Simultaneous localisation and mapping for laparoscopic liver navigation : a comparative evaluation study

Conference Paper · February 2021

DOI: 10.1117/12.2582121

11 authors, including:

Reuben Docea
German Cancer Research Center
12 PUBLICATIONS **92** CITATIONS

SEE PROFILE

Micha Pfeiffer
German Cancer Research Center
40 PUBLICATIONS **700** CITATIONS

SEE PROFILE

Sebastian Bodenstedt
National Center for Tumor Diseases (NCT) Dresden, Dresden, Germany
136 PUBLICATIONS **3,468** CITATIONS

SEE PROFILE

Fiona Kolbinger
Purdue University West Lafayette
101 PUBLICATIONS **1,830** CITATIONS

SEE PROFILE

Simultaneous Localisation and Mapping for Laparoscopic Liver Navigation - A Comparative Evaluation Study

Reuben Docea[a], Micha Pfeiffer[a], Sebastian Bodenstedt[a], Fiona R. Kolbinger[a,b], Lukas Höller[c], Ines Wittig[c], Ralf-Thorsten Hoffmann[a,e], Esther G.C. Troost[a,d,f], Carina Riediger[b], Jürgen Weitz[a,b], and Stefanie Speidel[a]

[a]Translational Surgical Oncology, National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany; Helmholtz Association / Helmholtz-Zentrum Dresden – Rossendorf (HZDR), Dresden, Germany
[b]Department for Visceral, Thoracic and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
[c]National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany; Helmholtz Association / Helmholtz-Zentrum Dresden – Rossendorf (HZDR), Dresden, Germany
[d]Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
[e]Department of Radiology, Institute and Policlinic of Diagnostic and Interventional Radiology, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
[f]OncoRay - National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden - Rossendorf, Dresden, Germany

ABSTRACT

Computer-Assisted Surgery (CAS) aids the surgeon by enriching the surgical scene with additional information in order to improve patient outcome. One such aid may be the superimposition of important structures (such as blood vessels and tumors) over a laparoscopic image stream. In liver surgery, this may be achieved by creating a dense map of the abdominal environment surrounding the liver, registering a preoperative model (CT scan) to the liver within this map, and tracking the relative pose of the camera. Thereby, known structures may be rendered into images from the camera perspective. This intraoperative map of the scene may be constructed, and the relative pose of the laparoscope camera estimated, using Simultaneous Localisation and Mapping (SLAM). The intraoperative scene poses unique challenges, such as: homogeneous surface textures, sparse visual features, specular reflections and camera motions specific to laparoscopy. This work compares the efficacies of two state-of-the-art SLAM systems in the context of laparoscopic surgery, on a newly collected phantom dataset with ground truth trajectory and surface data. The SLAM systems chosen contrast strongly in implementation: one sparse and feature-based, ORB-SLAM3,[1–3] and one dense and featureless, ElasticFusion.[4] We find that ORB-SLAM3 greatly outperforms ElasticFusion in trajectory estimation and is more stable on sequences from laparoscopic surgeries. However, when extended to give a dense output, ORB-SLAM3 performs surface reconstruction comparably to ElasticFusion. Our evaluation of these systems serves as a basis for expanding the use of SLAM algorithms in the context of laparoscopic liver surgery and Minimally Invasive Surgery (MIS) more generally.

Article DOI: https://doi.org/10.1117/12.2582121
Further author information: (Send correspondence to R.D.)
R.D.: E-mail: reuben.docea@nct-dresden.de
M.P.: E-mail: micha.pfeiffer@nct-dresden.de

Keywords: SLAM, Minimally Invasive Surgery, Augmented Reality, ElasticFusion, ORB-SLAM3, Trajectory Estimation, Surface Reconstruction, Surgical Navigation

1. INTRODUCTION

Laparoscopic liver surgery poses many benefits over open liver surgery, such as lower morbidity and better cost-effectiveness.[5] However, it is especially difficult to perform for several reasons: it complicates hand-eye coordination, it is difficult to anticipate the locations of significant blood vessels, and the surgeon cannot palpate the liver to feel for tumors. We thus aim for a Computer-Assisted Surgery (CAS) system to aid liver navigation by highlighting such structures of interest.

At present, surgical navigation often uses patient-specific preoperative computed tomography (CT) or magnetic resonance imaging (MRI) models. These models are used both during planning and operation, and serve to facilitate the surgery.[6] However, the surgeon must mentally fit what they observe from the preoperative models onto what they see during surgery. Navigation is usually done using the laparoscopic video stream, complemented by ultrasound, or in rare cases intraoperative CT.[7,8] The CAS system for which we aim requires the knowledge of the laparoscope pose, and a map of the intraoperative liver surface, to which the preoperative liver model is registered. Currently, the pose of the laparoscope is tracked intraoperatively using an optical tracking device.[9] Such devices frequently lose tracking, require additional setup time, and are limited to non-bendable devices. As the name suggests, Simultaneous Localisation and Mapping (SLAM) methods build a map of the environment, within which they localise the camera.[10] Of these, Visual SLAM (vSLAM) methods use camera sensor data alone. Sparse SLAM methods provide pose information,[1–3] while dense SLAM methods additionally produce a dense map.[4,11] With stereo laparoscopic data being commonly available, we look to vSLAM methods to provide the laparoscope pose and, if possible, the surface of the liver.

Recently, SLAM has received increasing attention due to it's great potential in applications such as autonomous driving, exploration of new environments and, in a similar vein as this work, Augmented Reality (AR).[10] However, it is seldom used in surgery. Sparse methods such as ORB-SLAM3[1–3] are feature-based, and perform pose-graph optimisation and bundle adjustment on loop closures, emphasising camera pose estimation. As for dense SLAM methods, recent significant works begin with KinectFusion,[12] from which Kintinuous[13] is built. Kintinuous, as with many sparse methods, features a pose-graph. A contrasting dense method is the map-centric ElasticFusion, which optimises a surfel-based map rather than a pose-graph, and is able to handle loopy trajectories.[4]

The intraoperative scene poses unique challenges: homogeneous surface textures, sparse visual features, a deforming environment, specular reflections from endoscopic light sources, small mapped volumes and camera motion patterns specific to laparoscopy, to name a few. Only a handful of works apply SLAM to minimally invasive surgery (MIS). Grasa et al.[14] apply extended kalman filter (EKF) SLAM to monocular sequences from the abdominal cavity. Other works apply ORB-SLAM to MIS, such as Mahmoud et al.,[15] who build upon the method to provide a dense scene reconstruction. Similarly, Song et al. extend ORB-SLAM2[1,2] to a dense deformable configuration they refer to as MIS-SLAM.[16]

In this work, we evaluate two state-of-the-art vSLAM systems, ORB-SLAM3[1–3] and ElasticFusion,[4] which are sparse and dense methods, respectively. We conduct this evaluation in the context of laparoscopic liver navigation using a new phantom dataset with ground truth trajectories and surfaces (which were collected at the NCT imaging platform), as well as a qualitative assessment on two sequences from real laparoscopic liver surgeries. We take the metrics of trajectory estimation error and surface reconstruction error as tools to assess each algorithm in relation to the intricacies of SLAM in MIS. Given the nascent status of SLAM in MIS, our evaluation of contrasting SLAM algorithms on data from both phantom and human laparoscopic liver surgeries will serve as a base for further development in this area.

2. METHODS

2.1 ElasticFusion

ElasticFusion by Whelan et al.[4] takes as input RGB-D image pairs, which we provided using a stereo laparoscope and depth estimation as in Yang et al.[17] In a similar strategy to many SLAM methods, tracking and mapping are

separate processes. ElasticFusion does not implement pose-graph optimisation, but rather optimises the depth and colour of a surfel-based model of the environment, drawing on work from Keller et al.[18] These surfels are grouped as either active or inactive based on the time that has elapsed since they were last updated, δ_t. With each incoming RGB-D pair, a camera pose, P_S, is estimated by registering them against the active portion of the map by a combined geometric-photometric loss-minimisation procedure. A randomised fern encoding based place recognition system from Glocker et al.[19] is used to detect loop closures, upon the detection of which a non-rigid deformation of the surfel map is performed.

ElasticFusion receives depth data in *uint16* format, where the value represents the distance in millimeters. In the application of laparoscopic liver surgery, the maximum distance of structures from the camera is likely to be around 30cm, which is 10% of the default depth range for ElasticFusion with millimeter input. Given the smaller scale of the abdomen, depth data was provided to ElasticFusion in 0.1mm scale. Lastly, to reduce computation time of depth estimation, Yang et al. introduce a maximum disparity up to which the method will compute. This value was left unchanged, which corresponded to around 2cm in depth.

2.2 ORB-SLAM3

ORB-SLAM3 is a versatile method, usable with monocular, stereo and RGB-D input.[1–3] Of these, we used the stereo configuration as this reflects the type of data acquired from our laparoscopic setup. The system utilises three main threads: one for tracking, a second for local mapping and a third for the detection of loop closures and the optimisation of a pose graph. A fourth thread is also used for performing full bundle adjustment (BA) following loop closure. As suggested by the name, ORB-SLAM3 uses ORB features[20] for place recognition, due to their efficiency in computation and efficacy in recall. The system maintains a covisibility graph, whose nodes represent keyframes and edges the numbers of shared observations, which is refined through pose graph optimisation. For place recognition, loop detection and relocalisation, ORB-SLAM3 implements a novel algorithm, which checks for geometrical consistency and local consistency in candidate keyframes. Furthermore, within ORB-SLAM3, a procedure to handle the creation and merging of multiple maps is implemented.

While internally ORB-SLAM3 uses a sparse representation of the map, to obtain a dense output we used a method based on that recommended by the ORB-SLAM2 authors:[2] From each camera pose ORB-SLAM3 provides, we used the estimated disparity[17] of the corresponding stereo image pair to project the left image into 3D space. To reduce noise, the space occupied by each new point was first checked to determine whether it was already occupied by a point from previous frames, and whether that point had a similar hue. If so, the previous point's internal counter was raised by one to indicate that it had been detected in multiple camera frames. If no previous point was found, the new point was added. After fusion, we only accepted points into the final dense representation of the map if they had been seen a certain number of times: in this case, 30. The dense reconstruction that we performed was done post-hoc, after ORB-SLAM3 had been run on a sequence and with the full trajectory that it output. Here, the use of Robotic Operating System (ROS) to pass data between processes may have introduced timing delays. However, the impact on results is expected to be negligible.

2.3 Phantom Data

Phantom data was acquired using the imaging platform at NCT Dresden (Figures 1 & 2), along with an experimental setup consisting of: the OpenHELP phantom, a human body model for open and laparoscopic surgery research;[21] a silicone liver replica positioned in different ways, giving a set of six distinct deformations, D; various materials in place of other organs such as the stomach and fatty tissue. More specifically, the different deformations were achieved by changing the position of the silicone liver, rearranging different substitute organs or materials, and pulling or compressing the liver with cloths in a way that mimics the falciform ligament which suspends it. A rigid stereo laparoscope (KARL STORZ TIPCAM 1*) with a 30-degree angled tip and a baseline of 6mm was used to record multiple sweeping views over the abdomen at 30fps with a resolution of 960x540. This recording process was repeated for every deformation $d \in D$, in each case yielding a set of recordings R. Between each recording $r \in R$, the intensity of the light source attached to the laparoscope and the starting position for the recording were varied. An optical tracking device, the *Polaris Spectra*†, was used to record

*https://www.karlstorz.com/de/en/index.htm

†https://www.ndigital.com/medical/products/polaris-family/

Figure 1: Experimental setup in the NCT imaging platform.

poses for both the phantom and laparoscope in 3D space. A hand-eye calibration[22] from a laparoscope-attached optical marker M_{camera} to the left laparoscope camera allowed the camera pose to be tracked using the determined offset, P_{cal}. For each deformation, a CT scan of the full phantom was taken with slice thickness 0.6mm. During all recordings, a Robot Operating System (ROS)[‡] implementation of ORB-SLAM2[2] was used to pace recordings, in an attempt to ensure that they were not overly difficult to track - a set sequences which contain movement that is too fast is not useful for this evaluation as the difficulty of the sequences would likely cause the methods to fail and, therefore, evaluating their accuracy while operational is more difficult. As with the surface reconstruction in Section 2.2, timing issues associated with passing data between processes with ROS may have occurred. Although in all cases the influence of this is believed to be negligible, in one of the recordings used in Section 2.4, poses were extracted with a lower frequency and thus fewer poses were acquired (though still high in number).

2.4 Parameter Selection

Initially, there were 33 individual recordings, r, from the six different deformations of the phantom. Six of these were rejected due to the presence of a 'lid' on the phantom belly which was not scanned in CT, or due to prolonged occlusions of the optically tracked markers. This left 27 recordings, which were spread approximately evenly over the six deformations D. Of the remaining recordings, one was selected at random from each of the three deformations which had the lowest number of recordings. These were used to tune the parameters of the algorithms by carrying out a parameter sweep in two stages. All parameter values used and the results of either stage can be found in Appendices A.1 and A.2.

In the first stage: default parameters of the algorithms were varied by $\pm 20\%$ for parameters where it wasn't known how they might impact performance; default parameters where there was good reason to believe their increase or decrease would result in better performance were trialed at +20% and +40%, or -20% and -40%, respectively; and default parameters affected by image resolution were scaled proportionally with how the number of pixels of our laparoscope images compared to the number of pixels of the resolution the default parameters of either algorithm were specified for. For each setting s of a total of 81 settings S, the respective algorithms were run on each recording r three times. The mean translation error (absolute euclidean distance), E_{mean}, for each of the three runs was computed, and the median E_{median} of these was kept. For each $s \in S$, the mean of E_{median}

[‡]https://www.ros.org/

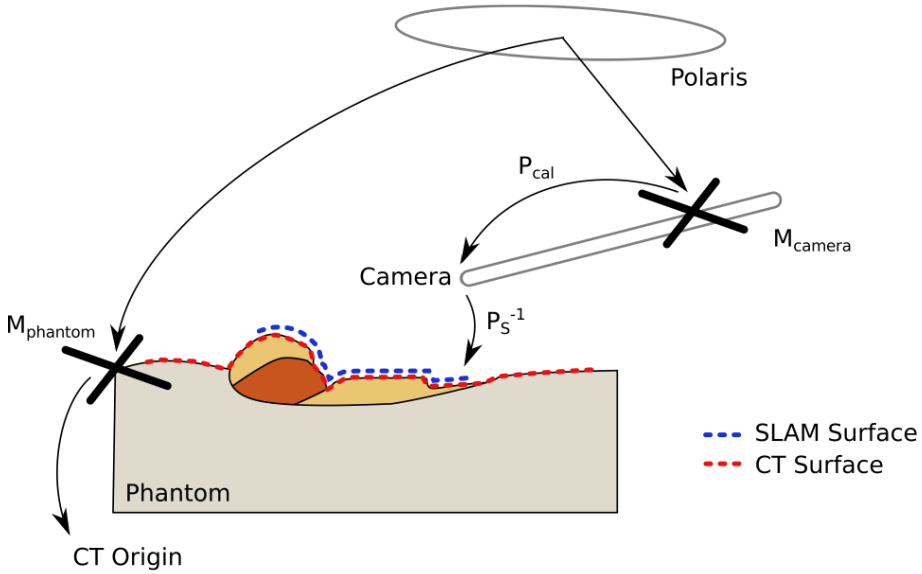

Figure 2: Diagram of experimental setup. The markers $M_{phantom}$ and M_{camera}, attached to the phantom torso and the laparoscope, respectively, were tracked by the *Polaris* optical tracking system. The offset between $M_{phantom}$ and the origin of the CT scanner was calibrated by identifying five small Teflon markers in both the CT scan and by using an optically tracked stylus. A hand-eye calibration was used to determine the offset P_{cal} between optical camera center and M_{camera}. The camera pose estimated by the SLAM system for each time frame, P_S, gave the offset between the SLAM surface and the camera. By chaining these transformations together, all data sources were mapped into a single, common coordinate system.

across the three sequences was computed to give an average error, $\bar{E}_{median}$. The settings, which produced the lowest $\bar{E}_{median}$, were taken forward to the next stage. The second stage of the parameter sweep followed the same procedure as the first stage, with the difference that the carried-forward parameters were only varied ±10% of the default values for all parameters. The exception to this rule was again for those which were modified in accordance with pixel count.

Although pose errors are composed of a translation and rotation error, the choice of minimising the mean translation error was due to translation and rotation errors being difficult to reconcile, and also on the assumption that the two errors are correlated. To constrain the parameter search, only parameters of the algorithms which do not relate solely to loop closure were optimised. This was also partly due to the fact that only three sequences were used in the parameter sweep, which limited the number of loop closures that the algorithms may be required to handle. Thus, any parameters that may have otherwise been chosen, which relate to loop closure only, could be overly influenced by these few examples.

2.5 Quantitative Evaluation

In this section, the algorithms, using the best parameters determined through the parameter sweep, were evaluated on the remaining 16 recordings from the three remaining deformations.

2.5.1 Trajectory Evaluation

Due to occlusions or otherwise, pose estimates for the camera given by *Polaris* were not available for each frame. The coordinate systems were aligned by setting the first concurrent poses from the tracking system and SLAM algorithm equal to one another, and the changes in pose relative to this time point were compared for subsequent concurrent pairs. Here, the translation and rotation (absolute angular error) errors were computed. The algorithms were run on each recording three times and, as in Section 2.4, the trial representing E_{median} was taken as the representative result for that recording. The trajectory of the same E_{median} trial was also taken forward for surface reconstruction with ORB-SLAM3, to be used in Section 2.5.2.

2.5.2 Surface Reconstruction Evaluation

To evaluate the SLAM surface reconstructions, they were compared to the CT reference data. Since the CT data were in CT-space and the SLAM systems constructed their surfaces in their own coordinate system (usually with the origin at the camera pose of the first video frame), this required a transformation of both data sources into a common coordinate system. Following Figure 2, transformations from both systems to the Polaris coordinate system could be determined: The hand-eye calibration P_{cal} linked the current camera pose to the camera marker M_{camera}, the position of which was tracked by the Polaris and was thus known in Polaris space. The Phantom torso contained 5 Teflon markers which were manually localized on the CT scan as well as via a tracked stylus, allowing the calculation of a transformation from CT origin to the phantom marker $M_{phantom}$. This marker's position was again tracked by the Polaris system, completing the transformation chain from the CT data to the Polaris coordinate system. As in Section 2.5.1, poses for the camera (P_S and M_{camera}) were chosen by identifying the first time frame at which both were available.

To obtain reference surface models from the CT data of each deformation $d \in D$, voxels belonging to the phantom were first extracted via a threshold filter. This binary segmentation was meshed to create a surface polygon model. To reduce the redundantly high number of polygons and to reduce noise, the number of surface elements was reduced to 10% of the original. The resulting models also contained sub-surface elements which could never be seen by the SLAM algorithm and were thus manually selected and discarded.

Due to small errors in trajectory estimation, hand-eye calibration, tracking and timing, this surface registration process was likely to result in a not quite perfect alignment. Therefore, an Iterative Closest Point (ICP) algorithm was run to further refine the rigid transformation. Examples of the aligned CT scans and SLAM-generated surfaces can be seen in Figure 3. With both surfaces in the same coordinate system, for each point in the SLAM surface the distance to the closest point on the aligned CT surface was recorded to determine the overall surface reconstruction error.

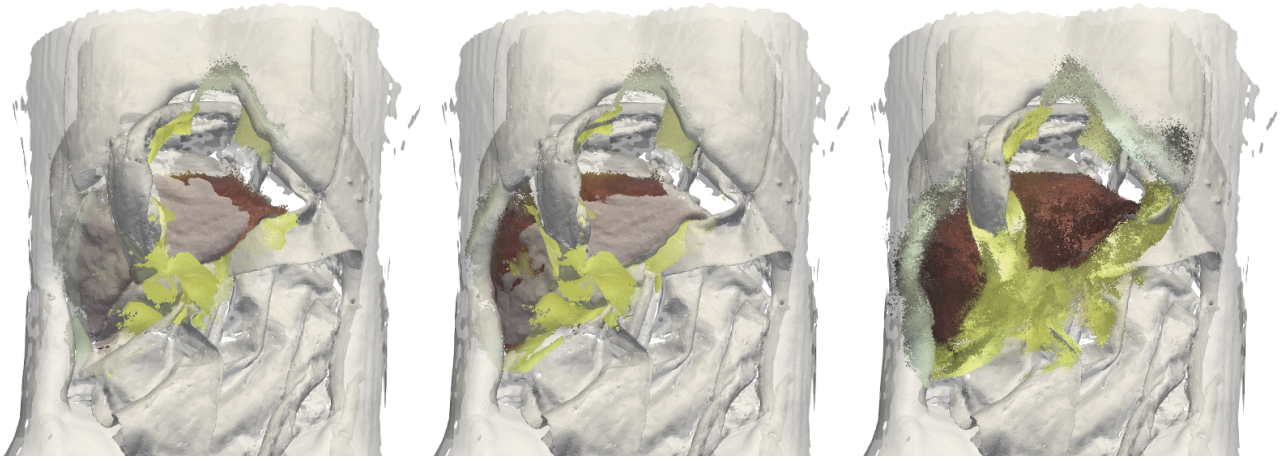

Figure 3: Surface reconstructions for Deformation 1, Sequence 2. From left to right: ElasticFusion result (aligned via Polaris tracking information only), Elastic Fusion result (aligned via tracking *and* ICP), ORB-SLAM3 result (aligned via tracking *and* ICP).

2.6 Qualitative Evaluation on Human Laparoscopic Liver Surgery

As no ground truth trajectory or surface data were available for stereo sequences from real laparoscopic liver surgeries, a qualitative evaluation was performed for the algorithms at hand. This involved observing the behaviour of the algorithms on data from two human in-vivo exploratory videos. This means that there was endoscope motion, drastic lighting changes and breathing motion, but no further tissue manipulation by the surgeons. The laparoscope used in data acquisition for these experiments was different from that used to acquire data with the phantom, and had a lower resolution of 640x512. Furthermore, the images provided as input to the

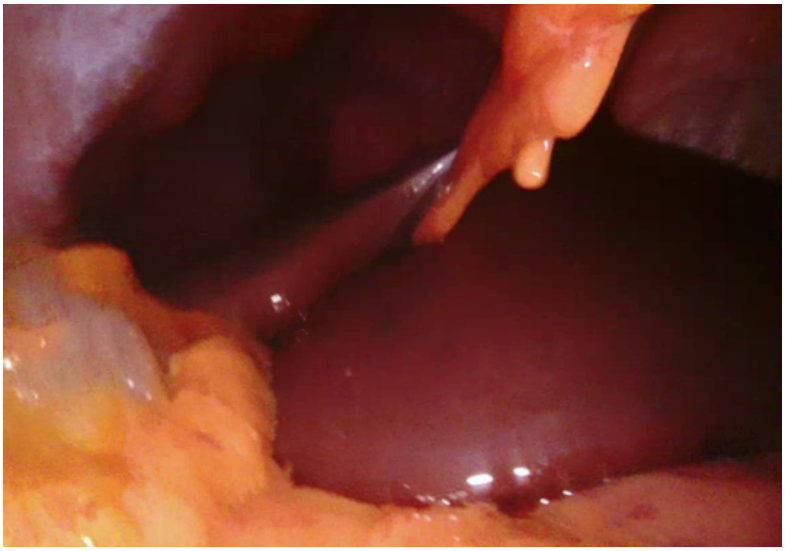

Figure 4: Sample rectified image from human in-vivo exploratory videos, as used in Section 2.6.

SLAM algorithms, after rectification, cropping, and the adjustment of calibration parameters, had resolutions of 536x373 in width and height. An example of the images which were input to the SLAM algorithms can be seen in Figure 4.

The local Institutional Review Board (ethics committee at the Technische Universität Dresden) reviewed and approved this study (approval number: BO-EK-137042018). Written informed consent was waived.

3. RESULTS AND DISCUSSION

3.1 Parameter Sweep

Through the parameter sweep process, parameters were found that reduce the mean $\bar{E}_{median}$ error across the three sequences used in the parameter sweep. In the case of ElasticFusion, the error was reduced from 58.5mm to 24.1mm; while for ORB-SLAM3 the error fell from 5.1mm to 4.6mm. The parameters trialed in the sweep stages, and those found to be best for both algorithms, can be found in Appendices A.1 and A.2. The timing issues associated with ROS and mentioned in Section 2.3 are unlikely to have impacted the results, as more than 500 tracked poses were available for each sequence.

3.2 Quantitative Evaluation

3.2.1 Trajectory Evaluation

The results from having ran the algorithms on the remaining 16 recordings can be seen in Table 1. In almost every case, ORB-SLAM3 achieved a considerably lower error in both translation and rotation. The average translation errors across the 16 sequences were 24.5mm and 4.7mm for ElasticFusion and ORB-SLAM3, respectively. As for errors in rotation, the averages across the 16 sequences were 12.9 degrees and 2.0 degrees for ElasticFusion and ORB-SLAM3, respectively. A representative sample for the typical performance of either algorithm can be seen in Figure 5.

Of the 16 sequences, the only sequence for which ElasticFusion determined itself to be lost was that of Deformation 3, Sequence 3 in Table 1. This loss of tracking only lasted for around 4.4% of the sequence. It appears that the larger errors that ElasticFusion suffered did not result from the algorithm becoming 'lost', but from it tracking incorrectly. This is illustrated in Figure 5. Here, ElasticFusion began by tracking well and following the ground truth pose. At a certain point, however, it deviated from the ground truth where ORB-SLAM3 continued to follow. Presumably, this resulted from the optimal transform computed through ElasticFusion's geometric-photometric loss being more ambiguous for the section where this deviation occured.

The deviation lead to the map being built out incorrectly, and ultimately the map and trajectory were disjoint with respect to ground truth.

In the parameter sweep, the assumption was made that error in translation and error in rotation are correlated. With respect to Table 1, the correlations between these metrics (computed with Pearson's correlation) were found to be 0.994 (p=1.09e-16) and 0.482 (p=0.0430) for ElasticFusion and ORB-SLAM3, respectively. The statistical significance of these correlation tests ($p<0.05$) confirm that there is a strong relationship between the two metrics. Interestingly, however, the strength of the relationship is much greater for ElasticFusion than for ORB-SLAM3, which may in part be due to the way ElasticFusion happens to generate disjoint maps: the building out of a warped map, and a failure to detect this as incorrect, lead to an offset in both rotation and translation during frame-to-model tracking and ultimately to a higher error in both metrics. This did not occur for ORB-SLAM3 as it did not sufficiently misinterpret translations within the sequence, although this may be a result of the sequences being recorded under the 'supervision' of ORB-SLAM2.[2]

Table 1: Mean and Standard deviation of translation errors (mm) and rotation errors (rad) on 16 sequences. Values in bold represent the better result.

		ElasticFusion		ORB-SLAM3	
		Translation	Rotation	Translation	Rotation
Deformation	Sequence	Mean ± Std	Mean ± Std	Mean ± Std	Mean ± Std
1	1	62.0 ± 20.8	0.597 ± 0.550	**5.2 ± 2.1**	**0.020 ± 0.023**
1	2	8.6 ± 9.9	0.117 ± 0.233	**5.6 ± 3.1**	**0.024 ± 0.027**
1	3	10.7 ± 5.9	0.100 ± 0.072	**4.0 ± 1.7**	**0.029 ± 0.025**
1	4	4.0 ± 2.5	0.033 ± 0.016	**3.8 ± 2.1**	**0.032 ± 0.022**
1	5	32.9 ± 24.0	0.254 ± 0.293	**5.9 ± 2.0**	**0.050 ± 0.051**
1	6	5.5 ± 2.5	0.039 ± 0.037	**3.9 ± 1.4**	**0.026 ± 0.028**
2	1	15.7 ± 18.0	0.114 ± 0.145	**3.2 ± 1.7**	**0.029 ± 0.029**
2	2	8.5 ± 4.6	0.046 ± 0.050	**6.3 ± 2.1**	**0.028 ± 0.016**
2	3	82.0 ± 68.5	0.762 ± 0.834	**3.7 ± 1.7**	**0.026 ± 0.031**
2	4	18.1 ± 8.2	0.168 ± 0.177	**7.4 ± 1.2**	**0.038 ± 0.012**
2	5	8.7 ± 5.1	0.106 ± 0.128	**5.6 ± 2.5**	**0.023 ± 0.023**
3	1	23.7 ± 14.0	0.176 ± 0.175	**5.2 ± 2.6**	**0.028 ± 0.033**
3	2	13.1 ± 8.5	0.158 ± 0.126	**2.7 ± 0.9**	**0.091 ± 0.011**
3	3	8.5 ± 6.6	0.104 ± 0.151	**4.7 ± 1.9**	**0.014 ± 0.016**
3	4	45.4 ± 47.8	0.418 ± 0.425	**3.5 ± 1.4**	**0.029 ± 0.035**
3	5	79.6 ± 44.9	0.766 ± 0.888	**9.5 ± 3.0**	**0.109 ± 0.033**
	Grand Mean	24.5	0.226	4.7	0.035

3.3 Surface Reconstruction Evaluation

Table 2 shows the point-to-point surface reconstruction errors on all 16 sequences for both ElasticFusion and ORB-SLAM3. Similarly, Figure 6 contains box plots of the same surface reconstruction errors. From Figure 6, it seems that the error in surface reconstruction of ElasticFusion varied more greatly between sequences. Although ElasticFusion had a higher mean reconstruction error across all sequences than ORB-SLAM3 (2.99mm as opposed to 2.33mm), the very low errors it achieved for a handful of sequences suggests that it fulfills its authors' objective of building a more consistent map. However, as evidenced by the greater variability in error, this high accuracy and consistency is dependent on a good trajectory estimation. Despite the difference in the grand means across

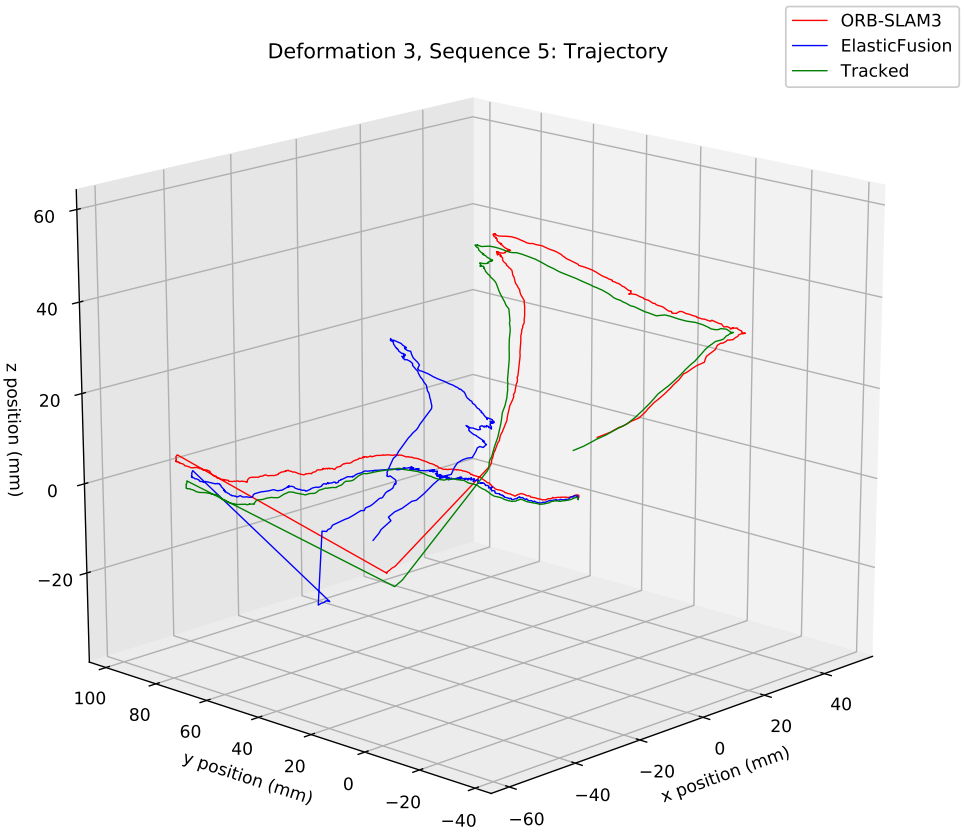

Figure 5: Representative trajectories for ElasticFusion and ORB-SLAM3, alongside the corresponding ground-truth optically-tracked trajectory for Deformation 3, Sequence 5. ElasticFusion begins by tracking well before suddenly doing so erroneously.

Table 2: Mean, standard deviation and maximum surface reconstruction errors (in mm) of the two methods for each sequence. Values in bold represent the better result.

		ElasticFusion		ORB-SLAM3	
Deformation	Sequence	Mean ± Std	Max	Mean ± Std	Max
1	1	4.92 ± 4.28	28.91	**3.47 ± 2.81**	22.75
1	2	2.75 ± 3.64	35.02	**2.61 ± 2.24**	27.15
1	3	**1.14 ± 1.08**	10.92	1.87 ± 1.66	24.25
1	4	**0.57 ± 0.55**	10.32	1.52 ± 1.13	7.75
1	5	**2.36 ± 2.27**	28.61	3.00 ± 2.61	25.57
1	6	**1.02 ± 0.83**	13.36	1.69 ± 1.40	12.35
2	1	2.83 ± 2.51	23.93	**1.77 ± 1.53**	13.91
2	2	**1.39 ± 1.61**	14.49	2.21 ± 1.70	11.90
2	3	6.37 ± 7.45	81.14	**1.68 ± 1.44**	20.68
2	4	**1.16 ± 1.19**	11.55	3.48 ± 3.02	25.35
2	5	**1.05 ± 1.20**	15.05	1.74 ± 1.39	15.43
3	1	3.98 ± 4.55	33.87	**2.61 ± 2.55**	20.65
3	2	7.55 ± 7.14	70.60	**3.03 ± 2.97**	28.67
3	3	**1.42 ± 1.24**	17.78	2.04 ± 2.17	22.21
3	4	6.56 ± 7.11	49.59	**2.21 ± 2.26**	22.25
3	5	2.73 ± 2.34	17.52	**2.28 ± 2.44**	28.52
	Grand Mean	2.99		2.33	

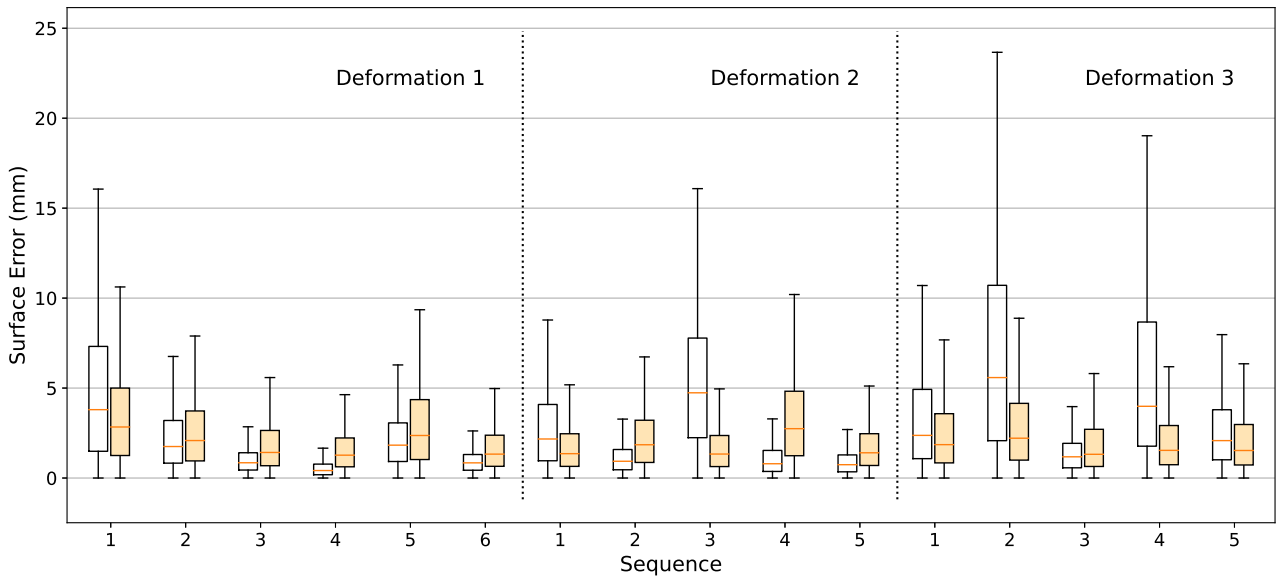

Figure 6: Quantitative evaluation of the surface reconstruction error for each sequence when compared to surface models extracted from corresponding CT scans: ElasticFusion (white); ORB-SLAM3 (orange).

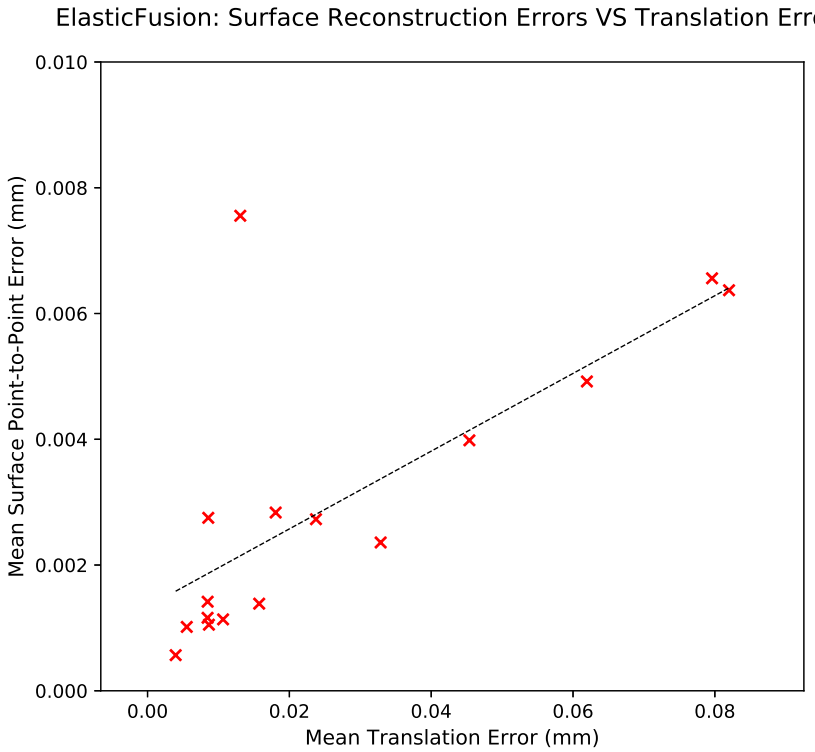

Figure 7: Translation Error VS Surface Error for ElasticFusion

all 16 sequences for both algorithms, a Wilcoxon Signed-Rank test between these found no statistically significant difference (p=0.501).

Figure 7 shows mean surface reconstruction error plotted against mean translation error on all 16 sequences for ElasticFusion. The figure suggests a linear relationship between the two metrics, and a Pearson's correlation test finds that they are strongly correlated with a value of 0.726 (p=1.47e-03). This implies that, by finding a way to improve tracking for ElasticFusion, it is likely possible to also improve its surface reconstruction capabilities. However, the same analysis does not find this to be the case for ORB-SLAM3 (p=0.739), where it may be the case that trajectory estimation is not the limiting factor in surface reconstruction performance.

3.4 Qualitative Evaluation

In the first sequence, ElasticFusion began by tracking while the camera was still. However, as soon as the camera was translated to the right, ElasticFusion determined itself to have lost tracking. The system then appeared to continue to follow the camera pose well, but never relocalised itself when revisiting the previously mapped area. Ultimately, the system failed to track entirely.

ElasticFusion began by steadily tracking the second sequence. However, it soon reached a point where it made an error in estimating the transform from one pose to the next, causing it to deviate significantly from the correct solution. This failure may again have been due to ambiguity in the solution to the geometric-photometric optimisation procedure, or errors in the generated map which impacted frame-to-model tracking. As the algorithm did not pick up on this tracking failure, the map-generation did not cease, and a loop was never closed when it returned to previously seen areas. The resulting map was very disjoint. It should be noted that the parameters specified for ElasticFusion through the parameter sweep were for a greater resolution, and this may in part have been responsible for its worse performance here.

In the case of ORB-SLAM3, it successfully tracked for the entirety of the first sequence, apparently experiencing no issues. It is interesting to note that, from around 15cm away, almost no features were picked up on

the liver by ORB-SLAM3. This is likely due to the fact that the liver itself is dark and its features are low in contrast.

For the second trajectory, ORB-SLAM3 successfully tracked the majority of the sequence. When the camera was translated near to a surface, however, the system lost track. Similarly, when the laparoscope was translated quickly, the system again lost track. These are both difficult cases for the algorithms to deal with, and a loss of tracking is to be expected given that the observed scene changes rapidly. In either case, the system initialised a new map to track and, in a couple of cases, when returning to previously tracked areas, ORB-SLAM3 successfully merged maps. As a result, a more comprehensive map was produced. This property of ORB-SLAM3's, multi-map generation and merging, is very useful for maximising the quantity of mapping done in difficult-to-track sequences. During surgery, where surgeons move an endoscope in ways which are not ideal for SLAM to track, ORB-SLAM3's multi-map method would enable the surgeon to use such a system much more naturally than if they had to return to previously mapped areas every time tracking was lost.

4. CONCLUSIONS

We find in this work that while ElasticFusion produces accurate and consistent maps of a human phantom, its ability to do so is highly dependent on its success in tracking the current sequence. It additionally struggles to detect previously visited areas and also to detect that it has lost tracking. The trajectory estimation procedure of ORB-SLAM3, based on salient features points, is more robust than the frame-to-model geometric-photometric method of ElasticFusion. In the context of laparoscopic liver surgery, the more erroneous camera pose estimation of ElasticFusion is likely to hamper the usefulness of a CAS system which renders important structures into the laparoscope view. Conversely, the very reliable tracking of ORB-SLAM3 and its ability to consistently detect previously visited areas, consequently enable it to merge maps and maximise the use of incoming data. This behaviour is very useful during surgery where movement patterns can be difficult to track. The dense output that can be obtained is competitive with that of ElasticFusion, but in its current implementation cannot update on loop-closure: which, alongside noise-robustness, is a property of ElasticFusion's surface reconstruction that could be benefited from.

The data used in this study contained many challenging aspects inherent to endoscopic data, such as: lighting and contrast changes, small field-of-view, blurriness and endoscopic lens distortion. It did not, however, incorporate larger real-time deformations which occur when surgeons manipulate an organ. With recent work being done on non-rigid SLAM systems,[16,23] these situations should be explored further in future work.

APPENDIX A. IMPLEMENTATION DETAILS

Here we include the parameter values used in, and resulting from, the parameters sweep procedure described in Section 2.4.

A.1 ElasticFusion

Parameters in Tables 3 and 4:

A) Surfel Confidence Threshold

B) Relative Iterative Closest Point (ICP) vs RGB weight

C) Local Loop Closure Residual Threshold

D) Local Loop Closure Inlier Threshold

E) Local Loop Closure Covariance Threshold

Table 3: Parameter values used in the first parameter sweep for ElasticFusion. Default parameters are in italics and the parameters determined to be best are in bold.

	A	B	C	D	E
Value 1	*10*	**8**	4e-05	*35000*	8e-06
Value 2	12	*10*	*5e-05*	47000	*1e-05*
Value 3	**14**	12	**6e-05**	**59000**	**1.2e-05**

Table 4: Parameter values used in the second parameter sweep for ElasticFusion. Parameters determined to be best are in bold.

	A	B	C	D	E
Value 1	13	**7**	5.5e-05	53000	**1.1e-05**
Value 2	**14**	8	**6e-05**	59000	1.2e-05
Value 3	15	9	6.5e-05	**65000**	1.3e-05

A.2 ORB-SLAM3

Parameters in Tables 5 and 6:

A) Number of ORB features extracted per image

B) Scale factor between levels in ORB scale pyramid

C) Number of levels in ORB scale pyramid

D) Initial response threshold for FAST detector

E) Minimum response threshold for FAST detector

Table 5: Parameter values used in the first parameter sweep for ORB-SLAM3. Default parameters are in italics and the parameters determined to be best are in bold.

	A	B	C	D	E
Value 1	**1600**	1.12	*8*	16	**5**
Value 2	*2000*	1.16	10	***20***	*7*
Value 3	2400	***1.20***	**12**	24	9

Table 6: Parameter values used in the second parameter sweep for ORB-SLAM3. Parameters determined to be best are in bold.

	A	B	C	D	E
Value 1	1400	1.18	11	18	4
Value 2	1600	**1.20**	**12**	**20**	**5**
Value 3	**1800**	1.22	13	22	6

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding for this reasearch by the State of Saxony via Sächsische Aufbaubank (SAB) in the scope of the ARAILIS project (100400076). This measure is co-financed with tax funds on the basis of the budget passed by the Saxon state parliament.

REFERENCES

[1] Mur-Artal, R., Montiel, J. M. M., and Tardos, J. D., "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics* **31**(5), 1147–1163 (2015).

[2] Mur-Artal, R. and Tardós, J. D., "Orb-slam2: an open-source slam system for monocular, stereo and rgb-d cameras," *IEEE Transactions on Robotics* **33**(5), 1255–1262 (2017).

[3] Campos, C., Elvira, R., Rodríguez, J. J. G., Montiel, J. M., and Tardós, J. D., "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics* **37**(6), 1874–1890 (2021).

[4] Whelan, T., Leutenegger, S., Salas-Moreno, R., Glocker, B., and Davison, A., "Elasticfusion: Dense slam without a pose graph," Robotics: Science and Systems (2015).

[5] Fretland, A., Aghayan, D., and Edwin, B., "Long-term survival after laparoscopic versus open resection for colorectal liver metastases," *Journal of Clinical Oncology* **37**(18_suppl), LBA3516–LBA3516 (2019).

[6] Palomar, R., Cheikh, F. A., Edwin, B., Beghdadhi, A., and Elle, O. J., "Surface reconstruction for planning and navigation of liver resections," *Computerized Medical Imaging and Graphics* **53**, 30 – 42 (2016).

[7] Pfeiffer, M., Riediger, C., Weitz, J., and Speidel, S., "Learning soft tissue behavior of organs for surgical navigation with convolutional neural networks," *International journal of computer assisted radiology and surgery* **14**(7), 1147–1155 (2019).

[8] Teatini, A., Pelanis, E., Aghayan, D. L., Kumar, R. P., Palomar, R., Fretland, Å. A., Edwin, B., and Elle, O. J., "The effect of intraoperative imaging on surgical navigation for laparoscopic liver resection surgery," *Scientific Reports* **9**(1), 1–11 (2019).

[9] Thompson, S., Schneider, C., Bosi, M., Gurusamy, K., Ourselin, S., Davidson, B., Hawkes, D., and Clarkson, M. J., "In vivo estimation of target registration errors during augmented reality laparoscopic surgery," *International journal of computer assisted radiology and surgery* **13**(6), 865–874 (2018).

[10] Fuentes-Pacheco, J., Ruiz-Ascencio, J., and Rendón-Mancha, J. M., "Visual simultaneous localization and mapping: a survey," *Artificial intelligence review* **43**(1), 55–81 (2015).

[11] Kerl, C., Sturm, J., and Cremers, D., "Dense visual slam for rgb-d cameras," in [*2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*], 2100–2106, IEEE (2013).

[12] Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., and Fitzgibbon, A., "Kinectfusion: Real-time dense surface mapping and tracking," in [*2011 10th IEEE International Symposium on Mixed and Augmented Reality*], 127–136, IEEE (2011).

[13] Whelan, T., Kaess, M., Johannsson, H., Fallon, M., Leonard, J. J., and McDonald, J., "Real-time large-scale dense rgb-d slam with volumetric fusion," *The International Journal of Robotics Research* **34**(4-5), 598–626 (2015).

[14] Grasa, O. G., Civera, J., and Montiel, J., "Ekf monocular slam with relocalization for laparoscopic sequences," in [*2011 IEEE International Conference on Robotics and Automation*], 4816–4821, IEEE (2011).

[15] Mahmoud, N., Cirauqui, I., Hostettler, A., Doignon, C., Soler, L., Marescaux, J., and Montiel, J., "Orbslam-based endoscope tracking and 3d reconstruction," in [*International workshop on computer-assisted and robotic endoscopy*], 72–83, Springer (2016).

[16] Song, J., Wang, J., Zhao, L., Huang, S., and Dissanayake, G., "Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing," *IEEE Robotics and Automation Letters* **3**(4), 4068–4075 (2018).

[17] Yang, G., Manela, J., Happold, M., and Ramanan, D., "Hierarchical deep stereo matching on high-resolution images," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 5515–5524 (2019).

[18] Keller, M., Lefloch, D., Lambers, M., Izadi, S., Weyrich, T., and Kolb, A., "Real-time 3d reconstruction in dynamic scenes using point-based fusion," in [*2013 International Conference on 3D Vision-3DV 2013*], 1–8, IEEE (2013).

[19] Glocker, B., Shotton, J., Criminisi, A., and Izadi, S., "Real-time rgb-d camera relocalization via randomized ferns for keyframe encoding," *IEEE transactions on visualization and computer graphics* **21**(5), 571–583 (2014).

[20] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G., "Orb: An efficient alternative to sift or surf," in [*2011 International conference on computer vision*], 2564–2571, Ieee (2011).

[21] Kenngott, H., Wünscher, J., Wagner, M., Preukschas, A., Wekerle, A., Neher, P., Suwelack, S., Speidel, S., Nickel, F., Oladokun, D., et al., "Openhelp (heidelberg laparoscopy phantom): development of an open-source surgical evaluation and training tool," *Surgical endoscopy* **29**(11), 3338–3347 (2015).

[22] Shah, M., Eastman, R. D., and Hong, T., "An overview of robot-sensor calibration methods for evaluation of perception systems," in [*Proceedings of the Workshop on Performance Metrics for Intelligent Systems*], 15–20 (2012).

[23] Lamarca, J., Parashar, S., Bartoli, A., and Montiel, J., "Defslam: Tracking and mapping of deforming scenes from monocular sequences," *IEEE Transactions on Robotics* (2020).