ORIGINAL ARTICLE

Learning how to robustly estimate camera pose in endoscopic videos

Michel Hayoz[1] · Christopher Hahne[1] · Mathias Gallardo[1] · Daniel Candinas[2] · Thomas Kurmann[3] · Maximilian Allan[3] · Raphael Sznitman[1]

Abstract

Purpose Surgical scene understanding plays a critical role in the technology stack of tomorrow's intervention-assisting systems in endoscopic surgeries. For this, tracking the endoscope pose is a key component, but remains challenging due to illumination conditions, deforming tissues and the breathing motion of organs.

Method We propose a solution for stereo endoscopes that estimates depth and optical flow to minimize two geometric losses for camera pose estimation. Most importantly, we introduce two learned adaptive per-pixel weight mappings that balance contributions according to the input image content. To do so, we train a Deep Declarative Network to take advantage of the expressiveness of deep learning and the robustness of a novel geometric-based optimization approach. We validate our approach on the publicly available SCARED dataset and introduce a new in vivo dataset, *StereoMIS*, which includes a wider spectrum of typically observed surgical settings.

Results Our method outperforms state-of-the-art methods on average and more importantly, in difficult scenarios where tissue deformations and breathing motion are visible. We observed that our proposed weight mappings attenuate the contribution of pixels on ambiguous regions of the images, such as deforming tissues.

Conclusion We demonstrate the effectiveness of our solution to robustly estimate the camera pose in challenging endoscopic surgical scenes. Our contributions can be used to improve related tasks like simultaneous localization and mapping (SLAM) or 3D reconstruction, therefore advancing surgical scene understanding in minimally invasive surgery.

Keywords Camera pose estimation · Endoscopic surgery · Deep declarative network

Introduction

Camera pose estimation is a well-established computer vision problem at the core of numerous applications of medical robotic systems for minimally invasive surgery (MIS). With a variety of methods proposed in recent years, most approaches have focused on Simultaneous Localization and Mapping (SLAM) and Visual Odometry (VO) frameworks to solve the pose estimation problem. Well-established techniques such as ORB-SLAM2 [1] and ElasticFusion [2] have shown great promise in rigid scenes. More recently, non-rigid cases in MIS using monocular [3–5] and stereo-

scopic cameras [6–8] have also been studied. Yet to this day, pose estimation in typical MIS settings remains particularly difficult due to deformations caused by instruments and breathing, self or instrument-based occlusions, texture-less surfaces and tissue specularities.

In this work, we tackle the problem of pose estimation in such difficult cases when using a stereo endoscopic camera system.

This allows depth estimation to be computed from parallax, which has been shown to improve robustness of SLAM methods [1]. In contrast to [6, 7] which assume the tissue is smooth and locally rigid, respectively, we avoid making assumptions on the tissue deformation and topology. Instead, we propose a dense stereo VO framework that handles tissue deformations and the complexity of surgical scenes. To do this, our approach leverages geometric pose optimization by inferring where to look at in the scene. At its core, our method uses a Deep Declarative Network (DDN) [9] to enable backpropagation of gradients through the pose optimization.

✉ Michel Hayoz
   michel.hayoz@unibe.ch

1   ARTORG Center, University of Bern, Bern, Switzerland

2   Department of Visceral Surgery and Medicine, Inselspital, Bern, Switzerland

3   Applied Research, Intuitive Surgical, Sunnyvale, USA

More specifically, we propose to integrate two adaptive weight maps with the role of balancing the contribution of two geometrical losses and the contribution of each pixel toward each of these losses. We learn these adaptive weight maps using a DDN with the goal of solving the pose estimation problem, inspired by the recent DiffPoseNet [10] approach. Similarly to theirs, our method exploits the expressiveness of neural networks while leveraging robustness from the geometric optimization approach. This allows our method to adapt the contribution of the image region depending on the image content, for each loss but also between the two losses. We thoroughly evaluate our approach by characterizing its performance in comparison with the state-of-the-art on various practical scenarios: rigid scenes, breathing, scanning and deforming sequences. This validation is performed on two different datasets, and we show that our method allows for more precise pose estimation in a wide range of cases.

Method

In the following, we present our depth-based pose estimation approach from an optimization perspective. We first derive our method in terms of context-specific adaptive weight maps in the pose estimation optimization and then show how to learn these from data in an end-to-end way using DDNs to facilitate differentiation [9]. Our proposed method is illustrated in Fig. 1, and we provide a notation overview in Table 1.

For all images in an image sequence, we first employ RAFT [11] to establish correspondences between frames for both the stereo and temporal domain. This model allows disparity and optical flow estimation to be computed simultaneously, based on the observation that both share similar constraints on relative pixel displacements. From these estimates, we extract the horizontal component of the parallax flow $\mathcal{F}'_t$ at time t as the stereo disparity to compute depth maps $\mathcal{D}_t$. As we would typically expect large vertical disparities in areas of low-texture or for de-calibrated stereo endoscopes, we use this parallax flow $\mathcal{F}'_t$ as input for the weight map estimation described in the following.

Pose estimation

In contrast to most existing VO methods, we estimate the camera motion exclusively based on a geometric loss function given that photometric consistency is entailed in optical flow estimation. We thus compute a 2D residual function based on a single-depth map by,

$$r_{2D}(\mathbf{p}_t, \mathbf{x}) = \sqrt{\frac{1}{XY}} \left\| \pi_{2D}\left(\exp(\mathbf{p}_t)\, \pi_{3D}(\mathcal{D}_t, \mathbf{x}) \right) - (\mathbf{x} + \mathcal{F}_t(\mathbf{x})) \right\|_2 , \tag{1}$$

where $\pi_{2D}(\exp(\mathbf{p}_t)\, \pi_{3D}(\mathcal{D}_t, \mathbf{x}))$ is the pixel location based on depth projection and the relative camera pose $\mathbf{p}_t$ that aligns views $\mathcal{I}_t^{(l)}$ to $\mathcal{I}_{t-1}^{(l)}$. We scale these residuals by the image dimensions X and Y to make values independent of the image size. Note that we normalize depth maps by the maximum expected depth value, such that rotation and translation components of $\mathbf{p}_t$ have the same order of magnitude and thus equally contribute to the residuals, which is important for a well-conditioned optimization. Ideally, a projected pixel position coincides with the optical flow when the observed scene is rigid and when flow and depth maps are correct.

While minimizing $r_{2D}(\mathbf{p}_t, \mathbf{x})$ helps to reliably estimate the camera motion in rigid scenes, detection of deformations is most effective by looking at the displacement of points in 3D space. To address this need, we propose to leverage the depth map at $t - 1$ and introduce a 3D residual function by,

$$r_{3D}(\mathbf{p}_t, \mathbf{x}) = \left\| \exp(\mathbf{p}_t)\, \pi_{3D}(\mathcal{D}_t, \mathbf{x}) - \pi_{3D}(\mathcal{D}_{t-1}, \mathbf{x} + \mathcal{F}_t(\mathbf{x})) \right\|_2 , \tag{2}$$

which measures the point-to-point alignment of the re-projected depth maps. As opposed to [2], we avoid using a point-to-plane distance as it is less constrained on planar surfaces such as organs (*e.g.*, liver). While a known weakness of the point-to-point distance is its sensitivity to noise in regions with large perspectives, we mitigate this effect by combining 2D and 3D residuals. Intuitively, we expect $r_{2D}(\mathbf{p}_t, \mathbf{x})$ to be most accurate when camera motion is large and $r_{3D}(\mathbf{p}_t, \mathbf{x})$ when deformations are significant. Similar to [11], we use bilinear sampling to warp point clouds from $\pi_{3D}(\mathcal{D}_{t-1}, \mathbf{x})$ to $\pi_{3D}(\mathcal{D}_{t-1}, \mathbf{x} + \mathcal{F}_t(\mathbf{x}))$, using our optical flow estimates.

In contrast to conventional scalar-weighted sum of residuals, we propose to weigh each residual using a dedicated weight map that is inferred from the image data. The final residual is computed as,

$$r(\mathbf{p}_t, \mathbf{x}) = \omega_{2D}(\mathbf{x})\, r_{2D}(\mathbf{p}_t, \mathbf{x}) + \omega_{3D}(\mathbf{x})\, r_{3D}(\mathbf{p}_t, \mathbf{x}) , \tag{3}$$

where $\omega_{2D}(\mathbf{x})$ and $\omega_{3D}(\mathbf{x})$ are the per-pixel weight maps for the 2D and 3D residuals, respectively.

At its core, our hypothesis is that we can learn how the weight maps should combine the contributions of both $\omega_{2D}(\mathbf{x})$ and $\omega_{3D}(\mathbf{x})$ even in challenging situations where tissue deformations are taking place. That is, the role of $(\omega_{2D}(\mathbf{x}), \omega_{3D}(\mathbf{x}))$ is to (1) weigh relative focus based on the context of tissue deformations, (2) weigh reliable residual functions (2D vs 3D) given a motion pattern and (3) balance the scale of the loss. In "Learning the weight maps" section, we detail how we learn a model to infer these weight maps.

Fig. 1 Overview of our proposed VO framework, which can be sub-divided into 3 parts (from left to right): (1) optical flow and depth are computed using RAFT [11], (2) weight maps computed and used in (3) to estimate the camera pose $\mathbf{p}_t^\star$. Weight maps ($\omega_{2D}(\mathbf{x})$, $\omega_{3D}(\mathbf{x})$) are learned via backpropagation using the Deep Declarative Network (DDN) [9]
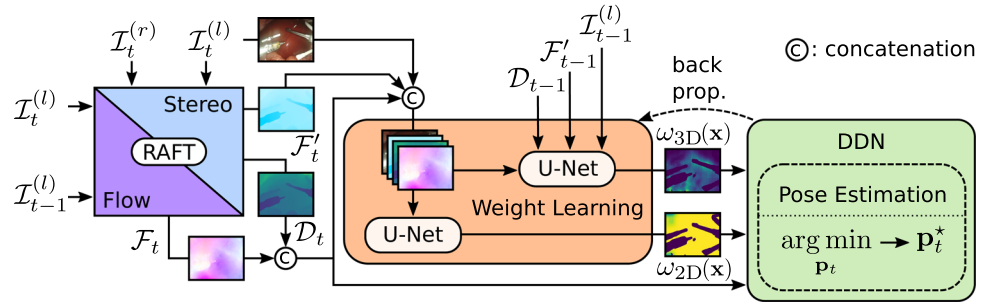
Optimization: To compute the relative pose $\mathbf{p}_t^\star \in \mathfrak{se}(3)$, we then optimize,

$$\mathbf{p}_t^\star = \operatorname*{argmin}_{\mathbf{p}_t} \left\{ \sum_{\mathbf{x} \in \Omega} r(\mathbf{p}_t, \mathbf{x})^2 \right\}, \tag{4}$$

in a Nonlinear Least-Squares (NLLS) problem. Here, Ω is a set containing all spatial image coordinates $\mathbf{x}$. We choose to optimize the pose in the Lie algebra vector space $\mathfrak{se}(3)$ because this is a unique representation of the pose and has the same number of parameters as degrees of freedom. NLLS problems are typically solved iteratively using a second-order optimizer. To do this, we use the quasi-Newton method L-BFGS [12] due to its fast convergence properties and computational efficiency. Identical to [10], we simply chain relative camera poses to obtain the full trajectory.

Learning the weight maps

In Eq. (3), we propose to *learn* residual weight maps $\omega_{2D}(\mathbf{x})$ and $\omega_{3D}(\mathbf{x})$, as determining these otherwise is not trivial.

To this end, we train a separate encoder–decoder network, denoted by $g(\cdot)$, for each weight map. The input to these networks is all the elements used to compute residuals,

$$\omega_{2D}(\mathbf{x}) = g\left(\mathbf{x}, \mathcal{I}_t^{(l)}, \mathcal{D}_t, \mathcal{F}_t, \mathcal{F}_t', \boldsymbol{\theta}_{2D}\right), \tag{5}$$

$$\omega_{3D}(\mathbf{x}) = g\left(\mathbf{x}, \mathcal{I}_t^{(l)}, \mathcal{D}_t, \mathcal{F}_t, \mathcal{F}_t', \mathcal{I}_{t-1}^{(l)}, \mathcal{D}_{t-1}, \mathcal{F}_{t-1}', \boldsymbol{\theta}_{3D}\right), \tag{6}$$

where $\boldsymbol{\theta}_{2D}$ and $\boldsymbol{\theta}_{3D}$ are the network parameters that are learned at training time. For $g(\cdot)$, we employ a 3-layer UNet [13] with Sigmoid activation function to ensure outputs in [0, 1].

To train $g(\cdot)$, we aim to learn weight maps that lead to improved pose estimation by minimizing the ℓ^1 supervised training loss,

$$\mathcal{L}_{\text{train}} = \|\mathbf{p}_t^\star - \mathbf{p}_t^{(\text{gt})}\|_1, \tag{7}$$

where $\mathbf{p}_t^{(\text{gt})}$ is the ground-truth pose. Because the pose optimization in Eq. (4) is not directly differentiable, we leverage a DDN [9] to enable end-to-end learning. This approach uses

Table 1 Summary of used notation

Symbol	Description
$t \in \mathbb{Z}$	Time frame index
$\mathcal{I}_t^{(l)} \in \mathbb{R}^{X \times Y \times 3}$	Rectified left stereo image at time t
$\mathcal{D}_t \in \mathbb{R}^{X \times Y}$	Depth map w.r.t. to left image at time t
$\mathcal{F}_t \in \mathbb{R}^{X \times Y}$	Optical flow from $\mathcal{I}_t^{(l)}$ to $\mathcal{I}_{t-1}^{(l)}$
$\mathcal{F}_t' \in \mathbb{R}^{X \times Y}$	Parallax flow displacement across stereo images
$\mathbf{x} \in \mathbb{Z}^2$	Pixel index in 2D Cartesian coordinate system
$\mathbf{p}_t \in \mathfrak{se}(3) \subset \mathbb{R}^6$	Relative pose from t to $t-1$ in Lie algebra space
$\mathbf{p}_t^\star \in \mathfrak{se}(3) \subset \mathbb{R}^6$	Relative pose solution in Lie algebra space
$\exp(\mathbf{p}) : \mathfrak{se}(3) \rightarrow \mathrm{SE}(3)$	Matrix exponential from Lie algebra to Lie group
$\pi_{2D}(\mathbf{v}) : \mathbb{R}^4 \rightarrow \mathbb{R}^2$	Projection of homogeneous 3D to 2D coordinates
$\pi_{3D}(\mathcal{D}_t, \mathbf{x}) : \mathbb{R}^{X \times Y} \times \mathbb{Z}^2 \rightarrow \mathbb{R}^4$	Re-projection of 2D to 3D homogeneous coordinates
$\mathcal{F}_t(\mathbf{x}) : \mathbb{Z}^2 \rightarrow \mathbb{R}^2$	Optical flow function across temporal domain
$\omega_{2D}(\mathbf{x}) : \mathbb{Z}^2 \rightarrow [0, 1]$	Learned per-pixel weight for 2D residuals
$\omega_{3D}(\mathbf{x}) : \mathbb{Z}^2 \rightarrow [0, 1]$	Learned per-pixel weight for 3D residuals
$\|\cdot\|_n : \mathbb{R}^m \rightarrow \mathbb{R}^+$	ℓ^n Vector norm

implicit differentiation of Eq. (4) to compute the gradients of the weight map parameters ($\boldsymbol{\theta}_{2D}$, $\boldsymbol{\theta}_{3D}$) with respect to $\mathcal{L}_{train}$. Therefore, the only requirements are that (1) the function to be optimized $\sum_{\mathbf{x} \in \boldsymbol{\Omega}} r(\mathbf{p}_t, \mathbf{x})^2$ is twice differentiable and (2) we find a local or global minimum in the forward pass.

Experiments

Datasets

We evaluate our method on two separate stereo video datasets: one containing rigid MIS scenes and another containing non-rigid scenes:

SCARED dataset [14]: consists of 9 in vivo porcine subjects with 4 sequences for each subject. The dataset contains a video stream captured using a da Vinci Xi surgical robot and camera forward kinematics. All sequences show rigid scenes without breathing motion or surgical instruments. We split the dataset into training (d2, d3, d6, d7) and testing sequences (d1, d8, and d9) where we exclude d4 and d5 due to bad camera calibrations.

StereoMIS: Additionally, we introduce a new in vivo dataset also recorded using a da Vinci Xi surgical robot. Similarly to [14], ground-truth camera poses are generated from the endoscope forward kinematics and synchronized with the video feed. While we expect errors in the absolute camera pose due to accumulated errors in the forward kinematics, relative camera motions are expected to be accurate. It consists of 3 porcine (P1, P2, and P3) and 3 human subjects (H1, H2, and H3) with a total of 16 recorded sequences. We denote sequences with Px_y where Px is the subject and y the sequence number. Sequence durations range from 50 s to 30 min. They contain challenging scenes with breathing motions, tissue deformations, resections, bleeding, and presence of smoke. We assign P1 and H1 to the training set and the rest is kept for testing.

To provide a finer grained performance characterization of methods with this data, we extract from each video a number of short sequences that visually depict one of several possible settings:

- *breathing:* only depicts breathing deformations and contains no camera or tool motion,
- *scanning:* includes camera motion in addition to breathing deformations,
- *deforming:* comprises tissue deformations due to breathing and manipulation or resection of tissue, while the camera is static.

In practice, we select 88 different, non-overlapping, and 150-frames-long sequences from P2, P3, H2, H3 and assigned

each to one of the above categories or *surgical scenarios* (see supplementary material for more information).

Implementation details

Segmentation of surgical instruments

For all experiments, we mask out surgical instrument pixels by setting corresponding residuals to 0. To do this, we use the DeepLabv3+ architecture [15] trained on the EndoVis2018 segmentation dataset [16] to generate instrument masks for each frame. Additionally, we mask out specularities, by means of maximum intensity detection, as they cause optical flow estimations to be ill-defined.

Training and inference

First, we classify all training frames from the SCARED and StereoMIS training sequences into "moving" and "static" based on the camera forward kinematics. We then randomly sample 4000 frames from each sequence keeping a balance between moving and static frames. For each sampled frame, we generate a sample pair with an interval of 1 to 5 frames. We use the forward kinematics of the camera as the ground truth poses change between two frames in a sample pair. Note that forward kinematics entail minor deviations that may propagate during training. We randomly assign 80% of the sample pairs to the training set and 20% to the validation set.

For all experiments, we resize images to half resolution (512x640 pixels). We use a batch size of 8 and the Adam optimizer with learning rate 10^{-5}. We train for 200 epochs and perform early stopping on the validation loss. We implement our method using PyTorch and train on a NVIDIA RTX3090 GPU. We reach 6.5 frames per second at test time. RAFT is trained on the FlyingThings3D dataset, and we do not perform any fine-tuning.

Metrics and baseline methods

We use trajectory error metrics as defined in [17], namely the absolute trajectory error ATE-RMSE to evaluate the overall shape of the trajectory and the relative pose errors, RPE-trans and RPE-rot, to evaluate relative pose changes from frame to frame. The ATE-RMSE is sensitive to drift and depends on the length of the sequence, whereas the RPE measures the average performance for frame-to-frame pose estimation.

As no stereo SLAM method dedicated for MIS has open-source code or is evaluated on a public dataset with trajectory ground truth, we compare our method to two general state-of-the art rigid SLAM methods that contain loop closure and are based on the rigid scene assumption:

- ORB-SLAM2 [1], a sparse SLAM leveraging bundle adjustment to compensate drift,
- ElasticFusion [2], a dense SLAM and as such closer to our proposed method.

In addition, we compare our method to [8] on the frames of the SCARED dataset for which they reported performances. For fair comparison, we use the same input depth maps for all methods.

Results

Surgical scenarios and ablation study: Table 2 reports the performance of our approach on the StereoMIS surgical scenarios. To show the importance of learning the weight maps we perform an ablation study where we evaluate the impact of (1) constant weights, denoted by *ours (w/o weight)*, where $\omega_{2D}(\mathbf{x}) = \omega_{3D}(\mathbf{x}) = 1$ for each $\mathbf{x}$; (2) our method with only 2D-residuals, denoted by *ours (only 2D)*; and (3) using only 3D-residuals, denoted by *ours (only 3D)*.

Our proposed method outperforms the baselines in all scenarios. Improvements in *breathing* and *scanning* are partly due to correct identification of errors in the optical flow and depth estimation as well as optimal balancing of the 2D and 3D residuals. Indeed, exploiting the complementary properties of 2D and 3D residuals improves the average performance. The fact that ours (only 3D) outperforms ours (only 2D) in *breathing* and *deforming* supports our intuition that it is easier to learn tissue deformations from the 3D residuals. Contrarily, in *scanning* where the optical flow is dominated by the camera motion, the 2D residuals lead to a more accurate pose estimation.

In general, it is not possible to detect or completely compensate the breathing motion on a frame-to-frame basis in our proposed optimization scheme as we cannot completely disambiguate the camera and tissue motion. However, the method learns which regions are more affected by breathing deformations and consequently assigns a smaller weight to those regions.

We note that the weight maps in Fig. 2 (see *breathing* rows) support our claims that the weight maps have low values in the dark regions (A) where we expect the optical flow to be inaccurate and where tissue is moving most (B). The *scanning* example also illustrates that the weight maps have a different response depending on the motion pattern and deformation. Note that the presence of surgical instruments has no influence on the weight maps in *scanning*, as no tissue interaction takes place. As expected, the largest improvements can be seen in the *deforming* scenario. Inspecting the two last rows in Fig. 2 reveals that regions where the instruments deform tissue (C) are correctly ignored in the pose estimation. Similarly, the region occluded by smoke (D) has low values in

the weight maps. Additionally, we observe that ω_{2D} usually has 100 times larger magnitude than ω_{3D} compensating for the different scale of the 2D and 3D residuals.

Results on full test StereoMIS sequences: Table 3 shows the pose estimation performance on the complete sequences in the StereoMIS testset. As the sequences are much longer than in the scenario experiment, accumulation of drift results in a large ATE-RMSE for all methods. Even though our frame-to-frame approach does not include any bundle adjustment or regularization over time, it still has the lowest ATE-RMSE on average. The reason for this good performance is reflected in the relative metrics RPE-trans and RPE-rot, where our method outperforms all others by almost a factor of three and five, respectively. Our method robustly estimates the pose in challenging situations, whereas ORB-SLAM2 fails in two sequences (H2_0, P2_5). Figure 3 shows two example trajectories. P2_7 does not include any tool–tissue interactions and consists of smooth camera motions. Its trajectory illustrates the drift of our method which results in an ATE-RMSE of 9.28 versus 3.76mm for ORB-SLAM2. On the other hand, P3_0 consists of strong tissue deformations and abrupt camera movements. Despite visible drift, we can see that our method is able to follow the abrupt movements. The small-scale oscillations in the trajectories are due to breathing motion. The trajectories of all test sequences and evaluation results excluding frames where the SLAM methods fail can be found in the supplementary materials.

Results on SCARED dataset: Wei et al. reported ATE-RMSE results for rigid surgical scenes of the SCARED dataset in a frame-to-model approach [8]. For the sake of fair comparison, we extend our method to SLAM by accommodating a surfel map model denoted by *ours (frame2model)*, which is equivalent to that used in [8]. Specifically, we replace input images $\mathcal{I}_{t-1}^{(l)}, \mathcal{I}_{t-1}^{(r)}$ by rendered images from the surfel map. Note, we can only adopt this approach for the SCARED dataset, as the surfel map model assumes scene rigidity. Results are provided in Table 4.

Conclusion

We proposed a visual odometry method for robust pose estimation in the challenging context of endoscopic surgeries. To do so, we learn adaptive weight maps for two geometrical residuals to leverage pose estimation performance on common surgical scenes including breathing motion and tissue deformations. Thanks to a performance analysis in common scenarios, we observed the complementary action of the 2D/3D residuals and the strong contribution of their specific weighting at pixel level. This results in better performances compared to state-of-the-art methods, on average and in the most challenging cases. We believe that our contributions are

Table 2 The ATE-RMSE (mean±std mm) for the different scenarios from the StereoMIS dataset with average over sequences (microavg.) and scenarios (macroavg.)

Scenario	Breathing	Scanning	Deforming	Microavg.	Macroavg.
# Sequences	17	60	9		
Camera motion		✓			
Breathing	✓	✓	✓		
Tool interactions			✓		
ORB-SLAM2 [1]	2.35 ± 1.81	3.26 ± 1.65	4.29 ± 2.30	3.19 ± 1.81	3.30 ± 0.97
ElasticFusion [2]	1.94 ± 0.93	4.04 ± 3.46	6.47 ± 8.64	3.88 ± 4.12	4.15 ± 2.27
Ours (w/o weight)	1.65 ± 0.97	3.01 ± 1.60	4.67 ± 2.13	2.91 ± 1.74	3.11 ± 1.51
Ours (only 2D)	1.15 ± 0.72	3.01 ± 1.66	2.83 ± 1.41	2.62 ± 1.66	2.33 ± 1.03
Ours (only 3D)	**0.78 ± 2.03**	7.02 ± 5.86	2.72 ± 1.90	5.34 ± 5.64	3.51 ± 3.20
Ours (2D & 3D)	1.01 ± 0.59	**2.89 ± 2.33**	**2.23 ± 1.07**	**2.45 ± 2.12**	**2.04 ± 0.95**
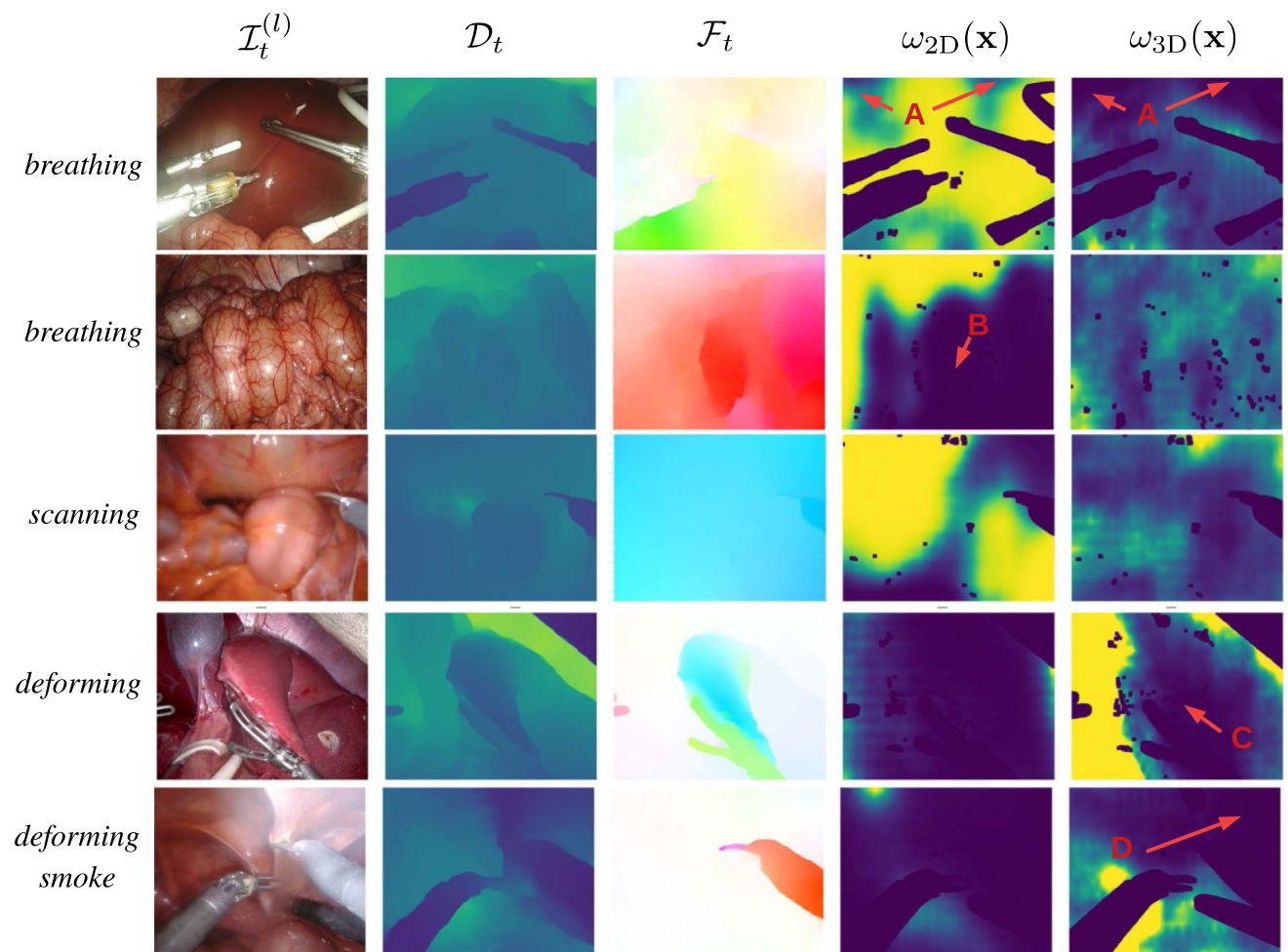
Fig. 2 Exemplary results for 5 different scenarios. Surgical instruments and specularities are masked out. From left to right: left image, its depth map, its optical flow displacement as well as its weights $\omega_{2D}(\mathbf{x})$ and $\omega_{3D}(\mathbf{x})$. Weight maps are normalized (lowest value in dark blue and highest value in yellow). Best viewed in color

Table 3 Pose estimation results on full StereoMIS test sequences for ORB-SLAM2 [1], ElasticFusion [2], and **ours**. Metrics are reported in (mean±std) when applicable

	H2	H3	P2	P3	Macroavg
ATE-RMSE (mm)					
ORB-SLAM2 [1]	18.0	**9.1**	14.0	21.4	15.6 ± 5.3
ElasticFusion [2]	30.8	72.1	33.6	37.7	43.6 ± 19.3
Ours	**10.9**	21.2	**13.8**	**8.8**	**13.7 ± 5.4**
RPE-trans (mm)					
ORB-SLAM2 [1]	0.20 ± 0.43	0.24 ± 0.25	0.35 ± 0.46	0.54 ± 0.47	0.33 ± 0.13
ElasticFusion [2]	0.87 ± 1.11	0.56 ± 1.03	0.81 ± 1.11	0.71 ± 0.79	0.74 ± 0.12
Ours	**0.10 ± 0.27**	**0.10 ± 0.18**	**0.16 ± 0.32**	**0.19 ± 0.31**	**0.14 ± 0.04**
RPE-rot (deg)					
ORB-SLAM2 [1]	0.16 ± 0.36	0.16 ± 0.22	0.19 ± 0.24	0.28 ± 0.27	0.20 ± 0.05
ElasticFusion [2]	0.73 ± 1.06	0.41 ± 0.96	0.50 ± 1.11	0.38 ± 0.40	0.51 ± 0.14
Ours	**0.04 ± 0.20**	**0.04 ± 0.13**	**0.07 ± 0.14**	**0.05 ± 0.10**	**0.05 ± 0.01**
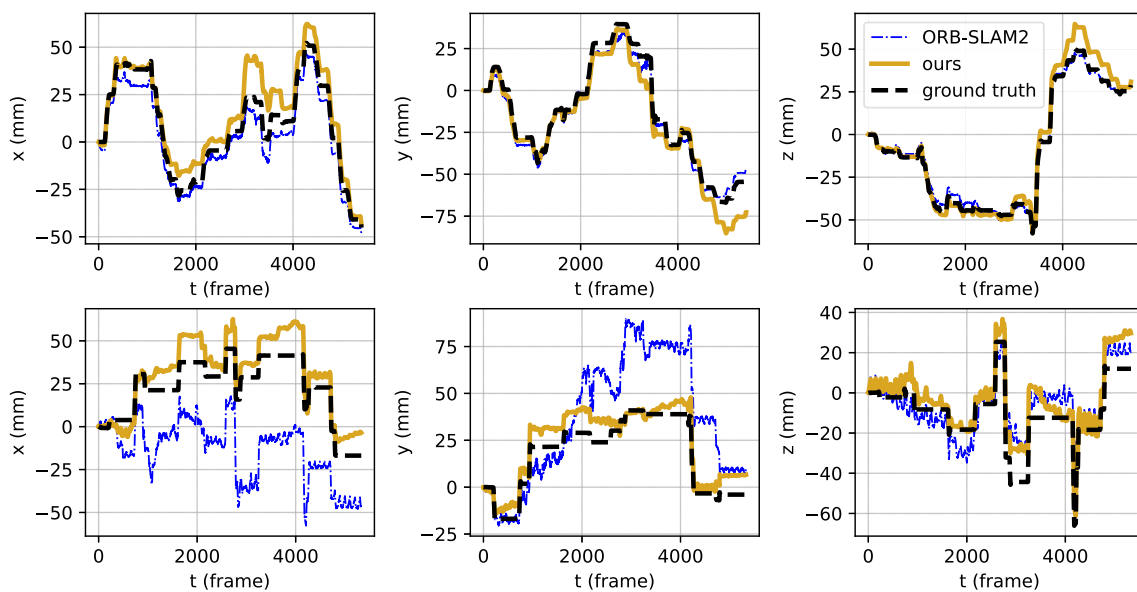
Fig. 3 Two example trajectories of the *StereoMIS* test set. **Top:** trajectory of P2_7. **Bottom:** trajectory of P3_0

Table 4 The ATE-RMSE in mm for SCARED sequences reported by [8] and microaverage over all SCARED test sequences (SCARED avg.) using surfel maps

	d1_k2	d8_k1	d9_k1	d9_k3	Avg	SCARED avg
ORB-SLAM2 [1]	0.91	2.97	4.33	3.79	3.00	2.34 ± 1.24
ElasticFusion [2]	1.02	3.62	4.30	3.36	3.08	2.91 ± 1.77
Wei et al. [8]	0.74	2.47	4.07	1.54	2.21	–
Ours (frame2model)	**0.37**	**2.08**	**2.04**	**0.84**	**1.33**	**1.38 ± 0.93**

beneficial for some SLAM components, e.g., map building, and therefore downstream applications in MIS. Future work will focus on drift and breathing compensation.

Supplementary information Appendix A: details on *StereoMIS*. Appendix B: trajectories of *StereoMIS* test set. Appendix C: Results on full *StereoMIS* sequences. Appendix D: trajectories of *SCARED* test set.

Declarations

Conflict of interest Authors declare that they have no conflict of interest.

Ethics approval All applicable international, national, and institutional guidelines for the care and use of animals were followed. All procedures performed in studies involving animals were in accordance with the ethical standards of the institution at which the studies were conducted.

Consent Not applicable.

Availability of data, materials, and code *StereoMIS* porcine data- https://doi.org/10.5281/zenodo.7727692. Code and models - https://github.com/aimi-lab/robust-pose-estimator.

References

1. Mur-Artal R, Tardós JD (2017) Orb-slam2: an open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Trans Rob 33(5):1255–1262. https://doi.org/10.1109/TRO.2017.2705103

2. Whelan T, Salas-Moreno RF, Glocker B, Davison AJ, Leutenegger S (2016) Elasticfusion: real-time dense slam and light source estimation. Int J Rob Res 35(14):1697–1716. https://doi.org/10.1177/0278364916669237

3. Lamarca J, Montiel JMM (2018) Camera tracking for slam in deformable maps. In: Computer vision-ECCV 2018 workshops, pp 730–737. https://doi.org/10.1007/978-3-030-11009-3_45

4. Gómez-Rodríguez JJ, Lamarca J, Morlana J, Tardós JD, Montiel JMM (2021) Sd-defslam: semi-direct monocular slam for deformable and intracorporeal scenes. In: 2021 IEEE international conference on robotics and automation (ICRA), pp 5170–5177. https://doi.org/10.1109/ICRA48506.2021.9561512

5. Liu X, Li Z, Ishii M, Hager GD, Taylor RH, Unberath M (2022) Sage: slam with appearance and geometry prior for endoscopy. In: 2022 International conference on robotics and automation (ICRA), pp 5587–5593. https://doi.org/10.1109/ICRA46639.2022.9812257

6. Song J, Wang J, Zhao L, Huang S, Dissanayake G (2018) Mis-slam: real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing. IEEE Robot Autom Lett 3(4):4068–4075. https://doi.org/10.1109/LRA.2018.2856519

7. Zhou H, Jayender J (2021) EMDQ-SLAM: real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos. In: Medical image computing and computer assisted intervention—MICCAI 2021, pp 331–340 . https://doi.org/10.1007/978-3-030-87202-1_32

8. Wei R, Li B, Mo H, Lu B, Long Y, Yang B, Dou Q, Liu Y, Sun D (2023) Stereo dense scene reconstruction and accurate localization for learning-based navigation of laparoscope in minimally invasive surgery. IEEE Trans Biomed Eng 70(2):488–500. https://doi.org/10.1109/TBME.2022.3195027

9. Gould S, Hartley R, Campbell D (2022) Deep declarative networks. IEEE Trans Pattern Anal Mach Intell 44(8):3988–4004. https://doi.org/10.1109/TPAMI.2021.3059462

10. Parameshwara CM, Hari G, Fermüller C, Sanket NJ, Aloimonos Y (2022) Diffposenet: direct differentiable camera pose estimation. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 6835–6844. https://doi.org/10.1109/CVPR52688.2022.00672

11. Teed Z, Deng J (2020) RAFT: recurrent all-pairs field transforms for optical flow. In: Computer vision-ECCV 2020, pp 402–419. https://doi.org/10.1007/978-3-030-58536-5_24

12. Liu DC, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. In: Mathematical programming, vol 45, pp 503–528. https://doi.org/10.1007/BF01589116

13. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015, pp 234–241. Springer https://doi.org/10.1007/978-3-319-24574-4_28

14. Allan M, McLeod AJ, Wang CC, Rosenthal J, Hu Z, Gard N, Eisert P, Fu KX, Zeffiro T, Xia W, Zhu Z, Luo H, Jia F, Zhang X, Li X, Sharan L, Kurmann T, Schmid S, Sznitman R, Psychogyios D, Azizian M, Stoyanov D, Maier-Hein L, Speidel S (2021) Stereo correspondence and reconstruction of endoscopic data challenge. arXiv:2101.01133

15. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder–decoder with atrous separable convolution for semantic image segmentation. In: Computer vision—ECCV 2018, pp 833–851. https://doi.org/10.1007/978-3-030-01234-2_49

16. Allan M, Kondo S, Bodenstedt S, Leger S, Kadkhodamohammadi R, Luengo I, Fuentes F, Flouty E, Mohammed A, Pedersen M, et al (2020) 2018 Robotic scene segmentation challenge. arXiv:2001.11190

17. Ozyoruk KB, Gokceler GI, Bobrow TL, Coskun G, Incetan K, Almalioglu Y, Mahmood F, Curto E, Perdigoto L, Oliveira M, Sahin H, Araujo H, Alexandrino H, Durr NJ, Gilbert HB, Turan M (2021) Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Med Image Anal 71:102–112. https://doi.org/10.1016/j.media.2021.102058