Initialization of Monocular Visual Navigation for Autonomous Agents Using Modified Structure from Small Motion

Juan-Diego Florez[1], Mehregan Dor[1], Panagiotis Tsiotras[1]

***Abstract*— We propose a standalone monocular visual Simultaneous Localization and Mapping (vSLAM) initialization pipeline for autonomous space robots. Our method, a state-of-the-art factor graph optimization pipeline, extends Structure from Small Motion (SfSM) to robustly initialize a monocular agent in spacecraft inspection trajectories, addressing visual estimation challenges such as weak-perspective projection and center-pointing motion, which exacerbates the bas-relief ambiguity, dominant planar geometry, which causes motion estimation degeneracies in classical Structure from Motion, and dynamic illumination conditions, which reduce the survivability of visual information. We validate our approach on realistic, simulated satellite inspection image sequences with a tumbling spacecraft and demonstrate the method's effectiveness over existing monocular initialization procedures.**

I. INTRODUCTION

A. Problem Statement

Accurate estimation of the relative pose and 3D map of a non-cooperative resident space object (RSO) enables the real-time guidance and control required for missions such as satellite repair and active debris removal and is crucial for safe inspection and proximity operations [1]–[5]. This work addresses the initialization of a relative navigation pipeline on an autonomous chaser spacecraft tracking a non-cooperative RSO, without prior knowledge of the RSO's kinematics, dynamics, or 3D structure.

Monocular visual Simultaneous Localization and Mapping (vSLAM) systems provide real-time estimates, remain robust in dynamic environments, and operate with low power consumption and mass. Unlike other approaches, they do not require ranging (LIDAR), time-of-flight detection or structured light projection (RGB-D cameras), or specialized calibration (stereo cameras). Although monocular vSLAM systems are prone to scale and depth ambiguities, proper initialization and use of advanced estimation algorithms enable accurate relative motion, 3D mapping, and dynamic motion characterization [6], [7].

Initializing RSO inspection trajectories is challenging due to the weak-perspective projection caused by large operating ranges [8], which limits depth variation and exacerbates the bas-relief ambiguity. Under these conditions, small rotations, small translations, and depth scaling yield similar 2D projections, thereby complicating 3D mapping. The ambiguity is further pronounced in center-pointing motions, where reduced apparent lateral motion further complicates the disambiguation of motion components.

Both feature-based and pixel-based detection approaches face additional challenges. For instance, dynamic illumination conditions in space limit the survivability of tracked features and reduce the reliability of photometric error in pixel-based approaches. RSOs often have dominant planar geometries (e.g., solar panel arrays), leading to degeneracies in fundamental matrix estimation [9], and low-texture areas that reduce pixel-intensity variation and hinder pixel-based methods.

To address these challenges, we enhance the Structure from Small Motion (SfSM) [10], [11] framework to develop a monocular vSLAM initialization module that is robust to the ambiguities and degeneracies of inspection trajectories, enabling rapid convergence to accurate relative poses estimates and a high quality 3D mapping solution.

B. Related Work

Sparse, feature-based visual estimation methods are well-suited to online space-bound applications due to their computational efficiency over dense methods and their robustness to dynamic lighting conditions. Assuming a known calibrated camera intrinsic matrix, the classical feature-based initialization, often found in Structure from Motion (SfM) schemes, involves estimating motion from two-view geometry [12]. The 5-point algorithm [13], commonly used in SLAM initialization, estimates the essential matrix for relative pose estimation and is often paired with Random Sample Consensus (RANSAC) for outlier rejection. The essential matrix is decomposed into four possible relative pose solutions, refined through 3D point triangulation and cheirality checks. However, dominant planes in the scene can cause degeneracies in fundamental and essential matrix estimation [9], resulting in failed pose recovery and mapping errors. In such cases, homography-based methods [8] are preferred.

Model selection strategies can enhance the robustness of visual estimation pipelines. For example, ORB-SLAM [14] solves for both fundamental matrix and homography transformation during initialization, using an error metric to select the best-fitting model. Similarly, the USAC_FM_8PTS algorithm [15], [16] applies DEGENSAC [9], [17] to detect planar degeneracies and select the most appropriate model, and uses LOSAC [18] to refine the model through local optimization. However, defining a unified metric to compare different models can yield inconsistent results [19], as higher

This work has been supported by Verus Research under AFRL contract No. FA9453- 23-C-A025 and by NSF award FRR-2101250.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Air Force Research Laboratory.

[1]School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA {jdflorez, mehregan.dor, tsiotras}@gatech.edu

degree-of-freedom (DoF) models may produce lower errors than simpler models without necessarily providing accurate motion recovery, particularly in noisy conditions.

Model-based methods require sufficient parallax between queried frames to generate accurate estimates. Consequently, pipelines using this approach delay initialization until enough parallax is present to resolve depth, necessitating a wide baseline for estimation; hence, they are known as "delayed initialization" and "wide-baseline" methods. However, at operating ranges with weak-perspective projection, achieving sufficient parallax requires significant relative motion between successive camera frames. Accordingly, delayed initialization faces a timing dilemma—delaying too long risks losing feature tracks due to changing illumination conditions over a large baseline motion, while initializing early without sufficient parallax leads to unreliable estimates. Thus, wide-baseline approaches are not amenable to small-motion image sequences [20].

Sensor-fusion approaches, like visual-inertial SLAM [21], [22], may mitigate some limitations of purely visual methods. However, inertial measurements are ineffective for estimating motion in free-fall—typical of un-powered inspection orbits [7]—and therefore do not provide useful cues for pose initialization in RSO inspection trajectories.

Structure from Small Motion (SfSM) [10] offers a non-delayed, purely visual approach that leverages small apparent motion between consecutive frames. Rather than waiting for large baselines to accumulate, SfSM utilizes the available visual information early, thereby reducing the risk of losing track of critical features to changing illumination conditions and motion. Thus, SfSM is especially effective for early initialization, where the robust feature tracking is crucial. SfSM is well-suited to handle small-motion scenarios, avoiding the dependence on wide-baseline parallax and enabling faster, more reliable initialization in spacecraft inspection arcs.

To improve the robustness of SfSM under the small-motion, small-angle rotation assumption, Ha et al. [11] propose a three-step estimation pipeline: first, camera rotations are recovered using RANSAC; second, camera translations and 3D point inverse depths are estimated through a restricted bundle adjustment (BA); and finally, a full BA is performed. Hence, the required initialization to the full BA optimization is recovered through the first two steps, enabling reliable convergence in favorable image sequences. Our work extends the three-step SfSM framework to robustly initialize under the challenging visual estimation conditions of RSO inspection trajectories.

C. Contributions

In this work, we develop a three-step SfSM monocular initialization pipeline that modifies the small-motion assumption to achieve a robust and accurate initial map and relative trajectory solution for space vSLAM applications. We extend the approach in [11] to ensure robust initialization in RSO inspection trajectories, which are challenged by weak-perspective projection, center-pointing motion, dominant planar geometry, and dynamic illumination. We validate the performance of our initialization method using realistic simulated image sets.

Our specific contributions to the SfSM procedure are as follows:

- We redefine the small-motion, small-angle rotation assumptions of previous SfSM pipelines to address weak-perspective projection and center-pointing motion, where the apparent motion of small translations is comparable to that of small rotations.
- We re-parameterize inverse depth to guarantee the cheirality condition and ensure numerical stability.
- We reformulate 3D landmarks to account for image feature quantization by allowing in-camera-plane variation of their coordinates.

II. Methodology

The proposed initialization pipeline takes as input a SLAM front-end output consisting of feature-tracks that contain the pixel coordinates of m point features tracked across n frames, and processes them in a three-step pipeline comprising:

1) Rotation and scaled translation estimation
2) Translation and inverse depth estimation
3) Bundle adjustment

A. Notation

For convenience, we set the 0-th camera frame as the reference frame and align it with the world frame. The camera motion from the reference frame to the i-th frame is described by a rigid transformation consisting of a rotation matrix $\mathbf{R}_i \in \mathrm{SO}(3)$ and a translation vector $\mathbf{r}_i \in \mathbb{R}^3$. The coordinates of the j-th 3D landmark expressed in the i-th camera frame are denoted as $\mathbf{y}_{ij}$, and follow the relation

$$\mathbf{y}_{ij} = \mathbf{R}_i \mathbf{y}_{0j} + \mathbf{r}_i, \tag{1}$$

where $\mathbf{y}_{0j} \triangleq [X_j, Y_j, Z_j]^\top$ is the 3D coordinate vector of the j-th landmark expressed in the reference frame (index 0).

For each j-th point and i-th image frame, we define the homogeneous pixel coordinates $\mathbf{p}_{ij} = [u_{ij}, v_{ij}, 1]^\top$. We transform these coordinates into the camera coordinate frame using $\mathbf{x}_{ij} = \mathbf{K}^{-1}\mathbf{p}_{ij}$, where $\mathbf{K}$ is the pre-calibrated intrinsic matrix and $\mathbf{x}_{ij} = [x_{ij}, y_{ij}, 1]^\top$. The projected and normalized coordinate vector $\mathbf{x}_{ij}$ of the 3D landmark $\mathbf{y}_{ij}$ is given by $\mathbf{x}_{ij} = \langle \mathbf{y}_{ij} \rangle$, where $\langle \cdot \rangle$ denotes the normalization by the third component, such that $\langle [x, y, z]^\top \rangle = [x/z, y/z, 1]^\top$ for $z \neq 0$. Measured quantities are denoted using the hat operator $\hat{(\cdot)}$.

B. Step 1: Rotation and Scaled Translation Estimation

Given m sequences $(\hat{\mathbf{p}}_{0j}, \hat{\mathbf{p}}_{1j}, \ldots, \hat{\mathbf{p}}_{nj})$, $j = 1, \ldots, m$ of image point measurements $\hat{\mathbf{p}}_{ij}$ matched and tracked across image frames $i = 1, \ldots, n$, we apply the RANSAC algorithm to obtain the best-fitting rotation estimate and the corresponding inlier set between the reference frame and each of the frames $i = 1, \ldots, n$.

As in [11], we initially assume that the 3D coordinate of each landmark lies at an unknown depth along the back-projection of its corresponding image point. Consequently,

the j-th landmark, with expected coordinate vector $\mathbf{y}_{0j}$, is parameterized by the camera frame coordinate measurement $\hat{\mathbf{x}}_{0j} = \mathbf{K}^{-1}\hat{\mathbf{p}}_{0j}$ and an estimated inverse depth w_j, such that $\mathbf{y}_{0j} = \hat{\mathbf{x}}_{0j}/w_j$. Thus, the inverse depth w_j is the sole DoF determining the landmark's position. Using this inverse depth parameterization, we rewrite Eq. (1), yielding

$$\mathbf{y}_{ij} = \mathbf{R}_i \frac{\hat{\mathbf{x}}_{0j}}{w_j} + \mathbf{r}_i. \tag{2}$$

Under the small motion assumption, we apply the first-order approximation $\mathbf{R}_i \approx I_3 + [\boldsymbol{\theta}_i]_\times$ [10], where I_3 is the 3×3 identity matrix, $[\cdot]_\times$ denotes the skew-symmetric matrix of a 3D vector, and $\boldsymbol{\theta}_i = [\theta_{i1}, \theta_{i2}, \theta_{i3}]^\top$ and represents the rotation vector of the i-th camera frame. Hence,

$$\mathbf{R}_i = \begin{bmatrix} 1 & -\theta_{i3} & \theta_{i2} \\ \theta_{i3} & 1 & -\theta_{i1} \\ -\theta_{i2} & \theta_{i1} & 1 \end{bmatrix}. \tag{3}$$

From Eq. (2), we obtain the expected normalized camera frame coordinates $\mathbf{x}_{ij} = [x_{ij}, y_{ij}, 1]^\top$, which correspond to the 3D point $\mathbf{y}_{ij}$. Thus, $\mathbf{x}_{ij} = \langle \mathbf{y}_{ij} \rangle$, or, in scalar form,

$$\begin{aligned} x_{ij} &= \frac{\hat{x}_{0j} - \theta_{i3}\hat{y}_{0j} + \theta_{i2} + w_j r_{i1}}{-\theta_{i2}\hat{x}_{0j} + \theta_{i1}\hat{y}_{0j} + 1 + w_j r_{i3}}, \\ y_{ij} &= \frac{\theta_{i3}\hat{x}_{0j} + \hat{y}_{0j} - \theta_{i1} + w_j r_{i2}}{-\theta_{i2}\hat{x}_{0j} + \theta_{i1}\hat{y}_{0j} + 1 + w_j r_{i3}}. \end{aligned} \tag{4}$$

By virtue of normalization and essential to the derivation of Eq. (4), we have $\langle \mathbf{R}_i \frac{\hat{\mathbf{x}}_{0j}}{w_j} + \mathbf{r}_i \rangle = \langle \mathbf{R}_i \hat{\mathbf{x}}_{0j} + w_j \mathbf{r}_i \rangle$.

Applying the small-translation assumption from [11]—where small rotations dominate apparent motion and scene points are distant—we assume $w_j \mathbf{r}_i \approx \mathbf{0}$, thereby simplifying Eq. (4). However, the center-pointing motion of RSO inspection trajectories implies that both translation and rotation can contribute similarly to the apparent motion of points in the image. Consequently, $\mathbf{R}_i \hat{\mathbf{x}}_{0j} \sim w_j \mathbf{r}_i$, invalidating the small-translation assumption.

Accurately quantifying the rotation and translation contributions to the apparent motion is crucial for overcoming the bas-relief ambiguity. Under weak-perspective projection, landmark depth variations are small compared to their average distance from the camera, allowing us to approximate the individual inverse depths w_j with their mean value $\bar{w}$, such that $w_j \approx \bar{w}$ for each landmark j.

We further assume that, from the reference frame's perspective, landmarks are tightly clustered at a large distance along the camera's boresight and are re-parameterized as

$$\langle \mathbf{y}_{ij} \rangle = \langle \mathbf{R}_i \hat{\mathbf{x}}_{0j} + \bar{\mathbf{r}}_i \rangle, \tag{5}$$

where $\bar{\mathbf{r}}_i \triangleq [\bar{r}_{i1}, \bar{r}_{i2}, \bar{r}_{i3}] = \bar{w}\mathbf{r}_i$ represents the scaled translation. The uniform depth assumption serves as an effective initialization strategy. We mitigate the degeneracy in rotation estimation around the optical axis introduced by the assumption by RANSAC for robust estimation and explicit depth refinement in subsequent steps. Using this re-parameterization, the expected projected landmark coordinate $\mathbf{x}_{ij}$ has components:

$$\begin{aligned} x_{ij} &= \frac{\hat{x}_{0j} - \theta_{i3}\hat{y}_{0j} + \theta_{i2} + \bar{r}_{i1}}{-\theta_{i2}\hat{x}_{0j} + \theta_{i1}\hat{y}_{0j} + 1 + \bar{r}_{i3}}, \\ y_{ij} &= \frac{\theta_{i3}\hat{x}_{0j} + \hat{y}_{0j} - \theta_{i1} + \bar{r}_{i2}}{-\theta_{i2}\hat{x}_{0j} + \theta_{i1}\hat{y}_{0j} + 1 + \bar{r}_{i3}}. \end{aligned} \tag{6}$$

Following a RANSAC iteration and given a predetermined pixel distance threshold μ, we select the pair of rotation and scaled translation vectors $(\boldsymbol{\theta}_i^{\mathrm{s}}, \bar{\mathbf{r}}_i^{\mathrm{s}})$ from the sample set that best maximizes the cardinality of the inlier set $\mathcal{M}_i \subset \{1, \ldots, m\}$, so that $\|\hat{\mathbf{p}}_{ij} - \mathbf{K}\mathbf{x}_{ij}\| < \mu$ for all $j \in \mathcal{M}_i$. Each step 1 optimizer pair $(\boldsymbol{\theta}_i^*, \bar{\mathbf{r}}_i^*)$ minimizes the associated cost $\sum_{j \in \mathcal{M}_i} \|\hat{\mathbf{x}}_{ij} - \mathbf{x}_{ij}\|^2$, implying that $\hat{\mathbf{x}}_{ij} - \mathbf{x}_{ij} = \mathbf{0}$ for all $j \in \mathcal{M}_i$ by first-order optimality conditions. Substituting the expected point $\mathbf{x}_{ij}$ using Eq. (6), we obtain a system of equations linear in $\boldsymbol{\theta}_i$ and $\bar{\mathbf{r}}_i$, yielding

$$A_{ij} \begin{bmatrix} \theta_{i1} & \theta_{i2} & \theta_{i3} & \bar{r}_{i1} & \bar{r}_{i2} & \bar{r}_{i3} \end{bmatrix}^\top = \begin{bmatrix} \hat{x}_{0j} - \hat{x}_{ij} \\ \hat{y}_{0j} - \hat{y}_{ij} \end{bmatrix}, \tag{7}$$

where

$$A_{ij} \triangleq \begin{bmatrix} \hat{x}_{ij}\hat{y}_{0j} & -\hat{x}_{ij}\hat{x}_{0j} - 1 & \hat{y}_{0j} & -1 & 0 & \hat{x}_{ij} \\ \hat{x}_{ij}\hat{y}_{0j} + 1 & -\hat{y}_{ij}\hat{x}_{0j} & -\hat{x}_{0j} & 0 & -1 & \hat{y}_{ij} \end{bmatrix}.$$

The system in Eq. (7) can be solved in a least-squares sense using standard techniques such as matrix decomposition.

To achieve robust motion estimation, we perform RANSAC between the reference frame and each subsequent frame in the sequence to account for outliers. For each frame pair, we solve Eq. (7) to estimate the relative rotations $\boldsymbol{\theta}_i$ and scaled translations $\bar{\mathbf{r}}_i$. Our RANSAC procedure uses three points for robust triangulation with an inlier-count selection criterion, achieving 99.9% confidence in $52(n-1)$ iterations.

C. Step 2: Translation and Inverse Depth Estimation

We proceed by fixing the rotations $\boldsymbol{\theta}_i^*$, estimated in section II-B, and initializing the translations as $\mathbf{r}_i^{(0)} \leftarrow \bar{\mathbf{r}}_i^*/\bar{w}$ and the inverse depths as $w_j^{(0)} \leftarrow \bar{w}$. These initial estimates are then refined by a BA procedure restricted to estimating only translations and inverse depths. We utilize the Georgia Tech Smoothing And Mapping Library (GTSAM) [23] to implement and solve the restricted BA via factor graph optimization and a Levenberg Marquardt (LM) solver. A custom factor is defined to enable simultaneous optimization of both $\mathbf{r}_i$ and w_j, encoding the image disparity residual.

To ensure all landmarks satisfy the cheirality condition, we reparameterize w_j using the soft-plus function $\mathtt{sp}(x) = \ln(1 + \exp(\alpha x))/\alpha$, such that $w_j = \mathtt{sp}(\omega_j)$, where $\omega_j \in \mathbb{R}$ is now the landmark-related optimization variable, initialized with $\omega_j^{(0)} \leftarrow \mathtt{sp}^{-1}(w_j^{(0)})$. We enforce numerical stability by applying the equation [24]

$$\mathtt{sp}(x) = \max\{0, x\} + \frac{\mathtt{log1p}(\exp(-|\alpha x|))}{\alpha}, \tag{8}$$

where $\mathtt{log1p}(x)$ is a numerically stable implementation of $\ln(1 + x)$, particularly for $x \to 0^+$.

We incorporate Eq. (8) into a custom factor $\boldsymbol{\varepsilon}_{ij}^{\mathrm{step2}}$ that encodes, as residual, the 2D disparity between the measured

image point $\hat{\mathbf{p}}_{ij}$ and the expected projection of the corresponding landmark. The residual is modeled as a function of the parameters $\mathbf{r}_i$ and ω_j, while keeping $\boldsymbol{\theta}_i^*$ fixed, as in

$$\boldsymbol{\varepsilon}_{ij}^{\text{step2}}(\mathbf{r_i}, \omega_j) = \hat{\mathbf{p}}_{ij} - \langle \mathbf{K} \left((I_3 + [\boldsymbol{\theta}_i^*]_\times) \, \hat{\mathbf{x}}_{0j} + \mathtt{sp}(\omega_j) \mathbf{r_i} \right) \rangle \,. \quad (9)$$

We apply Eq. (9) to construct the restricted BA problem for poses $i = 1, \ldots, n$ and landmarks $j = 1, \ldots, m$, as in

$$\min_{\{\mathbf{r}_i\}_{i=1}^n, \{\omega_j\}_{j=1}^m} \sum_{i=1}^{n} \sum_{j=1}^{m} \Omega(\|\boldsymbol{\varepsilon}_{ij}^{\text{step2}}(\mathbf{r}_i, \omega_j)\|_{\Sigma_{ij}}^2), \quad (10)$$

where $\Omega(x)$ is the robust Huber cost, $\|\boldsymbol{\varepsilon}\|_\Sigma = \sqrt{\boldsymbol{\varepsilon}^\top \Sigma^{-1} \boldsymbol{\varepsilon}}$, and Σ_{ij} is the covariance of the measurement j taken at pose i. We solve Eq. (10) using the LM iterative procedure.

The optimized ω_j^* values are converted to w_j^* using Eq. (8), ensuring positive inverse depths. The minimizers $\left(\{\mathbf{r}_i^*\}_{i=1}^n, \{w_j^*\}_{j=1}^m\right)$ resulting from Eq. (10) are used to initialize the procedure in Section II-D. The procedures in Sections II-B and II-C collectively avoid the solution ambiguity associated with pose estimation from essential matrix decomposition [13].

D. Step 3: Full Bundle Adjustment

Equipped with estimates for all parameters $\boldsymbol{\theta}_i^*$, $\mathbf{r}_i^*$, and w_j^*, we solve a full BA to compute the camera trajectory and the coordinates of the landmarks.

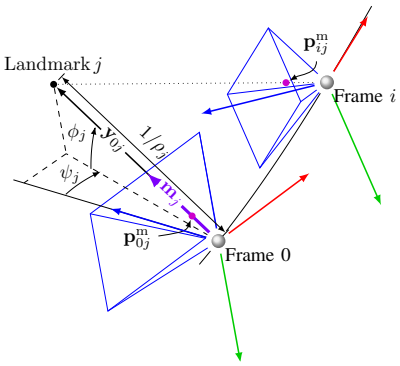


Fig. 1. Landmark parameterization using inverse depth w, azimuth ψ, and elevation ϕ.

Deviating from the landmark definition in Sections II-B and II-C, we reformulate the parameterization of the expected landmark coordinates to account for image feature quantization by introducing variables for in-camera-plane variation of the coordinates instead of directly exploiting the reference camera frame's image feature measurement. We adopt the inverse depth approach in [25], Fig. 1, re-parameterizing the 3D point $\mathbf{y}_{0j}$ as $\mathbf{y}_{0j} = \mathbf{m}_j / \rho_j$, where $\mathbf{m}_j = \mathbf{m}(\psi_j, \phi_j)$ is a unit directional vector from the reference camera frame to the j-th landmark. Here, $\mathbf{m}(\psi, \phi)$ is defined as $\mathbf{m}(\psi, \phi) = [\cos\phi \sin\psi, -\sin\phi, \cos\phi\cos\psi]^\top$, and $\rho_j = 1/\|\mathbf{y}_{0j}\|$ is the inverse depth, replacing the prior definition $w_j = 1/Z_j$. Accordingly, $\mathbf{m}_j$ is parameterized by the azimuth angle ψ_j and the elevation angle ϕ_j, which, for $\mathbf{y}_{0j} = [X_j, Y_j, Z_j]^\top$, are computed as $\psi_j = \arctan(X_j/Z_j)$ and $\phi_j = \arctan(-Y_j/\sqrt{X_j^2 + Z_j^2})$. As in Section II-C, we apply Eq. (8) in $\rho_j = \mathtt{sp}(\omega_j)$ to ensure ρ_j remains strictly positive, guaranteeing the cheirality condition.

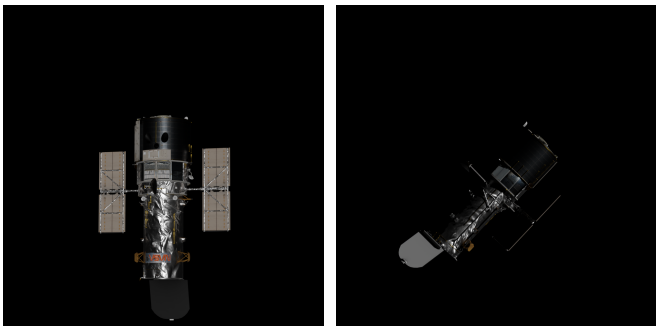


Fig. 2. Realistic synthetic images of the Hubble Space Telescope exhibiting specular and diffuse reflections and moving shadows.

We define a custom factor $\boldsymbol{\varepsilon}_{ij}^{\text{step3}}$ that encodes the residual described in Section II-C; however, the residual is now modeled as a function of the camera's rotation $\mathbf{R} \in \mathrm{SO}(3)$, the camera's translation $\mathbf{r} \in \mathbb{R}^3$, and the landmark parameters $\omega, \psi, \phi \in \mathbb{R}$, such that

$$\boldsymbol{\varepsilon}_{ij}^{\text{step3}}(\mathbf{R}, \mathbf{r}, \omega, \psi, \phi) = \hat{\mathbf{p}}_{ij} - \langle \mathbf{K} \left(\mathbf{R} \mathbf{m}(\psi, \phi) + \mathtt{sp}(\omega) \mathbf{r} \right) \rangle \,. \quad (11)$$

Thus, ρ_j, ψ_j, and ϕ_j determine the 3D location of landmark $\mathbf{y}_{0j}$.

Meanwhile, we maintain the reference frame at a fixed pose with $\mathbf{R}_0 = I_3$ and $\mathbf{r}_0 = \mathbf{0}$. Reformulating Eq. (11), for each reference frame measurement $\mathbf{x}_{0j}$, $j = 1, \ldots, m$, we constrain ψ_j and ϕ_j to yield an additional factor $\boldsymbol{\varepsilon}_{0j}^{\text{prior}}$. The factor encodes the residual described in Section II-C as a function of ψ_j and ϕ_j, as in

$$\boldsymbol{\varepsilon}_{0j}^{\text{prior}}(\psi_j, \phi_j) = \hat{\mathbf{p}}_{0j} - \langle \mathbf{K} \mathbf{m}(\psi_j, \phi_j) \rangle \,. \quad (12)$$

We use Eq. (11) and Eq. (12) to construct the BA problem for poses $i = 1, \ldots, n$ and landmarks $j = 1, \ldots, m$, as in

$$\min_{\substack{\{(\mathbf{R}_i, \mathbf{r}_i)\}_{i=1}^n \\ \{(\omega_j, \psi_j, \phi_j)\}_{j=1}^m}} \sum_{j=1}^{m} \Omega(\|\boldsymbol{\varepsilon}_{0j}^{\text{prior}}(\psi_j, \phi_j)\|_{\Sigma_{0j}}^2) + \sum_{i=1}^{n} \Omega(\|\boldsymbol{\varepsilon}_{ij}^{\text{step3}}(\mathbf{R}_i, \mathbf{r}_i, \omega_j, \psi_j, \phi_j)\|_{\Sigma_{ij}}^2), \quad (13)$$

where $\Omega(x)$, $\| * \|_\Sigma$ and Σ_{ij} are described in Section II-C. The minimizers $\left(\{(\mathbf{R}_i^{**}, \mathbf{r}_i^{**})\}_{i=1}^n, \{(\omega_j^{**}, \psi_j^{**}, \phi_j^{**})\}_{j=1}^m\right)$ to problem Eq. (13) serve as the vSLAM initialization solution.

III. Experiments

We evaluated our initialization pipeline on 101 synthetic small-motion image sequences generated in a custom simulation platform in Unreal Engine 5 [26]. Each sequence represents a weak-perspective, center-pointing RSO inspection trajectory and consists of 12 images captured at 10 frames per second using a simulated camera with a 14.9-degree field of view. The RSO is a tumbling, highly-detailed model of the Hubble Space Telescope (HST), positioned 100 meters from the camera. The simulated images of the tumbling HST, as

shown in Fig. 2, exhibit dynamic illumination conditions that closely mimic the challenges of inspection in the vicinity of a non-cooperative RSO.

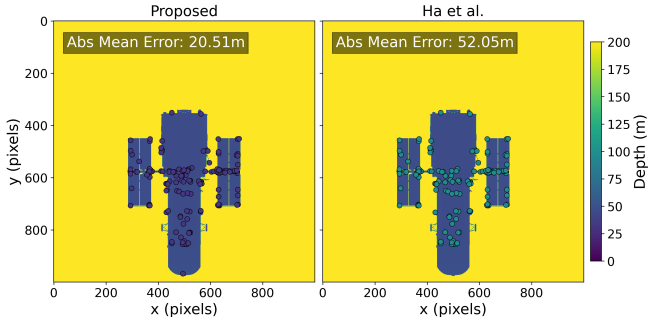


Fig. 3. The estimated landmark depths superimposed on the depth map of the HST model. The proposed method's depths (left) are estimated closer to the ground truth than the depths estimated by the method in [11].

We compare the proposed method with the SfSM pipeline developed by Ha et al. [11] and the USAC_FM_8PTS algorithm [15]. For consistency, we align the reference frame of the estimated trajectory with that of the ground truth trajectory, and we normalize the estimated translations and depths by the magnitude of the n-th frame translation.

For each sequence, we compute (a) the absolute trajectory error after aligning estimated and ground truth trajectories, given by $e_{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{x}_i^{\text{true}}\|_2$ after finding the optimal rigid alignment between trajectories; (b) the relative pose error between consecutive frames, consisting of translational error $e_{\text{RPEt}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|\mathbf{t}_{i,i+1} - \mathbf{t}_{i,i+1}^{\text{true}}\|_2$ and rotational error $e_{\text{RPEr}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \angle(R_{i,i+1}, R_{i,i+1}^{\text{true}})$; and (c) the depth error $\boldsymbol{\epsilon}_j^{\text{depth}} = Z_j - Z_j^{\text{true}}$ at the reference frame across all landmarks. The depth error is obtained by comparing the estimated depths with the scale-normalized depth map of the simulated RSO, shown in Fig. 3. To represent data trends across all sequences, we compute the mean of the root-mean-square of each error metric.

All evaluated methods were implemented in C++ using the GTSAM library and OpenCV [16]. We provide the pipelines with the same ORB [27] feature tracks from a shared front-end. Furthermore, both SfSM methods use identical LM optimizer parameters. The experiments were conducted on a machine equipped with an AMD Ryzen 7800X3D CPU and 32 GB of DDR5 RAM.

The results, summarized in Table I, demonstrate that the proposed method is significantly more robust than the other methods when initializing from realistic inspection trajectories that emulate a tumbling RSO. The proposed method successfully initialized in 77.2% of the sequences—a 47% improvement over Ha's approach, which only initialized in 52.5% of the sequences. Furthermore, in Fig. 4, the smaller height of the box plots for the proposed method compared to Ha's method indicates a lower spread in error across all SLAM variables, demonstrating more reliable and robust performance. Thus, we are able to overcome the bas-relief ambiguity more effectively and consistently when initializing in RSO inspection trajectories, as seen in Fig. 5.

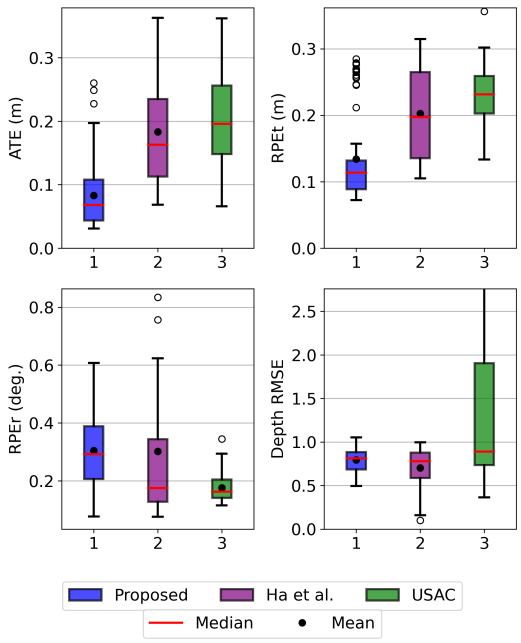
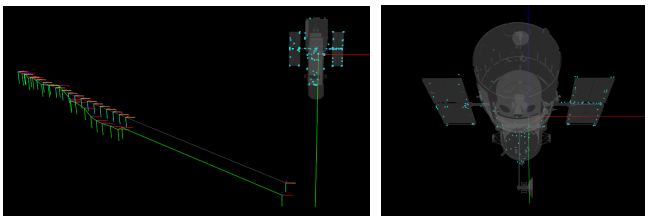


Fig. 4. Error-value distribution, means (black lines), and medians (red lines) for the error metrics comparing the proposed, Ha et al. [11], and USAC_FM_8PTS [15] methods. The image is cropped to show major data trends.



(a) Pose trajectory (b) 3D reconstruction

Fig. 5. SLAM solution one step after initialization, with estimated poses (RGB) closely aligned with ground truth poses (CYM) and estimated landmarks (cyan) coinciding with the ground truth HST object.

The USAC_FM_8PT method had the most favorable depth estimates in the 7% of the sequences in which it was able to initialize. Its success was limited to sequences featuring favorable illumination and high depth-of-field, highlighting why wide-baseline approaches, discussed in I-B, are ill-suited for initialization in small-motion RSO inspection sequences, particularly when compared to methods that explicitly account for weak-perspective projection and small apparent motion.

Although more robust and accurate, the proposed algorithm achieves its estimation improvements at the cost of an average 8.9 second increase in computation time compared to the competing SfSM method. The inverse depth and landmark re-parameterizations in step 3 enhance the robustness of the solution, but also introduce additional variables to the optimization, increasing the dimensionality of the search space. A key challenge arises from the vanishing gradient issue, which hinders convergence [28]. The problem is exacerbated by ambiguous observations and scale differences between the optimization variables. However, when considering the

TABLE I

COMPARISON OF INITIALIZATION METHODS ON 101 WEAK-PERSPECTIVE, CENTER-POINTING, SMALL-MOTION SEQUENCES

Method	Success Rate (%)	ATE (m)		RPE_t (m)		RPE_r (m)		Norm. Depth Error		Comp. Time (s)
		RMSE	Median	RMSE	Median	RMSE	Median	RMSE	Median	
Proposed	**77.2**	**0.096**	**0.067**	**0.105**	**0.105**	**0.349**	0.274	0.789	0.813	29.9
Ha et. al.	52.5	0.176	0.130	0.162	0.133	0.693	0.165	0.684	**0.758**	21.0
USAC	7.00	0.261	0.203	0.162	0.137	0.595	**0.154**	**0.646**	0.778	**0.126**

scale and range of RSO inspection trajectories, the added computation time is less impactful.

IV. CONCLUSIONS

In this work, we present a monocular initialization pipeline that extends the Ha et al. three-step SfSM method [11] to address visual estimation in RSO inspection trajectories, which are characterized by weak-perspective projection, center-pointing trajectories, dominant planar geometry, and dynamic illumination conditions. We demonstrate improved accuracy and robustness over both Ha's approach and the USAC_FM_8PTS wide-baseline method when evaluating on realistic simulated image sets.

Future work for the proposed method should focus on improving the computation time and robustness of the initializer, enabling its integration into monocular SLAM pipelines for real-time, time-critical applications. For instance, variable pre-conditioning may be used to increase the rate of convergence of step 3 of the pipeline, addressing the vanishing Jacobian issue. Other potential improvements include using simulated annealing [29] for better local minima exploration and applying convex relaxation methods [30] to improve nonlinear program initialization. Incorporating motion constraints for different RSO motion paradigms may also help further constrain the problem.

REFERENCES

[1] J. Ventura, M. Ciarcià, M. Romano, and U. Walter, "Fast and near-optimal guidance for docking to uncontrolled spacecraft," *Journal of Guidance, Control, and Dynamics*, vol. 40, no. 12, p. 3138–3154, Dec. 2017.

[2] B. Ma, Z. Jiang, Y. Liu, and Z. Xie, "Advances in space robots for on-orbit servicing: A comprehensive review," *Advanced Intelligent Systems*, vol. 5, no. 8, Apr. 2023.

[3] K. Albee, C. Oestreich, C. Specht, A. Terán Espinoza, J. Todd, I. Hokaj, R. Lampariello, and R. Linares, "A robust observation, planning, and control pipeline for autonomous rendezvous with tumbling targets," *Frontiers in Robotics and AI*, vol. 8, Sept. 2021.

[4] "ADRAS-J - Astroscale, Securing Space Sustainability — astroscale.com," 2024.

[5] G. Aglietti, B. Taylor, S. Fellowes, T. Salmon, I. Retat, A. Hall, T. Chabot, A. Pisseloup, C. Cox, Z. A., A. Mafficini, N. Vinkoff, K. Bashford, C. Bernal, F. Chaumette, A. Pollini, and W. Steyn, "The active space debris removal mission RemoveDebris. Part 2: In orbit operations," *Acta Astronautica*, vol. 168, 09 2019.

[6] M. Dor and P. Tsiotras, "ORB-SLAM applied to spacecraft non-cooperative rendezvous," in *2018 Space Flight Mechanics Meeting*. American Institute of Aeronautics and Astronautics, Jan. 2018.

[7] M. Dor, T. Driver, K. Getzandanner, and P. Tsiotras, "AstroSLAM: Autonomous monocular navigation in the vicinity of a celestial small body—theory and experiments," *The International Journal of Robotics Research*, June 2024.

[8] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[9] O. Chum, T. Werner, and J. Matas, "Two-view geometry estimation unaffected by a dominant plane," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 772–779.

[10] F. Yu and D. Gallup, "3D reconstruction from accidental motion," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2014.

[11] H. Ha, T.-H. Oh, and I. S. Kweon, "A closed-form solution to rotation estimation for structure from small motion," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 393–397, Mar. 2018.

[12] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, p. 1052–1067, June 2007.

[13] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 6, pp. 756–770, 2004.

[14] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[15] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, "USAC: A universal framework for random sample consensus," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 2022–2038, 2013.

[16] O. Team, "Open source computer vision library," https://opencv.org, n.d., version 4.x.

[17] Y. Jin, D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi, and E. Trulls, "Image matching across wide baselines: From paper to practice," *International Journal of Computer Vision*, vol. 129, no. 2, p. 517–547, Oct. 2020.

[18] K. Lebeda, J. Matas, and O. Chum, "Fixing the locally optimized ransac–full experimental evaluation," in *British machine vision conference*, vol. 2. Citeseer Princeton, NJ, USA, 2012.

[19] P. Torr, "An assessment of information criteria for motion model selection," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, ser. CVPR-97, vol. 49. IEEE Comput. Soc, 1997, p. 47–52.

[20] M. Dor, "Autonomous and robust monocular simultaneous localization and mapping-based navigation for robotic operations in space," PhD thesis, Georgia Institute of Technology, May 2024.

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

[22] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[23] F. Dellaert and G. Contributors, "borglab/gtsam," May 2022. [Online]. Available: https://github.com/borglab/gtsam)

[24] P. F. V. Wiemann, T. Kneib, and J. Hambuckers, "Using the softplus function to construct alternative link functions in generalized linear models and beyond," *Stat. Pap. (Berl)*, vol. 65, no. 5, pp. 3155–3180, July 2024.

[25] J. Civera, A. J. Davison, and J. Montiel, "Inverse depth parametrization for monocular SLAM," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 932–945, Oct. 2008.

[26] Epic Games, "Unreal engine." [Online]. Available: https://www.unrealengine.com

[27] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.

[28] J. Nocedal and S. J. Wright, "Large-scale unconstrained optimization," *Numerical optimization*, pp. 164–192, 2006.

[29] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[30] D. M. Rosen, C. DuHadway, and J. J. Leonard, "A convex relaxation for approximate global optimization in simultaneous localization and mapping," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 5822–5829.