.008	0.814/0.011	1.292/0.015	0.436/0.009	0.796/**0.011**
Ours + DP	0.869/0.011	0.711/0.012	0.725/0.008	0.820/0.010	1.318/0.016	0.447/0.008	0.815/**0.011**

Table 6. Initialization Window Scale Error (%) on TUM-VI After VI-BA (5 KFs, 0.5 s Window).

Algorithm	room1	room2	room3	room4	room5	room6	Average
DS 3D	4.693	2.807	0.433	3.774	4.165	2.324	3.033
DS 1D	2.577	2.216	0.563	1.364	4.010	2.177	2.151
DS 3D + DP	2.012	2.617	0.723	6.885	4.093	0.568	2.816
DS 1D + DP	0.616	2.023	19.871	2.518	8.837	3.529	6.232
Ours w/o RANSAC	0.404	1.450	0.534	1.349	3.753	1.746	**1.539**
Ours	0.471	2.746	0.513	0.490	4.600	2.310	1.855
Ours + DP	0.520	1.944	0.925	4.425	4.664	2.055	2.422

Table 7. Visual-Inertial Odometry Tracking ATE (deg/m) on TUM-VI (5 KFs, 0.5 s Window for init).

Algorithm	room1	room2	room3	room4	room5	room6	Average
DS 3D	1.478/0.140	0.809/0.036	1.384/0.056	2.047/0.178	0.971/0.047	1.809/0.482	1.417/0.156
DS 1D	1.285/0.096	0.789/0.037	1.319/0.051	2.305/0.203	0.958/0.049	1.782/0.432	1.406/0.145
DS 3D + DP	1.558/0.150	0.814/0.037	1.113/0.046	2.159/0.339	0.919/0.046	1.364/0.130	**1.321**/0.125
DS 1D + DP	1.250/0.065	0.820/0.039	1.715/1.028	1.561/0.083	2.692/0.713	0.955/0.073	1.499/0.334
Ours w/o RANSAC	1.257/0.101	0.797/0.037	1.317/0.051	1.691/0.185	1.019/0.053	1.842/0.361	**1.321**/0.131
Ours	1.417/0.184	0.806/0.038	1.354/0.049	2.125/0.296	1.099/0.057	1.836/0.360	1.440/0.164
Ours + DP	1.437/0.128	0.818/0.036	1.408/0.050	1.872/0.248	1.028/0.058	1.565/0.222	1.355/**0.124**

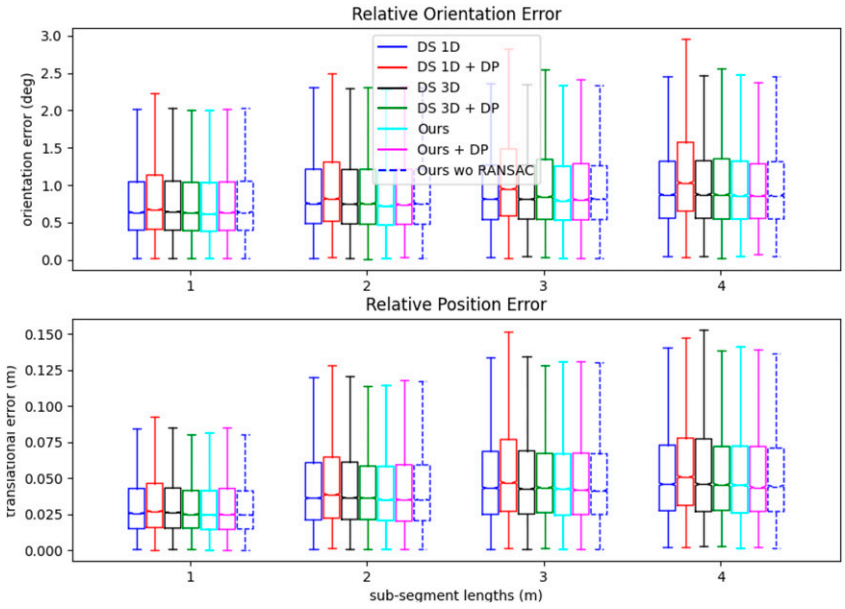

Figure 15. RPE for VIO tracking on TUM-VI with 5 KFs and 0.5 s window.

Table 8. Percent of Successful Initializations on TUM-VI (Averaged Over All Rooms) With 5 KFs and 0.5 s Window.

Algorithm	60 feats	45 feats	30 feats	15 feats
DS 3D	**100.0**	**100.0**	**100.0**	68.8
DS 1D	**100.0**	**100.0**	**100.0**	78.8
DS 3D + DP	**100.0**	**100.0**	**100.0**	67.5
DS 1D + DP	**100.0**	**100.0**	**100.0**	78.8
Ours w/o RANSAC	**100.0**	**100.0**	**100.0**	**88.8**
Ours	**100.0**	**100.0**	**100.0**	**88.8**
Ours + DP	**100.0**	**100.0**	**100.0**	87.5

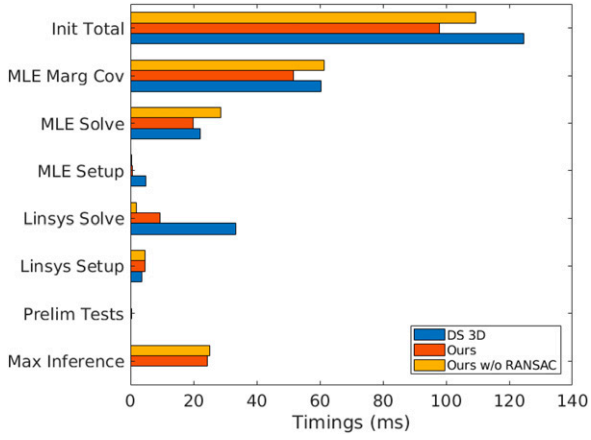

Figure 16. Timing results on a desktop device on the TUM-VI room1 sequence.

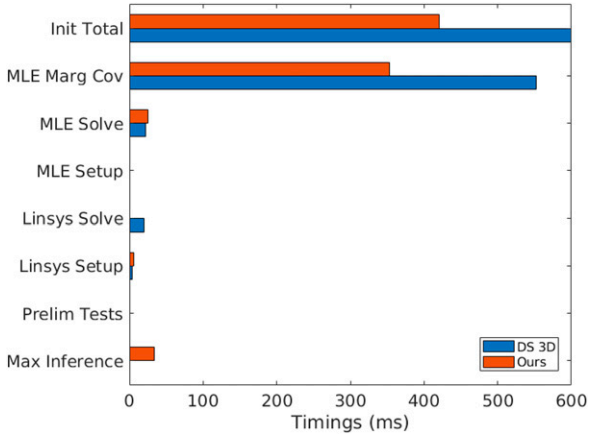

Figure 17. Timing results on an embedded Jetson Orin device on the TUM-VI room1 sequence.

7.3.3. Robustness to outliers. We additionally investigate how robust the proposed RANSAC method is to outliers. Given a set of features selected for initialization, a percent of them are selected to be outliers. All observations for these features are perturbed with a normally distributed 10px feature distribution. The mixture of inlier and outlier features is then fed into the rest of the initialization process.

Table 9. Initialization Window ATE (deg/m) From the Linear System on TUM-VI With Extreme Settings (5 KFs, 0.3 s Window).

Algorithm	room1	room2	room3	room4	room5	room6	Average
DS 3D	5.922/0.111	2.939/0.115	5.412/0.130	1.806/0.070	7.532/0.114	2.417/0.070	**4.338**/0.102
DS 1D	8.013/0.106	4.872/0.115	5.101/0.112	3.412/0.075	6.382/0.085	2.791/0.065	5.095/**0.093**
Ours w/o RANSAC	15.529/0.145	5.320/0.140	7.010/0.141	5.860/0.088	11.193/0.120	3.347/0.073	8.043/0.118
Ours	15.123/0.180	4.796/0.145	6.957/0.157	4.752/0.092	9.976/0.129	3.307/0.080	7.485/0.131
Ours (DA)	15.244/0.155	5.166/0.146	6.892/0.148	4.170/0.094	10.673/0.140	3.169/0.079	7.552/0.127

Table 10. Initialization Window ATE (deg/m) on TUM-VI After VI-BA With Extreme Settings (5 KFs, 0.3 s Window).

Algorithm	room1	room2	room3	room4	room5	room6	Average
DS 3D	1.475/0.026	1.002/0.011	2.021/0.019	0.673/0.024	1.545/0.017	0.738/0.014	1.243/0.018
DS 1D	2.548/0.020	0.940/0.011	2.167/0.020	1.142/0.013	2.848/0.020	0.556/0.008	1.700/0.015
DS 3D + DP	1.523/0.026	1.014/0.012	2.022/0.019	0.681/0.024	1.675/0.023	0.712/0.014	1.271/0.020
DS 1D + DP	2.458/0.020	0.943/0.009	2.083/0.019	1.327/0.013	1.964/0.014	0.570/0.008	1.557/0.014
Ours w/o RANSAC	1.670/0.010	0.660/0.007	1.345/0.010	1.879/0.017	1.149/0.010	0.695/0.010	1.233/**0.011**
Ours	1.546/0.014	0.841/0.008	1.829/0.016	0.839/0.010	1.459/0.010	0.681/0.012	**1.199**/**0.011**
Ours (DA)	2.106/0.014	0.814/0.007	1.673/0.012	0.822/0.016	1.247/0.010	0.652/0.008	1.219/**0.011**
Ours + DP	1.416/0.013	0.840/0.010	1.796/0.013	1.248/0.014	1.475/0.010	0.698/0.012	1.245/0.012

Table 11. Initialization Window Scale Error (%) on TUM-VI After VI-BA With Extreme Settings (5 KFs, 0.3 s Window).

Algorithm	room1	room2	room3	room4	room5	room6	Average
DS 3D	13.686	1.245	19.176	8.074	4.360	8.644	9.197
DS 1D	1.165	2.120	24.799	4.230	2.939	3.739	6.499
DS 3D + DP	13.913	1.113	19.183	8.699	0.973	9.304	8.864
DS 1D + DP	0.616	1.060	19.883	2.541	8.830	3.615	6.091
Ours w/o RANSAC	9.765	0.784	5.485	11.095	8.504	6.271	6.984
Ours	1.607	0.015	7.508	7.817	10.232	8.169	5.891
Ours (DA)	8.558	1.235	7.366	3.653	9.689	0.243	**5.124**
Ours + DP	1.361	2.076	7.274	15.155	10.870	5.175	6.985

Table 12. Visual-Inertial Odometry Tracking ATE (deg/m) on TUM-VI With Extreme Settings (5 KFs, 0.3 s Window for init).

Algorithm	room1	room2	room3	room4	room5	room6	Average
DS 3D	1.255/0.210	0.859/0.043	1.445/0.059	2.318/0.368	1.457/0.045	0.946/0.074	1.380/0.133
DS 1D	1.648/0.284	0.822/0.039	1.870/1.029	1.739/0.103	3.155/0.535	0.970/0.071	1.701/0.343
DS 3D + DP	1.246/0.205	0.835/0.043	1.535/0.061	2.298/0.311	1.427/0.050	0.976/0.076	1.386/0.124
DS 1D + DP	1.251/0.065	0.815/0.039	1.728/1.028	1.574/0.083	2.701/0.712	0.955/0.074	1.504/0.334
Ours w/o RANSAC	1.110/0.073	0.831/0.043	1.641/0.069	1.851/0.096	2.194/0.256	0.919/0.074	1.424/0.102
Ours	0.986/0.037	0.833/0.043	1.762/0.075	1.488/0.080	1.024/0.035	0.845/0.052	1.156/**0.054**
Ours (DA)	1.189/0.077	0.796/0.038	2.537/0.296	1.166/0.113	1.624/0.078	0.896/0.058	1.368/0.110
Ours + DP	0.988/0.036	0.882/0.043	1.476/0.069	1.717/0.086	0.963/0.036	0.845/0.053	**1.145**/**0.054**

Shown in Table 14, as the outlier percentage increases the proposed initialization method with and without the additional depth prior are the least affected by the added outliers. The proposed RANSAC method is able to robustly provide reliable initial guesses even in the case of 40% outlier features. We stress that this RANSAC formulation is only enabled by leveraging the affine-invariant depth map to ensure the state remains

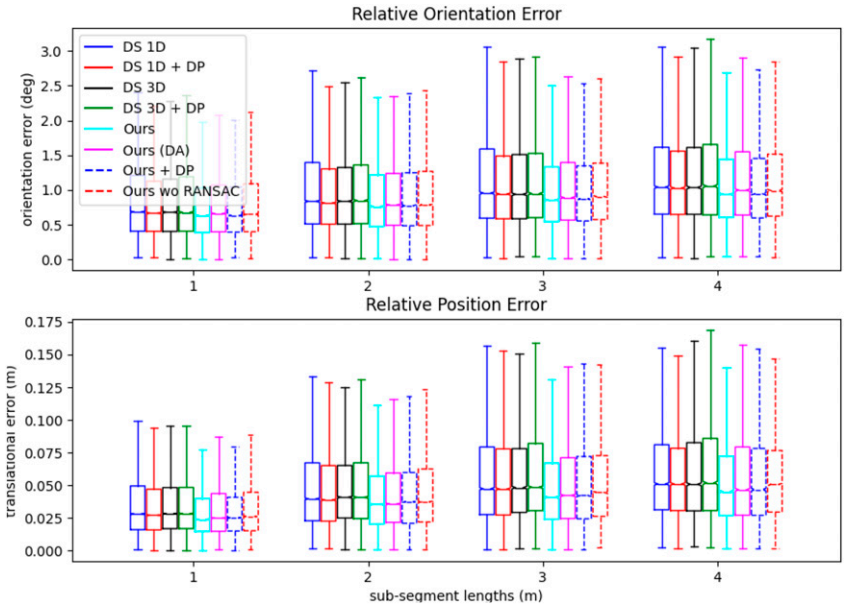

Figure 18. RPE for VIO tracking on TUM-VI with 5 KFs and 0.3 s window.

Table 13. Percent of Successful Initializations on TUM-VI (Averaged Over all Rooms) With 5 KFs and 0.3 s Window.

Algorithm	60 feats	45 feats	30 feats	15 feats
DS 3D	81.3	17.5	33.8	2.5
DS 1D	**100.0**	81.3	82.5	26.3
DS 3D + DP	78.8	16.3	32.5	2.5
DS 1D + DP	**100.0**	80.0	82.5	25.0
Ours w/o RANSAC	**100.0**	**98.8**	**97.5**	**55.0**
Ours	**100.0**	95.0	96.3	47.5
Ours + DP	**100.0**	95.0	96.3	50.0

independent to the number of features, and thus is unique to our formulation.

7.4. EuRoC MAV dataset

We next evaluate on the EuRoC MAV dataset (Burri et al., 2016). In this section, we also include a direct comparison to the state-of-the-art work by Zhou et al. (2022) (denoted as Zhou et al. (2022)). This comparison is only partial since the implementation of Zhou et al. (2022) is not open-sourced, thus we are forced to quote results from the paper where applicable. We measure the full orientation error and scale error over the whole trajectory rather than just the gravity and scale error over well-excited trajectory segments, and thus cannot directly compare to their orientation and scale. We selected the closest equivalent challenging configuration of of 5 KFs evenly spaced over a 0.5 s window.

7.4.1. Vicon Room sequences. We first evaluate the system on the EuRoC Vicon Room sequences. Similar to the TUM-VI dataset, these sequences take place in a small room equipped with a motion capture device. Figure 19 shows some qualitative results of the depth network's performance on these sequences. Despite the network being confused about some unusual strips on the floor, our method still performs well due to the incorporation of RANSAC to reject these bad depth points.

We report the results of the linear system solutions (no VI-BA refinement) in Table 15. Again, our method is less accurate than the baselines. All systems successfully initialized 100% of the time in this experiment.

Looking now to results which perform the VI-BA refinement after closed-form recovery, Tables 16 and 17 report the ATE and scale error, respectively. We can observe that our system outperforms all the baselines in the average case, and adding the depth prior helped in this case. We can see that the proposed system without RANSAC enabled (i.e., using all available measurements outlier or not) hurts the performance, while leveraging RANSAC has improved scale and ATE accuracy. All methods successfully initialized 65 out of 65 10 s windows in this experiment except for DS 3D + DP, which succeeded 64 times.

The results of VIO tracking are also reported in Table 18 and Figure 20. Our methods (with and without the depth prior in the VI-BA) are shown to be the most accurate out of

Table 14. Outlier Ablation Study of VIO Tracking ATE (deg/m) on TUM-VI Dataset After VI-BA With Extreme Settings (5 KFs and 0.3 s Window).

Outliers (%)	Algorithm	room1	room2	room3	room4	room5	room6	Average
5	DS 3D	1.169/0.054	0.894/0.043	1.361/0.052	2.843/0.999	1.455/0.070	1.003/0.069	1.454/0.215
	DS 1D	1.581/0.840	0.802/0.043	1.521/0.072	4.186/0.848	1.406/0.078	1.274/0.087	1.795/0.328
	DS 3D + DP	1.097/0.054	0.866/0.043	2.204/0.582	2.882/1.095	1.432/0.067	0.981/0.067	1.577/0.318
	DS 1D + DP	1.503/0.555	0.844/0.044	1.429/0.069	2.178/0.315	1.407/0.923	1.277/0.087	1.440/0.332
	Ours w/o RANSAC	1.012/0.059	0.797/0.042	1.967/0.092	2.214/0.188	1.409/2.343	0.854/0.045	1.375/0.461
	Ours	1.104/0.053	0.892/0.039	1.321/0.042	1.297/0.132	1.582/0.043	0.764/0.038	**1.160/0.058**
	Ours + DP	1.189/0.055	0.894/0.041	1.309/0.045	2.306/0.212	1.172/0.038	0.770/0.040	1.273/0.072
10	DS 3D	0.919/0.044	1.107/0.096	1.723/0.059	1.754/0.270	1.405/0.083	1.003/0.071	**1.318**/0.104
	DS 1D	1.496/0.428	0.794/0.043	1.576/0.086	1.520/0.498	1.519/2.235	1.222/0.094	1.355/0.564
	DS 3D + DP	1.017/0.049	1.101/0.098	1.648/0.059	1.880/0.278	1.538/0.089	1.039/0.073	1.371/0.108
	DS 1D + DP	2.522/3.080	0.866/0.047	1.605/0.085	1.231/0.103	1.856/0.952	1.271/0.097	1.559/0.727
	Ours w/o RANSAC	1.536/0.076	0.827/0.039	1.244/0.048	2.044/0.326	2.710/1.298	1.207/1.458	1.595/0.541
	Ours	1.585/0.071	0.908/0.033	1.271/0.048	2.236/0.430	3.626/0.237	0.717/0.056	1.724/0.146
	Ours + DP	1.626/0.076	0.947/0.035	1.228/0.049	1.836/0.164	3.616/0.223	0.703/0.068	1.660/**0.102**
25	DS 3D	3.760/8.652	0.964/0.082	2.233/1.848	2.922/0.663	1.816/0.357	1.171/0.105	2.144/1.951
	DS 1D	3.482/6.572	1.017/0.070	1.563/0.700	2.187/0.727	2.581/1.120	2.279/0.288	2.185/1.580
	DS 3D + DP	3.591/8.554	0.968/0.103	1.819/0.979	1.327/0.134	1.407/0.080	1.067/0.064	**1.697**/1.652
	DS 1D + DP	3.676/6.625	0.926/0.070	1.573/0.682	1.993/0.558	2.571/1.130	2.261/0.290	2.167/1.559
	Ours w/o RANSAC	3.224/4.804	1.199/0.086	1.578/0.137	3.341/0.291	4.613/3.992	2.173/2.651	2.688/1.993
	Ours	1.055/0.044	2.972/2.152	1.322/0.063	1.958/1.053	1.504/1.173	3.933/1.160	2.124/0.941
	Ours + DP	1.110/0.045	3.245/1.291	1.209/0.039	1.395/0.162	1.594/1.407	4.087/1.209	2.107/**0.692**
40	DS 3D	5.946/11.371	1.263/0.158	1.561/1.588	1.351/0.081	3.269/1.177	1.706/0.715	**2.516**/2.515
	DS 1D	7.338/14.876	2.012/0.236	2.559/0.144	2.588/0.536	2.622/2.350	4.458/3.226	3.596/3.561
	DS 3D + DP	5.737/11.216	1.574/0.171	1.578/1.586	2.851/0.161	4.240/5.030	1.700/0.709	2.947/3.145
	DS 1D + DP	8.117/14.498	1.912/0.177	3.176/0.264	2.085/0.331	2.701/2.373	4.471/3.227	3.744/3.478
	Ours w/o RANSAC	3.772/3.992	1.945/0.621	2.013/1.427	8.966/5.144	2.448/1.921	5.531/2.980	4.112/2.681
	Ours	1.490/0.066	1.486/0.601	3.124/3.827	5.515/3.284	1.880/0.066	4.683/2.300	3.030/**1.691**
	Ours + DP	1.453/0.070	1.913/0.666	3.145/3.942	6.164/6.042	3.138/2.426	2.599/1.600	3.069/2.458

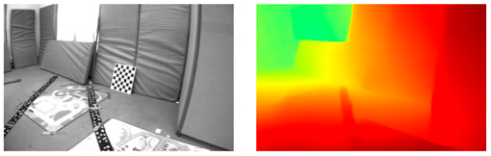

Figure 19. Qualitative result of the MiDaS Ranftl et al. (2022) v2.1 small on the EuRoC Vicon Room 1 sequence. The network can easily get confused about the unusual decor, such as on strips the floor.

the ones implemented on top of OpenVINS, while Zhou et al. (2022) is the most accurate out of all the methods. It is tough to know if this accuracy gain is due to the feature tracking front-end or a difference in evaluation due to the closed-source nature of Zhou et al. (2022) and the similar error magnitude levels achieved by all re-implemented methods which build on top of the open-sourced OpenVINS (Geneva et al., 2019).

Table 15. Initialization Window ATE (deg/m) From the Linear System on EuRoC Vicon Room (5 KFs, 0.5 s Window).

Algorithm	V101	V102	V103	V201	V202	V203	Average
DS 3D	1.725/0.099	2.660/0.133	5.614/0.097	1.404/0.035	2.832/0.099	3.719/0.116	**2.992/0.096**
DS 1D	1.767/0.104	2.769/0.149	5.998/0.109	1.428/0.038	3.107/0.115	4.209/0.142	3.213/0.110
Ours w/o RANSAC	1.892/0.125	5.358/0.225	7.256/0.157	1.682/0.061	4.943/0.164	4.796/0.194	4.321/0.154
Ours	1.949/0.123	4.903/0.219	7.269/0.152	1.656/0.059	4.470/0.156	5.452/0.189	4.283/0.150

Table 16. Initialization Window ATE (deg/m) on EuRoC Vicon Room After VI-BA (5 KFs, 0.5 s Window).

Algorithm	V101	V102	V103	V201	V202	V203	Average
DS 3D	1.317/0.042	0.797/0.021	1.611/0.018	0.931/0.010	1.454/0.028	1.933/0.040	1.340/0.027
DS 1D	1.034/0.029	0.817/0.019	1.544/0.017	0.940/0.009	1.671/0.027	1.697/0.029	1.284/0.022
DS + DP	1.322/0.041	0.867/0.022	1.989/0.026	0.940/0.010	1.470/0.029	1.131/0.036	1.286/0.028
Zhou et al. (2022)[a]	-/0.021	-/0.038	-/0.025	-/0.015	-/0.015	-/0.033	-/0.024
Ours w/o RANSAC	0.998/0.020	0.751/0.013	1.695/0.018	0.924/0.010	2.939/0.079	1.613/0.030	1.487/0.028
Ours	0.998/0.020	0.734/0.013	1.436/0.016	0.936/0.010	2.008/0.045	1.488/0.029	1.267/0.022
Ours + DP	0.969/0.019	0.780/0.013	1.477/0.017	0.926/0.010	1.786/0.024	1.119/0.036	**1.176/0.020**

[a]Results quoted from Table 1 in Zhou et al. (2022).

Table 17. Initialization Window Scale Error (%) on EuRoC Vicon Room After VI-BA (5 KFs, 0.5 s Window).

Algorithm	V101	V102	V103	V201	V202	V203	Avg
DS 3D	7.014	8.953	6.700	2.958	0.489	41.604	11.286
DS 1D	3.249	4.553	9.238	2.871	0.710	35.911	9.422
DS + DP	6.795	4.887	12.009	3.176	0.169	32.385	9.903
Ours w/o RANSAC	1.361	0.423	4.020	2.934	29.629	10.208	8.096
Ours	1.438	0.255	4.195	3.116	10.005	20.806	6.636
Ours + DP	2.661	0.113	3.580	2.956	1.142	25.462	**5.986**

Table 18. Visual-Inertial Odometry Tracking ATE (deg/m) on EuRoC Vicon Room (5 KFs, 0.5 s Window for init).

Algorithm	V101	V102	V103	V201	V202	V203	Average
DS 3D	1.821/1.101	1.279/0.096	2.961/0.424	1.630/0.074	1.912/0.099	4.756/4.479	2.393/1.046
DS 1D	1.394/0.168	1.273/0.097	3.498/0.504	1.589/0.073	2.016/0.099	4.995/5.793	2.461/1.122
DS + DP	2.070/0.962	1.389/0.095	3.144/1.229	1.673/0.075	1.975/0.110	4.970/2.968	2.537/0.907
Zhou et al. (2022)[a]	-/0.082	-/0.097	-/0.059	-/0.046	-/0.060	-/0.567	-/**0.152**
Ours w/o RANSAC	1.063/0.087	1.358/0.115	2.931/0.492	1.565/0.075	5.088/5.425	2.991/2.203	2.499/1.400
Ours	1.060/0.088	1.417/0.117	2.191/0.175	1.611/0.077	3.318/0.646	3.337/3.914	2.156/0.836
Ours + DP	1.070/0.089	1.422/0.105	2.269/0.171	1.574/0.073	2.053/0.106	4.192/3.115	**2.097**/0.610

[a]Results quoted from Table 3 in Zhou et al. (2022).

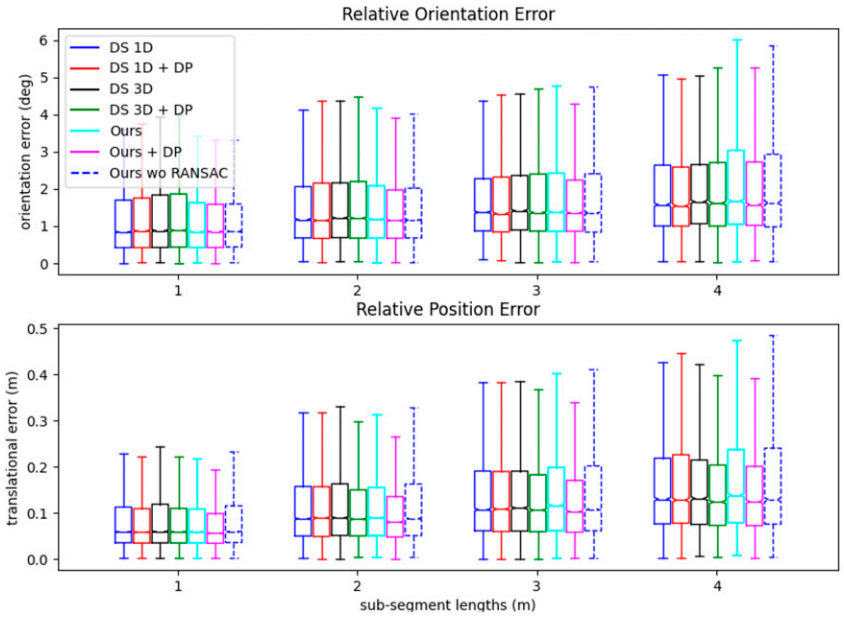

Figure 20. RPE for VIO tracking on EuRoC Vicon Room with 5 KFs and 0.5 s window.

Table 19. Percent of Successful Initializations on EuRoC Vicon Room (Averaged Over all Sequences) With 5 KFs and 0.5 s Window.

Algorithm	60 feats	45 feats	30 feats	15 feats
DS 3D	**100.0**	**100.0**	**100.0**	76.9
DS 1D	**100.0**	**100.0**	**100.0**	93.8
DS 3D + DP	**100.0**	**100.0**	**100.0**	76.9
DS 1D + DP	98.5	**100.0**	**100.0**	93.8
Ours w/o RANSAC	**100.0**	**100.0**	**100.0**	**95.4**
Ours	**100.0**	**100.0**	**100.0**	**95.4**
Ours + DP	**100.0**	**100.0**	**100.0**	**95.4**

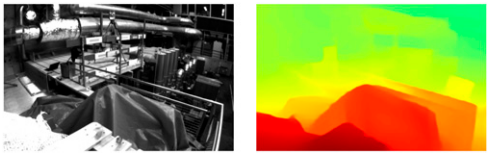

Figure 21. Qualitative result of the MiDaS Ranftl et al. (2022) v2.1 small on the EuRoC Machine Hall 1 sequence. Larger scale and small objects (e.g., pipes) make these sequences challenging for the depth network.

The robustness to reduced number of features is also reported in this section. Table 19 reports the results. It can be seen that all three of our methods (with and without RANSAC as well as with depth prior) are the most robust to reduced number of features in this case, especially with very low number of feature tracks available.

7.4.2. Machine Hall sequences. We additionally evaluate on the Machine Hall sequences of the EuRoC dataset. These sequences are more challenging than the Vicon Room due to the larger scale of the scene, and high amounts of clutter and small objects such as pipes which are challenging for low-resolution dense reconstruction methods. Qualitative results of the depth network's performance on these sequences can be seen in Figure 21.

In this section, we will show an interesting case where our method is less accurate by all measures except for VIO tracking accuracy, the most important metric, where it is tied for the best and even beats the state-of-the-art baseline Zhou et al. (2022). This shows that the other metrics, such as initialization window accuracy, may not always show the true performance of a visual-inertial initialization system.

First, the linear system is evaluated (without BA refinement). Table 20 reports the results. As usual, our linear system pose accuracy is worse than the baselines. All systems successfully initialized 100% of the time in this experiment.

We also evaluate the performance after performing the VI-BA. The results of the initialization window can be seen in Tables 21 and 22, which show the ATE and scale error, respectively. Our method is not the best in terms of the initialization window accuracy. However, in Table 23 it can be seen that our method has some of the best accuracy for VIO tracking—the best orientation and second best position. The RPE results of VIO tracking are reported in

Table 20. Initialization Window ATE (deg/m) From the Linear System on EuRoC Machine Hall (5 KFs, 0.5 s Window).

Algorithm	MH01	MH02	MH03	MH04	MH05	Average
DS 3D	1.507/0.071	2.062/0.066	3.127/0.175	1.877/0.169	2.380/0.215	**2.190**/**0.139**
DS 1D	1.626/0.077	2.229/0.073	3.327/0.198	1.982/0.178	2.505/0.225	2.334/0.150
Ours w/o RANSAC	2.179/0.110	2.853/0.102	4.300/0.305	2.263/0.197	2.957/0.273	2.910/0.197
Ours	2.080/0.101	2.747/0.097	4.408/0.286	2.302/0.197	2.807/0.270	2.869/0.190

Table 21. Initialization Window ATE (deg/m) on EuRoC Machine Hall After VI-BA (5 KFs, 0.5 s Window).

Algorithm	MH01	MH02	MH03	MH04	MH05	Average
DS 3D	1.143/0.021	1.028/0.007	1.190/0.033	1.056/0.020	0.970/0.027	1.077/**0.022**
DS 1D	1.119/0.018	1.064/0.008	1.149/0.034	1.099/0.023	0.974/0.028	1.081/**0.022**
DS 3D + DP	1.018/0.020	0.999/0.008	1.158/0.033	1.075/0.021	0.847/0.026	**1.019**/**0.022**
DS 1D + DP	1.068/0.016	0.970/0.007	1.148/0.034	1.140/0.024	0.855/0.035	1.036/0.023
Zhou et al. (2022)[a]	-/0.025	-/0.026	-/0.055	-/0.075	-/0.063	-/0.049
Ours w/o RANSAC	1.154/0.025	1.298/0.017	1.734/0.069	1.433/0.036	2.143/0.055	1.552/0.040
Ours	1.126/0.025	0.968/0.007	1.424/0.036	1.116/0.028	1.105/0.058	1.148/0.031
Ours + DP	1.812/0.035	1.750/0.020	1.850/0.052	1.471/0.040	1.142/0.058	1.605/0.041

[a]Results quoted from Table 1 in Zhou et al. (2022).

Table 22. Initialization Window Scale Error (%) on EuRoC Machine Hall After VI-BA (5 KFs, 0.5 s Window).

	MH01	MH02	MH03	MH04	MH05	Average
DS 3D	11.922	5.729	13.562	10.865	10.042	10.424
DS 1D	8.320	41.925	12.059	9.738	10.296	16.468
DS 3D + DP	0.611	7.494	13.657	10.597	3.925	**7.257**
DS 1D	8.025	7.616	12.646	1.648	15.847	9.156
Ours w/o RANSAC	52.709	12.462	137.833	12.413	102.917	63.667
Ours	52.444	3.868	0.134	1.152	217.949	55.109
Ours + DP	50.045	25.043	42.106	12.113	217.945	69.451

Table 23. Visual-Inertial Odometry Tracking ATE (deg/m) on EuRoC Machine Hall (5 KFs, 0.5 s Window for init).

Algorithm	MH01	MH02	MH03	MH04	MH05	Average
DS 3d	2.294/0.610	3.675/0.438	2.426/0.202	3.523/0.967	2.820/0.928	2.948/0.629
DS 1D	2.513/1.118	3.882/0.324	2.018/0.238	3.431/0.912	2.758/0.908	2.920/0.700
DS 3D + DP	2.524/0.260	3.704/0.342	2.515/0.202	3.705/1.590	3.153/0.813	3.120/0.641
DS 1D + DP	2.137/0.199	3.735/0.357	2.067/0.233	2.570/0.722	3.202/0.409	2.742/**0.384**
Zhou et al. (2022)[a]	-/0.543	-/0.071	-/1.299	-/0.124	-/0.910	-/0.589
Ours w/o RANSAC	2.712/0.942	4.520/0.394	3.114/0.358	2.969/0.750	4.434/1.916	3.550/0.872
Ours	2.282/0.184	3.970/1.204	2.006/0.164	2.368/0.438	2.888/0.864	2.703/0.571
Ours + DP	3.669/3.984	4.074/2.693	2.312/1.605	2.798/0.632	2.911/0.851	3.153/1.953

[a]Results quoted from Table 3 in Zhou et al. (2022).

Figure 22, where it can be seen that our method is comparable to the others. In general, our method outperforms Zhou et al. (2022) on the Machine Hall sequences. Not all methods initialized all of the time in this experiment. DS 1D, Ours, and Ours + DP initialized 64 out of 65 times successfully, while the others (including our method without RANSAC) successfully initialized 65 out of 65 times.

Table 24 reports the results of reducing the number of available feature tracks. It is clear that our method with and without the extra depth prior is more robust to a low number of feature tracks available than the others.

8. Discussion and limitations

While we have shown that the proposed method has state-of-the-art initialization performance on short time windows (0.3 s and 0.5 s), we admit that its performance diminishes as the initialization time window increases and more parallax/

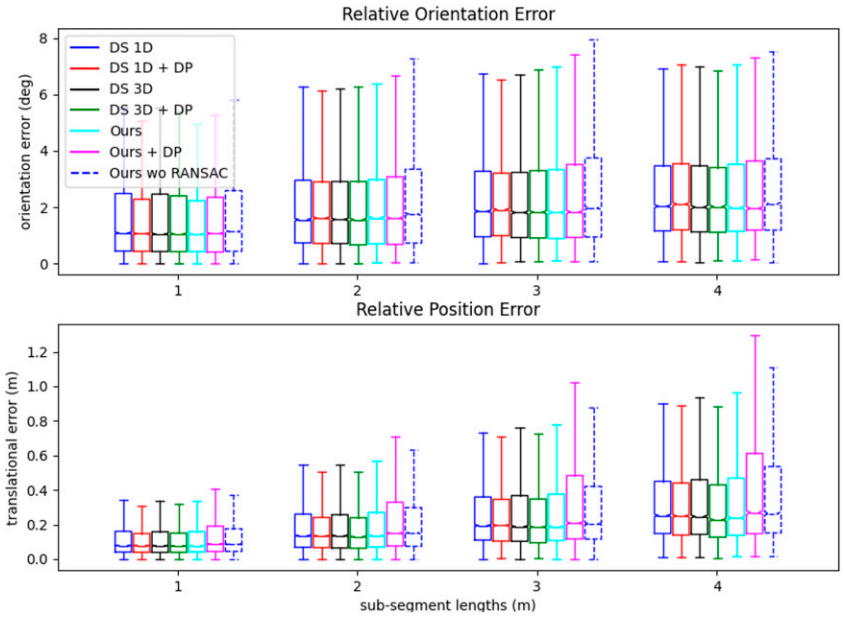

Figure 22. RPE for VIO tracking on EuRoC Machine Hall with 5 KFs and 0.5 s window.

Table 24. Percent of Successful Initializations on EuRoC Machine Hall (Averaged Over all Sequences) With 5 KFs and 0.5 s Window.

Algorithm	60 feats	45 feats	30 feats	15 feats
DS 3D	**100.0**	96.9	**100.0**	95.4
DS 1D	98.5	96.9	**100.0**	**100.0**
DS 3D + DP	**100.0**	96.9	**100.0**	95.4
DS 1D + DP	98.5	96.9	**100.0**	**100.0**
Ours w/o RANSAC	98.5	96.9	**100.0**	**100.0**
Ours	98.5	**98.5**	**100.0**	**100.0**
Ours + DP	98.5	**98.5**	**100.0**	**100.0**

excitation is available. We believe that this is due to the fact that our method relies on the learned *monocular* depth to aid in the *low excitation* cases, but as a consequence, cannot benefit from the classical triangulation that works very well when all the states are observable with sufficient baselines. If extremely fast monocular initialization is desired, then the proposed method reigns supreme, while if a longer initialization window is acceptable or stereo feature tracks are available, we would recommend to simply use a traditional method.

We additionally make no claim that the proposed method is able to initialize with *zero* excitation, since some motion and orientation change is required to recover scale. We also do not claim to improve any observability properties of the initialization problem—only that we can reduce the number of states required to be estimated, which improves the robustness to low number of feature tracks and lack of excitation while also being easily integrated into RANSAC for added robustness to outlier measurements.

9. Conclusions and future work

In this work, we have introduced a new state-of-the-art method to initialize monocular VIO extremely quickly and robustly with the help of a learned monocular depth network. As opposed to utilizing the learned depth in the VI-BA refinement step, we instead proposed to leverage it as known prior information in the fragile linear initialization stage—greatly reducing the number of parameters that need to be estimated. Not only does our method only require the depth to be predicted in one frame instead of all of them, it also conveniently allows for the entire linear initialization to be placed as a small minimal problem in a RANSAC loop—which robustifies the linear system that is already highly unstable outside of ideal conditions.

The proposed initialization method displays superior initialization accuracy and robustness in simulation and on two public benchmark datasets (EuRoC and TUM-VI) for short window initialization. Additionally, on TUM-VI our method shows an overall superior performance when initializing with only a 0.3 s window of data—which is the shortest ever reported. Adding the depth priors in the VI-BA on top of our method did not help in all cases, which shows that our method can simply be used on its own. While our method utilizes monocular depth to aid in initialization, it does not explicitly use it after to benefit the VIO performance as in Zuo et al. (2021) and Zhao et al.

(2022)—which would be an important point to improve upon in the future.

Acknowledgment

The authors would like to thank Mingyang Li (Google) and Chao Guo (Google) for their insightful discussion and comments during the course of this project.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by the University of Delaware (UD) College of Engineering, the NSF (IIS-1924897, MRI-2018905, SCH-2014264), and Google ARCore. Geneva is also partially supported by the NASA DE Space Grant Graduate Fellowship.

ORCID iDs

Nathaniel Merrill https://orcid.org/0000-0002-0579-7835
Patrick Geneva https://orcid.org/0000-0002-2179-3447
Chuchu Chen https://orcid.org/0000-0001-6903-6405

Note

1. We note that if one uses the quadratically-constrained least-squares (as we do in our system) this would remove 1DoF.

References

Agarwal S, Mierle K and Team TCS (2023) Ceres Solver. URL https://github.com/ceres-solver/ceres-solver

Bayard DS, Conway DT, Brockers R, et al. (2019) Vision-based navigation for the nasa mars helicopter. *AIAA Scitech 2019 Forum*: 1411.

Bloesch M, Burri M, Omari S, et al. (2017) Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback. *The International Journal of Robotics Research* 36(10): 1053–1072.

Burri M, Nikolic J, Gohl P, et al. (2016) The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*.

Campos C, Montiel JM and Tardós JD (2019) Fast and robust initialization for visual-inertial slam. In: *2019 International Conference on Robotics and Automation (ICRA)*. Piscataway: IEEE, 1288–1294.

Campos C, Montiel JM and Tardós JD (2020) Inertial-only optimization for visual-inertial initialization. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. Piscataway: IEEE, 51–57.

Campos C, Elvira R, Rodríguez JJG, et al. (2021) ORB-SLAM3: an accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics* 37(6): 1874–1890.

Camurri M, Ramezani M, Nobili S, et al. (2020) Pronto: a multi-sensor state estimator for legged robots in real-world scenarios. *Frontiers in Robotics and AI* 7: 68.

Chatfield AB (1997) *Fundamentals of High Accuracy Inertial Navigation*. Reston: AIAA.

Chen C, Yang Y, Geneva P, et al. (2022) *Visual-inertial-aided online mav system identification*. Piscataway: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

Concha A, Burri M, Briales J, et al. (2021) Instant visual odometry initialization for mobile ar. *IEEE Transactions on Visualization and Computer Graphics* 27(11): 4226–4235.

Dong-Si TC and Mourikis AI (2011) Closed-form solutions for vision-aided inertial navigation. In: *Technical report, Dept. of Electrical Engineering*. Riverside: University of California. URL: http://tdongsi.github.io/download/pubs/2011_VIO_Init_TR.pdf

Dong-Si TC and Mourikis AI (2012) Estimator initialization in vision-aided inertial navigation with unknown camera-imu calibration. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems)*. Piscataway: IEEE, 1064–1071.

Eckenhoff K, Geneva P and Huang G (2019) Closed-form preintegration methods for graph-based visual-inertial navigation. *The International Journal of Robotics Research* 38(5): 563–586.

Eisele J, Song Z, Nelson K, et al. (2019) Visual-inertial guidance with a plenoptic camera for autonomous underwater vehicles. *IEEE Robotics and Automation Letters* 4(3): 2777–2784.

Evangelidis G and Micusik B (2021) Revisiting visual-inertial structure-from-motion for odometry and slam initialization. *IEEE Robotics and Automation Letters* 6(2): 1415–1422.

Forster C, Carlone L, Dellaert F, et al. (2015) Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation Robotics: Science and Systems XI, Daegu, Republic of Korea, July 10 - July 14, 2023.

Geneva P and Huang G (2022) *Openvins State Initialization: Details and Derivations*. Newark: University of Delaware. Available: https://pgeneva.com/downloads/reports/tr_init.pdf

Geneva P, Eckenhoff K and Huang G (2019) A linear-complexity EKF for visual-inertial navigation with loop closures Proc. International Conference on Robotics and Automation, Montreal, Canada, 25-25 April 1997.

Geneva P, Eckenhoff K, Lee W, et al. (2020) OpenVINS: a research platform for visual-inertial estimation. Proc. Of the IEEE International Conference on Robotics and Automation, Paris, France, 25-25 April 1997. https://github.com/rpng/open_vins.

Hesch JA, Kottas DG, Bowman SL, et al. (2013) Consistency analysis and improvement of vision-aided inertial navigation. *IEEE Transactions on Robotics* 30(1): 158–176.

Hesch JA, Kottas DG, Bowman SL, et al. (2014) Camera-imu-based localization: observability analysis and consistency improvement. *The International Journal of Robotics Research* 33(1): 182–201.

Hruby P, Duff T, Leykin A, et al. (2022) Learning to solve hard minimal problems. In: *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition. . Piscataway: IEEE, 5532–5542.

Huang G (2019) Visual-inertial navigation: a concise review, *Proc. International Conference on Robotics and Automation*, Montreal, Canada, 25-25 April 1997.

Kaiser J, Martinelli A, Fontana F, et al. (2016) Simultaneous state initialization and gyroscope bias calibration in visual inertial aided navigation. *IEEE Robotics and Automation Letters* 2(1): 18–25.

Leutenegger S, Lynen S, Bosse M, et al. (2015) Keyframe-based visual-inertial odometry using nonlinear optimization. *The International Journal of Robotics Research* 34(3): 314–334.

Li M and Mourikis AI (2013) High-precision, consistent ekf-based visual-inertial odometry. *The International Journal of Robotics Research* 32(6): 690–711.

Li M and Mourikis AI (2014) A convex formulation for motion estimation using visual and inertial sensorsIn: Proceedings of the Workshop on Multi-View Geometry, Held in Conjunction with RSS, Berkeley, CA, July, 2014.

Liu S, Nie X and Hamid R (2022) Depth-guided sparse structure-from-motion for movies and tv shows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle WA, USA, Jun 21st, 2024, 15980–15989.

Lupton T and Sukkarieh S (2012) Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions. *IEEE Transactions on Robotics* 28(1): 61–76.

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

Martinelli A (2014) Closed-form solution of visual-inertial structure from motion. *International Journal of Computer Vision* 106(2): 138–152.

Merrill N, Geneva P, Katragadda S, et al. (2023) Fast monocular visual-inertial initialization leveraging learned single-view depth. In: Proc. Robotics: Science and Systems (RSS), Delft, Netherlands, Jul 15 – Jul 19, 2024.

Mourikis AI and Roumeliotis SI (2007) A multi-state constraint Kalman filter for vision-aided inertial navigation. In: Proceedings of the IEEE International Conference on Robotics and Automation. Rome, Italy, 13 May - 17 May 2024, 3565–3572.

Mur-Artal R and Tardós JD (2017a) ORB-SLAM2: an open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics* 33(5): 1255–1262.

Mur-Artal R and Tardós JD (2017b) Visual-inertial monocular slam with map reuse. *IEEE Robotics and Automation Letters* 2(2): 796–803.

Nistér D (2004) An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(6): 756–770.

Özaslan T, Loianno G, Keller J, et al. (2017) Autonomous navigation and mapping for inspection of penstocks and tunnels with mavs. *IEEE Robotics and Automation Letters* 2(3): 1740–1747.

Qin T and Shen S (2017) Robust initialization of monocular visual-inertial estimation on aerial robots. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Piscataway: IEEE, 4225–4232.

Qin T, Li P and Shen S (2018) VINS-Mono: a robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics* 34(4): 1004–1020.

Ranftl R, Lasinger K, Hafner D, et al. (2022) Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(3): 1623–1637.

Schubert D, Goll T, Demmel N, et al. (2018) The tum vi benchmark for evaluating visual-inertial odometry. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Piscataway: IEEE, 1680–1687.

Trawny N and Roumeliotis SI (2005) Indirect Kalman filter for 3D attitude estimation. In: *Technical Report*. Minnesota, USA: University of Minnesota, Dept. of Comp. Sci. & Eng.

Usenko V, Demmel N, Schubert D, et al. (2019) Visual-inertial mapping with non-linear factor recovery. *IEEE Robotics and Automation Letters* 5(2): 422–429.

Wu KJ, Guo CX, Georgiou G, et al. (2017) VINS on wheels. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. Piscataway: IEEE, 5155–5162.

Yang L, Kang B, Huang Z, et al. (2024) Depth anything: unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 13-19 June 2020.

Zhang Z and Scaramuzza D (2018) A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Piscataway: IEEE, 7244–7251.

Zhang Z, Gallego G and Scaramuzza D (2018) On the comparison of gauge freedom handling in optimization-based visual-inertial state estimation. *IEEE Robotics and Automation Letters* 3(3): 2710–2717.

Zhao M, Zhou D, Song X, et al. (2022) Dit-slam: real-time dense visual-inertial slam with implicit depth representation and tightly-coupled graph optimization. *Sensors* 22(9): 3389.

Zhou Y, Kar A, Turner E, et al. (2022) Learned monocular depth priors in visual-inertial initialization. In: *European Conference on Computer Vision*. Berlin: Springer Science+Business Media.

Zuñiga-Noël D, Moreno FA and Gonzalez-Jimenez J (2021) An analytical solution to the imu initialization problem for visual-inertial systems. *IEEE Robotics and Automation Letters* 6(3): 6116–6122.

Zuo X, Merrill N, Li W, et al. (2021) Codevio: visual-inertial odometry with learned optimizable dense depth. In: *Proc. of the IEEE International Conference on Robotics and Automation*. China: Xi'an.