RVO-MIS: Robust Visual Odometry for Minimally Invasive Surgery

Zhuo Wang[1] zwang570@arizona.edu
[1] *Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA*
Chiang-Heng Chien[2] Chiang-Heng_Chien@brown.edu
[2] *School of Engineering, Brown University, Providence, RI, USA*
Eungjoo Lee[1,3] eungjoolee@arizona.edu
[1] *Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA*
[3] *Department of Ophthalmology and Vision Science, University of Arizona, Tucson, AZ, USA*

Editors: Under Review for MIDL 2026

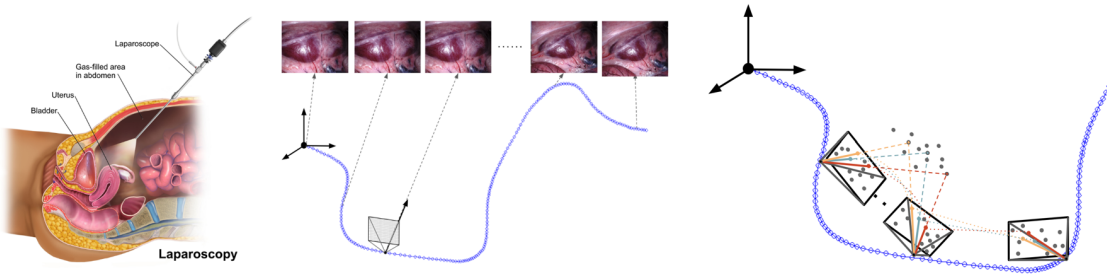
Abstract

Visual odometry (VO) in Minimally Invasive Surgery (MIS) scenarios plays a crucial role in current and future endoscopic surgical intervention assistance systems. However, MIS environments pose severely challenging situations for typical VO algorithms due to textureless environments, the movement of surgical instruments, different lighting angles, smoke generated during surgery, and organ deformation. Recent advances in this domain have increasingly incorporated deep learning-based depth estimation techniques into photometric tracking frameworks, aiming to address the inherent challenges posed by textureless regions. Yet, photometric tracking remains fragile, particularly in MIS scenes where specular reflections induce rapid and unpredictable illumination changes. In this paper, we propose a robust VO method based on feature point matching using M-Estimate Sample Consensus (MSAC) and Perspective-3-Point (P3P) absolute pose estimation to obtain accurate camera poses. To resolve the scale ambiguity, the scale of the absolute pose estimation is fixed by constructing a point cloud in the coordinate system of the first image through triangulating 3D points between keyframes. Evaluated on the SCARED dataset, our approach demonstrates consistently accurate camera pose estimation, achieving a translation ATE (RMSE) of 0.2970 cm in the best case. Quantitative results indicate that our method significantly outperforms established baseline methods in both translation and rotation metrics, validating its robustness in challenging MIS environments.

Keywords: Visual Odometry, Minimally Invasive Surgery, Feature-based Tracking

1. Introduction

Accurate camera pose estimation is an important component of navigation and guidance in Minimally Invasive Surgery (MIS). This surgical navigation system, capable of tracking the laparoscope and displaying the spatial relationship between the laparoscope and surrounding anatomical structures, can effectively reduce the risk of critical organ damage caused by excessive contact during surgery while enhancing the surgeon's spatial awareness. Compared to marker-based navigation systems, vision-based approaches demonstrate higher efficiency as they do not interrupt surgical procedures and have the potential to achieve real-time navigation (Ye et al., 2025; Liu et al., 2022). The objective is to accurately estimate the 6 degrees of freedom (DoF) camera motion from monocular video in MIS scenarios, as



(a) Laparoscopic surgery setup (b) 6-DoF motion estimation and trajectory (c) 3D triangulation and pose optimization

Figure 1: Illustration of the RVO-MIS framework in an MIS environment. (a) A typical laparoscopic surgery setup (image adapted from (Wikipedia, 2025)). (b) The visual odometry algorithm estimates the 6-DoF camera motion, generating a continuous patient-internal trajectory. (c) During keyframe updates, new 3D points are triangulated to serve as geometric anchors. The current frame's pose is then estimated via P3P and MSAC optimization based on the resulting 2D-3D correspondences.

illustrated in Figure 1. An accurate camera pose estimation is essential for downstream tasks such as three-dimensional (3D) reconstruction, Structure-from-Motion (SfM), Augmented Reality (AR), and Simultaneous Localization and Mapping (SLAM).

Most existing camera pose estimation methods are based on VO or SLAM frameworks, such as ORB-SLAM2 (Mur-Artal et al., 2015) and ElasticFusion (Whelan et al., 2016). There are also methods for monocular camera pose estimation, such as DefSLAM (Rodríguez et al., 2021), which are designed for non-rigid surgical scenes involving motions and tissue deformation. Building upon these monocular approaches, numerous deep learning-based monocular depth estimation methods have also emerged (Zhong et al., 2024; Bhat et al., 2023; Zhang et al., 2025; Yang et al., 2024) in recent years, which estimate depth images by inferring from the input RGB images. Depth estimation can be naturally embedded into traditional SfM and SLAM methods, further addressing the problem that traditional feature-based methods cannot recognize features in textureless regions. For example, Endo-Depth-and-Motion (Recasens et al., 2021), EndoSLAM (Ozyoruk et al., 2020), and LINGMI-MR (Yang et al., 2023) obtain estimated depth images through unsupervised learning methods. These estimated depth images are then used to compute camera poses via photometric tracking (e.g., based on depth-aware reprojection errors) or PoseNet architectures (Kendall et al., 2015) trained with depth-guided constraints.

Despite these advances, VO methods relying on estimated depth images inevitably suffer from errors introduced by depth estimation. Specifically, these methods involve projection- and reprojection-based photometric tracking using estimated depth images to compute relative pose (Recasens et al., 2021). Inaccurate depth estimation directly propagates errors into the camera pose estimation process. Moreover, traditional feature-based methods, while effective in outdoor and indoor environments, face significant challenges in complex MIS scenarios. In such settings, factors such as multi-angle lighting-induced reflections, surgical smoke, textureless backgrounds, and the dynamic movement of surgical instruments

create highly non-static environments, leading to insufficient or mismatched feature points that make such methods difficult to deploy effectively.

In this work, we propose a VO approach that incorporates deep learning-based methods for feature point extraction and matching with absolute pose estimation to establish metric scale in camera pose estimation. To reconcile the rigid-body assumption of P3P with non-rigid MIS deformations, we integrate MSAC to robustly filter out deforming tissues as outliers, ensuring stable tracking on quasi-rigid background structures. Recent studies have shown that such deep learning-based feature detection and matching methods demonstrate improved robustness in camera pose estimation performance compared to traditional approaches (Mackutė et al., 2024), although challenges remain in highly dynamic MIS environments. Our experimental results show that by integrating a deep learning-based feature detector, our algorithm achieves robust performance on challenging MIS sequences. Furthermore, our method outperforms depth estimation-based deep learning VO approaches by avoiding the inherent errors introduced by depth estimation. In contrast to the computationally intensive approaches presented in (Recasens et al., 2021; Ozyoruk et al., 2020; Yang et al., 2023; Hayoz et al., 2023) that require extensive GPU resources for scene-specific training of depth estimation networks, our framework employs deep learning only for feature extraction and matching. By leveraging pre-trained weights with demonstrated generalization capabilities (Mackutė et al., 2024), our method achieves robust feature detection and matching performance without requiring additional scene-specific fine-tuning.

2. Related Work

SLAM Methods: SLAM enables real-time tracking and mapping, a critical capability for MIS. SAGE (Liu et al., 2022) integrates learned priors with factor graph optimization to ensure robust reconstruction in textureless, illumination-varying environments. Another framework (Wu et al., 2022) combines medical bag-of-words with Poisson reconstruction, generating dense, detailed 3D models from sparse outputs. Addressing visual failure, ArthroSLAM (Marmol et al., 2018) utilizes a dynamically weighted Extended Kalman Filter (EKF) for continuous multi-sensor localization. Finally, feature-based methods (Deng et al., 2023) significantly improve tracking performance by combining K-means with SuperPoint (DeTone et al., 2018) for enhanced feature extraction. To address the generalization gap in deep learning-based SLAM, BodySLAM (Manni et al., 2024) achieves cross-domain generalization without fine-tuning by combining CycleGAN-based pose estimation with zero-shot depth prediction. Recently, Endo-2DTAM (Huang et al., 2025) leverages 2D Gaussian Splatting and surface normal-aware tracking to overcome multi-view inconsistencies, enabling high-fidelity, geometrically accurate reconstruction.

VO Methods: A hybrid approach (Song et al., 2021) integrating deep learning networks and geometric features was implemented in this scenario. Region classification and a two-stage pose refinement procedure are the two main components of this novel approach. It uses a Siamese network architecture (Bromley et al., 1993) and two identical PoseNet models (Kendall et al., 2015) to assess the similarity between the test image and its collected region. Pose is obtained via triangulation using region information. This data-efficient approach outperforms pure deep learning or geometry methods. Furthermore, sensor fusion is highlighted as a solution to the inherent scale drift and ambiguity of conventional monoc-

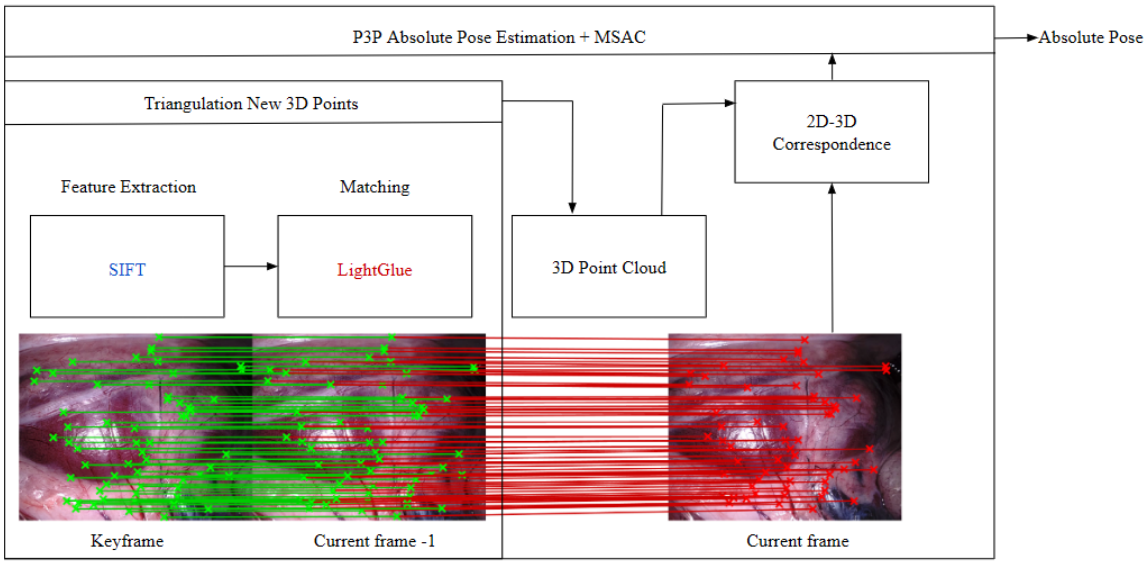


Figure 2: Overview of our proposed VO system: Our algorithm uses SIFT for feature point extraction and LightGlue for precise feature matching, followed by triangulation to reconstruct new 3D points; Absolute pose estimation is performed via the P3P algorithm with MSAC outlier rejection, using established 2D-3D correspondences between keyframe map points and current frame feature points.

ular VO. In this context, EndoVMFuseNet (Turan et al., 2017) uses a recurrent CNN to fuse 6DoF visual and 5DoF magnetic data without synchronization. Its energy reduction method integrates dense photometric alignment with sparse optical flow features. While sensor fusion enhances data richness, DPVO (Teed et al., 2023) maximizes efficiency by tracking sparse patches instead of dense flow. It combines a recurrent update operator with differentiable bundle adjustment, achieving the robustness of dense methods with significantly reduced computational and memory costs.

Depth Estimation Methods: (Chen et al., 2019) propose a cGAN-based framework using a U-Net generator for depth estimation. By enforcing geometric fidelity through adversarial training and fusing estimates within ElasticFusion (Whelan et al., 2016), it achieves robust real-time reconstruction. More recently, Yang et al. (Yang et al., 2023) introduces a geometry-aware framework based on MultiDepth (Watson et al., 2021). By employing a composite loss function targeting gradient and normal consistency, this approach significantly enhances geometric fidelity for complex anatomical features, achieving state-of-the-art performance on the EndoSLAM dataset (Ozyoruk et al., 2020).

3. Method

Overview: The core innovation of our method lies in reconstructing 3D point clouds through precise, numerous feature point detection and robust deep learning-based matching. We achieve accurate absolute pose estimation by establishing sufficient 2D-3D correspondences between keyframes and current frames, and then minimizing reprojection error using a robust PnP optimization framework (see Figure 2).

3.1. Notations

Let $\Gamma_{w,k}$ be a k-th 3D point in the world coordinate, K be the camera calibration matrix, $\mathcal{R}_i$ and $\mathcal{T}_i$ be the estimated absolute rotation matrix and translation vector of camera i, respectively. A 2D feature point γ_k with depth ρ_k gives rise to the k-th 3D point in the camera coordinate $\Gamma_{i,k} = \rho_k \gamma_k$ which relates the 3D point Γ_w in the world coordinate by

$$\Gamma_{i,k} = \mathcal{R}_i \Gamma_{w,k} + \mathcal{T}_i. \tag{1}$$

Denote $\gamma_{im,k}$ as the k-th image point in pixels so that $\gamma_{im,k} = K\gamma_k$.

3.2. Feature Extraction and Matching

We compared four feature extraction methods, *i.e.*, SIFT (Lowe, 2004), ORB (Rublee et al., 2011), SURF (Bay et al., 2006) and SuperPoint (DeTone et al., 2018). The results demonstrate that SIFT extracts a relatively larger number of accurate feature points, providing abundant candidates for subsequent feature matching. During the matching phase, we initially employed the VLfeat (Vedaldi and Fulkerson, 2010) library for SIFT points matching. The matching algorithm (Lowe, 2004) in the VLfeat library implements a combination of Lowe's ratio test and bidirectional matching. For each SIFT point in the first image, it calculates the 128-dimensional Euclidean distance to all feature points in the second image, identifying both the nearest and second-nearest neighbors. Matches are retained only when the nearest neighbor distance is smaller than the second-nearest neighbor distance divided by a predefined threshold. However, in this specific scenario, we observed that a significant proportion of SIFT points exhibited ambiguous matching characteristics, resulting in suboptimal performance of conventional descriptor-based matching methods. To address the limitations of conventional descriptor-based matching in this challenging scenario, we adopted LightGlue (Lindenberger et al., 2023), a state-of-the-art deep learning-based feature matcher. LightGlue is a feature matching framework that builds on SuperGlue (Sarlin et al., 2020). LightGlue leverages a graph neural network (GNN) to jointly reason about feature correspondences, incorporating both local appearance and geometric consistency. Compared to conventional feature-based VO methods that often fail in this challenging scenario, our approach demonstrates significantly improved robustness by incorporating advanced deep learning-based feature matching techniques.

3.3. Triangulate 3D Points

To obtain metrically-scaled continuous camera poses through absolute pose estimation, our method establishes a fixed scale reference by reconstructing 3D points from the first two frames. We estimate the relative pose between frame 1 and 2 using their 2D-2D correspondences, then reconstruct 3D points from 2D-2D correspondences that satisfy epipolar constraints. A 2D-2D correspondence is treated as inliers when their distance to the corresponding epipolar line is below 2 pixels. This reconstructed 3D point cloud is explicitly defined in frame 1's coordinate system, ensuring that all subsequent absolute pose estimations are inherently referenced to this initial frame.

3.4. Absolute Pose Estimation

Upon reconstructing 3D points from inlier 2D feature matches, we identify which 2D features in the keyframe have valid 3D correspondences. By matching 2D features between the current frame and the keyframe, we determine which observed points in the current frame correspond to these reconstructed 3D points (referred to as co-visible 3D points). To ensure robustness, we employ the MSAC to obtain the optimal absolute pose estimation. Subsequently, we refine the absolute pose estimation through energy minimization.

3.5. Energy Function

To refine the estimated absolute pose of camera i, an energy function $E(\mathcal{R}_i, \mathcal{T}_i)$ defined as a function of the absolute rotation $\mathcal{R}_i$ and absolute translation $\mathcal{T}_i$ which minimizes sum of squared *reprojection errors* is adopted, *i.e.*,

$$E(\mathcal{R}_i, \mathcal{T}_i) = \sum_{k=1}^{N} \left|\left| \gamma_{im,k} - \frac{K(\mathcal{R}_i \Gamma_{w,k} + \mathcal{T}_i)}{e_3^T K(\mathcal{R}_i \Gamma_{w,k} + \mathcal{T}_i)} \right|\right|, \tag{2}$$

so that the refined absolute pose $(\mathcal{R}_i^*, \mathcal{T}_i^*)$ is

$$(\mathcal{R}_i^*, \mathcal{T}_i^*) = \underset{\mathcal{R}_i, \mathcal{T}_i}{\operatorname{argmin}} \, E(\mathcal{R}_i, \mathcal{T}_i). \tag{3}$$

Minimizing $E(\mathcal{R}_i, \mathcal{T}_i)$ with respect to $(\mathcal{R}_i, \mathcal{T}_i)$ is done by the Levenburg-Marquardt algorithm. Note that the rotation matrix $\mathcal{R}_i$ is parameterized by the three Euler angles, and thus there are six unknowns in total, *i.e.*, three for rotation and three for translation.

3.6. Keyframe Update

Our keyframe update strategy triggers under two conditions: *(i)* when fewer than 55% of the current frame's co-visible 3D landmarks originate from the active keyframe, or *(ii)* when exceeding 15 frames since the last keyframe insertion.

3.7. Triangulate Newly Observed Point Features

Triangulating 2D point feature correspondences from two views requires the relative rotation and translation of the two cameras. This can be achieved by coordinate transformation of the current camera pose and the keyframe camera pose. Specifically, let $(\mathcal{R}_c, \mathcal{T}_c)$ and $(\mathcal{R}_k, \mathcal{T}_k)$ be the absolute poses of the current frame and the keyframe, respectively; the goal is to find the relative pose $(\mathcal{R}_{kc}, \mathcal{T}_{kc})$ so that a point Γ_c under the current camera coordinate is transformed to a point Γ_k under the keyframe camera coordinate by $\Gamma_k = \mathcal{R}_{kc}\Gamma_c + \mathcal{T}_{kc}$. From Equation 1, we have

$$\begin{cases} \Gamma_c = \mathcal{R}_c \Gamma_w + \mathcal{T}_c \\ \Gamma_k = \mathcal{R}_k \Gamma_w + \mathcal{T}_k, \end{cases} \tag{4}$$

where for simplicity we omit the index for the point. Now, Γ_w can be isolated in the first vector equation of Equation 4 as

$$\Gamma_w = \mathcal{R}_c^T \left(\Gamma_c - \mathcal{T}_c \right), \tag{5}$$

which can be plugged to the second vector equation of Equation 4, giving

$$\begin{aligned}\Gamma_k &= \mathcal{R}_k \mathcal{R}_c^T \left(\Gamma_c - \mathcal{T}_c\right) + \mathcal{T}_k \\ &= \mathcal{R}_k \mathcal{R}_c^T \Gamma_c + \mathcal{T}_k - \mathcal{R}_k \mathcal{R}_c^T \mathcal{T}_c.\end{aligned} \tag{6}$$

Thus, the relative pose $(\mathcal{R}_{kc}, \mathcal{T}_{kc})$ is

$$\begin{cases}\mathcal{R}_{kc} = \mathcal{R}_k \mathcal{R}_c^T \\ \mathcal{T}_{kc} = \mathcal{T}_k - \mathcal{R}_k \mathcal{R}_c^T \mathcal{T}_c.\end{cases} \tag{7}$$

The relative pose $(\mathcal{R}_{kc}, \mathcal{T}_{kc})$ provides epipolar constraint across the two frames as the essential matrix $E_{kc} = [\mathcal{T}_{kc}]_\times \mathcal{R}_{kc}$ can be easily found. This constraint is adopted to pick 2D-2D unconstructed point pairs that satisfy epipolar geometry from thresholding the Sampson error, *i.e.*, point-to-epipolar-line distance. These points are triangulated to form a group of *new* 3D points and are transformed to the world coordinate in order to keep the entire cloud of 3D points in the same coordinate system. This ensures continuous map expansion while maintaining metric scale consistency through geometric verification. The overall iterative procedure is summarized in Algorithm 1 in the appendix.

4. Experiments

Dataset: We evaluated our method on multiple sequences of the stereo correspondence and reconstruction of endoscopic data (SCARED) (Allan et al., 2021) dataset, including challenging scenarios with surgical instruments, smoke and reflections. The SCARED dataset consists of 9 in vivo porcine subjects, with 4 endoscopic video sequences captured for each subject using a da Vinci Xi surgical robotic system. All sequences feature rigid scenes without respiratory motion, providing stereo video streams synchronized with precise camera kinematic data. For experimental validation, we conducted comprehensive evaluations across multiple sequences.

Evaluation Metrics: We quantitatively and qualitatively evaluated our experimental results using evo (Grupp, 2017), a Python package for odometry and SLAM assessment. We adopt standard evaluation metrics for monocular visual odometry, including Absolute Trajectory Error (ATE) computed as the root-mean-square error (RMSE) between predicted trajectories and ground-truth trajectories with global alignment.

Following the existing evaluation method (Sturm et al., 2012), the estimated trajectory and the ground-truth trajectory are aligned using Horn's method, *i.e.*, a transformation containing a rotation matrix, a translation vector, and a scale so that the geometry of the estimated trajectory $\mathbf{P}_{1:n}$ can be as close to the ground-truth trajectory $\mathbf{Q}_{1:n}$ as possible. This alignment is represented by a 4×4 matrix $\mathbf{S}$. We also enforce the estimated trajectories to align to the origin, so that the first frame is identical.

Absolute trajectory error (ATE) at each timestep i is adopted as the evaluation metric which is defined as

$$\mathbf{F}_i \coloneqq \mathbf{Q}_i^{-1} \mathbf{S} \mathbf{P}_i, \tag{8}$$

which is a 4×4 matrix describing how "far apart" the estimation at timestamp i is from the ground-truth. To examine the performance on the rotation and the translation estimations,

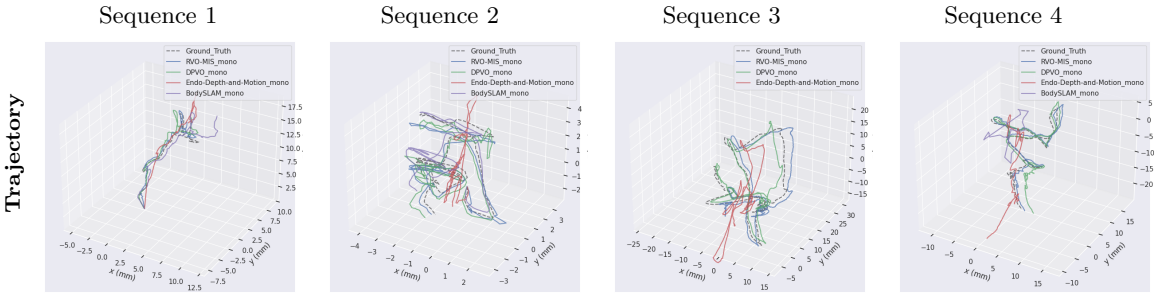


Figure 3: Trajectories comparison for the representative sequences in the SCARED dataset. Our proposed method (blue line) closely tracks the Ground Truth (black dashed line), outperforming established baselines (colored lines).

the rotation and translation parts of $\mathbf{F}_i$ are isolated as $\mathbf{F}_{i,\text{rot}}$ and $\mathbf{F}_{i,\text{trans}}$, respectively, and are represented by a single number as

$$\begin{cases} \text{ATE}(\mathcal{R}) = 2\left(3 - \text{trace}(\mathbf{F}_{i,\text{rot}})\right) \\ \text{ATE}(\mathcal{T}) = \left\|\mathbf{F}_{i,\text{trans}}\right\|, \end{cases} \tag{9}$$

where $\text{ATE}(\mathcal{R})$ describes the Frobenius norm of the rotation error, *i.e.*, $\|\mathbf{Q}_{i,\text{rot}} - \mathbf{P}_{i,\text{rot}}\|_F^2 = 2\left(3 - \text{trace}(\mathbf{F}_{i,\text{rot}})\right)$.

To aggregate errors across the entire trajectory, we define the root-mean-square error (RMSE) of translational components:

$$\text{RMSE}(\mathbf{F}_{1:n}) \coloneqq \left(\frac{1}{n}\sum_{i=1}^{n}\left\|(\mathbf{F}_{i,\text{trans}})\right\|^2\right)^{1/2}. \tag{10}$$

Results: A qualitative assessment of the trajectories is presented in Figure 3. Quantitative results are summarized in Table 1, while Figure 4 visualizes the ATE. To further analyze the error distribution, Figure 5 presents the Cumulative Distribution Function (CDF) of the translation ATE, demonstrating the superior accuracy of our method across most frames. Table 1 presents the RMSE ATE of translation and ATE of rotation for our method compared to state-of-the-art approaches across four representative sequences. For the quantitative metrics in Table 1 and the ATE visualization (Figure 4), we applied global alignment to adhere to standard error assessment protocols. In contrast, for the trajectory comparison against baseline methods (Figure 3), we utilized origin alignment. This strategy fixes a common starting pose for all methods, thereby intuitively demonstrating the reduced cumulative drift of RVO-MIS compared to the baselines.

For evaluation, we selected several state-of-the-art techniques for comparison. We compare our approach with SLAM methods (Mur-Artal et al., 2015; Rodríguez et al., 2021; Manni et al., 2024; Huang et al., 2025), VO methods (Teed et al., 2023) and deep learning based methods (Recasens et al., 2021). As demonstrated in Table 1 and Figure 3, our proposed method achieves superior performance both quantitatively and qualitatively and outperform SLAM, VO and deep learning method rely on estimated depth. The numerical results in Table 1 reveal that our approach consistently outperforms baseline methods across

Method	Sequence 1		Sequence 2		Sequence 3		Sequence 4	
	ATE ($\mathcal{T}$)	ATE ($\mathcal{R}$)	ATE ($\mathcal{T}$)	ATE ($\mathcal{R}$)	ATE ($\mathcal{T}$)	ATE ($\mathcal{R}$)	ATE ($\mathcal{T}$)	ATE ($\mathcal{R}$)
SLAM								
ORB-SLAM2 (M) (Mur-Artal et al., 2015)	0.5068	3.0534	0.9912	3.0798	9.6247	2.8640	4.4825	3.0183
ORB-SLAM2 (S) (Mur-Artal et al., 2015)	1.3091	0.3511	1.3623	3.0303	9.7053	3.0089	4.3626	2.9151
SD-DefSLAM (M) (Rodríguez et al., 2021)	1.0036	0.3247	1.7788	2.0058	6.2579	0.3455	3.5653	0.1761
BodySLAM (M) (Manni et al., 2024)	0.4504	0.1991	0.4447	0.1851	-	-	8.2481	0.7986
Endo-2DTAM (M) (Huang et al., 2025)	2.3290	2.2410	2.3187	1.1858	6.3792	1.9767	5.1394	2.3166
VO								
DPVO (M) (Teed et al., 2023)	0.3326	0.2466	0.3902	0.1763	**2.8271**	0.4607	0.9269	0.4405
Deep Learning								
EndoDepth (M) (Recasens et al., 2021)	1.1748	0.3688	1.4863	1.7176	9.2599	0.8804	4.2948	0.3484
Proposed								
RVO-MIS (M)	**0.2970**	**0.0523**	**0.3574**	**0.1302**	4.0381	**0.1261**	**0.6822**	**0.0548**

Table 1: Quantitative comparison of ATE (RMSE) on four representative sequences from the SCARED dataset. **Translation errors ($\mathcal{T}$) are reported in centimeters (cm)**, and **rotation errors ($\mathcal{R}$) are reported in degrees (°)**. Due to the inherent scale ambiguity of monocular methods, all estimated trajectories are aligned to the ground truth using Sim3 transformation (optimizing scale, rotation, and translation). Bold numbers indicate the best results, and underlined numbers indicate the second-best performances. ("M": Monocular, "S": Stereo; "-": Tracking Failure)

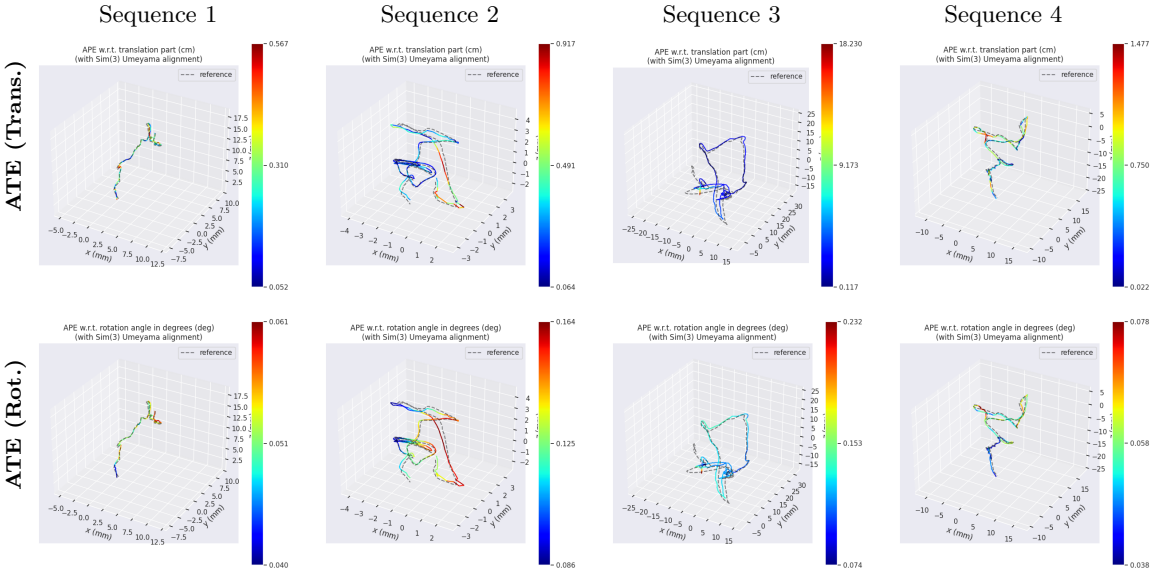


Figure 4: Visualization of ATE (top row: translation, bottom row: rotation).

key metrics—achieving the lowest ATE ($\mathcal{T}$ and $\mathcal{R}$) in 3 sequences. Notably, our algorithm maintains smooth and stable trajectories across all sequences. This consistent robustness highlights the effectiveness of the RVO-MIS framework for MIS navigation, particularly in dynamic environments where stable tracking is critical.

Run Time: The proposed method was implemented in Python to avoid cross-language interoperability overhead and leverage parallel computing capabilities. Experiments were performed on a high-performance server with dual AMD EPYC 9354 32-Core Processors (using 4 CPU cores for this study) and a single NVIDIA RTX A6000 GPU (48GB VRAM).

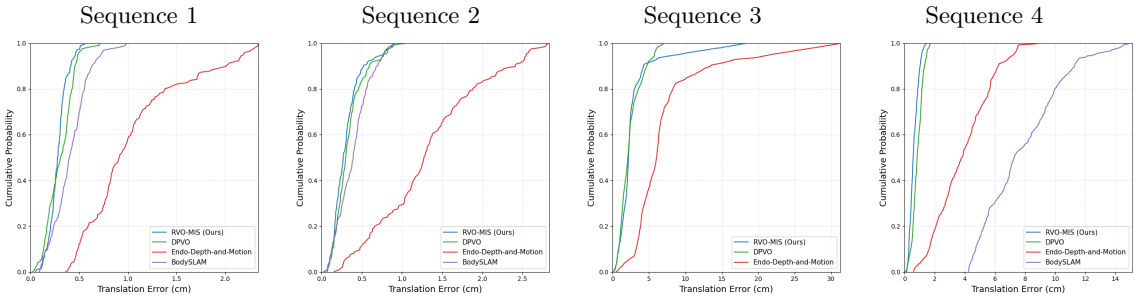


Figure 5: CDF of the translation ATE across representative sequences. The curves illustrate the cumulative distribution of per-frame errors. In Sequences 1, 2, and 4, our method (RVO-MIS, blue line) demonstrates superior accuracy. In Sequence 3, while DPVO achieves lower overall errors, our method maintains a consistent error distribution for the majority of frames, highlighting the varying challenges posed by different surgical scenes.

The system processed a total of 197 frames, with the cumulative runtime distribution dominated by feature matching and pose estimation. Specifically, feature matching accounted for 181.06s ($\approx 0.92s$ per frame), while essential matrix estimation and PnP pose recovery accounted for 73.15s ($\approx 0.37s$ per frame). In contrast, triangulation and Levenberg-Marquardt (LM) optimization remained computationally efficient, requiring only 0.54s and 0.69s in total, respectively. This distribution indicates that while the transition to a pure Python environment with multi-CPU acceleration has streamlined the pipeline, feature matching remains the primary bottleneck, prompting further research into lightweight feature matchers and optimization techniques to drastically reduce latency without compromising the system's robust accuracy.

5. Conclusions

Our method addresses the challenges of MIS environments by integrating deep learning-based feature extraction with MSAC and PnP pose estimation. This robust combination achieves state-of-the-art performance on the SCARED (Allan et al., 2021) dataset, outperforming both conventional SLAM frameworks and depth-estimation baselines. Notably, it delivers superior accuracy and requires significantly fewer computational resources than pure deep learning models, offering a precise and efficient solution for next-generation surgical navigation.

6. Future Works

Noting that DPVO (Teed et al., 2023) achieved the lowest quantitative ATE ($\mathcal{T}$) in Sequence 3, we plan to investigate the underlying causes in future work. To further refine our system, we propose three key improvements: implementing three-view feature matching for scale-aware estimation to mitigate cumulative errors; integrating advanced matchers such as LoFTR (Sun et al., 2021) to enhance PnP precision in MIS scenarios; and incorporating bundle adjustment optimization (Saha et al., 2025) to improve pose accuracy in challenging conditions.

References

Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, ECCV'06, page 404–417, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540338322. doi: 10.1007/11744023_32. URL https://doi.org/10.1007/11744023_32.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. URL https://arxiv.org/abs/2302.12288.

Jane Bromley, James Bentz, Leon Bottou, Isabelle Guyon, Yann Lecun, Cliff Moore, Eduard Sackinger, and Rookpak Shah. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7:25, 08 1993. doi: 10.1142/S0218001493000339.

Richard J. Chen, Taylor L. Bobrow, Thomas Athey, Faisal Mahmood, and Nicholas J. Durr. Slam endoscopy enhanced by adversarial depth prediction, 2019. URL https://arxiv.org/abs/1907.00283.

Liwei Deng, Zhen Liu, Tao Zhang, and Zhe Yan. Study of visual slam methods in minimally invasive surgery. *Mathematical Biosciences and Engineering*, 20(3):4388–4402, 2023. ISSN 1551-0018. doi: 10.3934/mbe.2023203. URL https://www.aimspress.com/article/doi/10.3934/mbe.2023203.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

Michael Grupp. evo: Python package for the evaluation of odometry and slam. https://github.com/MichaelGrupp/evo, 2017.

Michel Hayoz, Christopher Hahne, Mathias Gallardo, Daniel Candinas, Thomas Kurmann, Maximilian Allan, and Raphael Sznitman. Learning how to robustly estimate camera pose in endoscopic videos. *International journal of computer assisted radiology and surgery*, 18(7):1185–1192, 2023.

Yiming Huang, Beilei Cui, Long Bai, Zhen Chen, Jinlin Wu, Zhen Li, Hongbin Liu, and Hongliang Ren. Advancing dense endoscopic reconstruction with gaussian splatting-driven surface normal-aware tracking and mapping. *arXiv preprint arXiv:2501.19319*, 2025.

Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946, 2015. doi: 10.1109/ICCV.2015.336.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023.

Xingtong Liu, Zhaoshuo Li, Masaru Ishii, Gregory D. Hager, Russell H. Taylor, and Mathias Unberath. Sage: Slam with appearance and geometry prior for endoscopy. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 5587–5593, 2022. doi: 10.1109/ICRA46639.2022.9812257.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004. URL http://dblp.uni-trier.de/db/journals/ijcv/ijcv60.html#Lowe04.

Emilė Mackutė, Ahad Abdalla, Stuart Dickson, Kevin Dhaliwal, and Mohsen Khadem. On challenges of monocular pose estimation for endoluminal navigation. *Journal of Medical Robotics Research*, 09(03n04):2440009, 2024. doi: 10.1142/S2424905X24400099. URL https://doi.org/10.1142/S2424905X24400099.

G. Manni, C. Lauretti, F. Prata, R. Papalia, L. Zollo, and P. Soda. Bodyslam: A generalized monocular visual slam framework for surgical applications, 2024. URL https://arxiv.org/abs/2408.03078.

Andres Marmol, Peter Corke, and Thierry Peynot. Arthroslam: Multi-sensor robust visual localization for minimally invasive orthopedic surgery. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3882–3889, 2018. doi: 10.1109/IROS.2018.8593501.

Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, Hasan Sahin, Helder Araujo, Henrique Alexandrino, Nicholas J. Durr, Hunter B. Gilbert, and Mehmet Turan. Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos: Endo-sfmlearner, 2020.

David Recasens, José Lamarca, José M Fácil, JMM Montiel, and Javier Civera. Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robotics and Automation Letters*, 6(4):7225–7232, 2021.

Juan J Gómez Rodríguez, José Lamarca, Javier Morlana, Juan D Tardós, and José MM Montiel. Sd-defslam: Semi-direct monocular slam for deformable and intracorporeal scenes. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5170–5177, 2021.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011. doi: 10.1109/ICCV.2011.6126544.

Shreya Saha, Zekai Liang, Shan Lin, Jingpei Lu, Michael Yip, and Sainan Liu. Based: Bundle-adjusting surgical endoscopic dynamic video reconstruction using neural radiance fields, 2025. URL https://arxiv.org/abs/2309.15329.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Jingwei Song, Mitesh Patel, Andreas Girgensohn, and Chelhwon Kim. Combining deep learning with geometric features for image-based localization in the gastrointestinal tract. *Expert Systems with Applications*, 185:115631, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.115631. URL https://www.sciencedirect.com/science/article/pii/S0957417421010253.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 573–580, 2012. doi: 10.1109/IROS.2012.6385773.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021.

Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Mehmet Turan, Yasin Almalioglu, Hunter Gilbert, Alp Eren Sari, Ufuk Soylu, and Metin Sitti. Endo-vmfusenet: Deep visual-magnetic sensor fusion approach for uncalibrated, unsynchronized and asymmetric endoscopic capsule robot localization data. *CoRR*, abs/1709.06041, 2017. URL http://arxiv.org/abs/1709.06041.

Andrea Vedaldi and Brian Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1469–1472, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1874249. URL https://doi.org/10.1145/1873951.1874249.

Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel J. Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. *CoRR*, abs/2104.14540, 2021. URL https://arxiv.org/abs/2104.14540.

Thomas Whelan, Renato F Salas-Moreno, Ben Glocker, Andrew J Davison, and Stefan Leutenegger. Elasticfusion: Real-time dense slam and light source estimation. *The International Journal of Robotics Research*, 35(14):1697–1716, 2016. doi: 10.1177/0278364916669237. URL https://doi.org/10.1177/0278364916669237.

Wikipedia. Laparoscopy, 2025. URL https://en.wikipedia.org/wiki/Laparoscopy. [Online; accessed 11-March-2025].

Haibin Wu, Ruotong Xu, Kaiyang Xu, Jianbo Zhao, Yan Zhang, Aili Wang, and Yuji Iwahori. 3d texture reconstruction of abdominal cavity based on monocular vision slam for minimally invasive surgery. *Symmetry*, 14(2), 2022. ISSN 2073-8994. doi: 10.3390/sym14020185. URL https://www.mdpi.com/2073-8994/14/2/185.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2: A more capable foundation model for monocular depth estimation. *arXiv preprint arXiv:2406.09414*, 2024.

Yongming Yang, Shuwei Shao, Tao Yang, Peng Wang, Zhuo Yang, Chengdong Wu, and Hao Liu. A geometry-aware deep network for depth estimation in monocular endoscopy. *Engineering Applications of Artificial Intelligence*, 122:105989, 2023.

Ke Ye, Bai Chen, Jingyang Zhou, Jiahao Li, Haoqing Wu, Feng Ju, and Yang Wu. Enhanced visual slam for surgical robots with cylindrical scene recognition in digestive endoscopic procedures. *Measurement*, 250:117054, 2025. ISSN 0263-2241. doi: https://doi.org/10.1016/j.measurement.2025.117054. URL https://www.sciencedirect.com/science/article/pii/S0263224125004130.

Jiuling Zhang, Yurong Wu, and Huilong Jiang. Survey on monocular metric depth estimation. *Computers*, 14(11), 2025. ISSN 2073-431X. doi: 10.3390/computers14110502. URL https://www.mdpi.com/2073-431X/14/11/502.

Jiawei Zhong, Hongliang Ren, Qin Chen, and Hui Zhang. A review of deep learning-based localization, mapping and 3d reconstruction for endoscopy. *Journal of Micro and Bio Robotics*, 21, 12 2024. doi: 10.1007/s12213-024-00181-0.

Appendix A. Algorithm Details

Algorithm 1: RVO-MIS Pipeline

Input: Sequence of video frames $I_0, \ldots, I_N$, Camera Intrinsic Matrix K
Output: Camera Trajectory $\{(\mathcal{R}_i, \mathcal{T}_i)\}$, 3D Map $\mathcal{M}$

`// Initialization Phase`
Extract SIFT features for I_0 and I_1
Match features using LightGlue
Estimate relative pose and triangulate initial 3D points to form map $\mathcal{M}$
Set Keyframe $I_{kf} \leftarrow I_1$
`// Tracking Phase`
for $i \leftarrow 2$ **to** N **do**
    Extract SIFT features for current frame I_i
    Match features between I_i and I_{kf} using LightGlue
    Identify 2D-3D correspondences (co-visible points) based on $\mathcal{M}$
    `// Absolute Pose Estimation`
    Estimate initial pose $(\mathcal{R}_i, \mathcal{T}_i)$ using MSAC on 2D-3D pairs
    Refine $(\mathcal{R}_i, \mathcal{T}_i)$ by minimizing the energy Function $E(\mathcal{R}_i, \mathcal{T}_i)$ (Eq. 2) using Levenberg-Marquardt
    `// Map Management`
    $N_{cov} \leftarrow$ ratio of co-visible 3D landmarks
    $N_{gap} \leftarrow$ frame distance from I_{kf}
    if $N_{cov} < 0.55$ **or** $N_{gap} > 15$ **then**
        Calculate relative pose $(\mathcal{R}_{kc}, \mathcal{T}_{kc})$ using current and keyframe absolute poses (Eq. 7)
        Triangulate new 2D matches satisfying epipolar constraints
        Transform new points to world coordinates and add to $\mathcal{M}$
        Update Keyframe $I_{kf} \leftarrow I_i$
    end
end