Neural Fields for 3D Tracking of Anatomy and Surgical Instruments in Monocular Laparoscopic Video Clips

Beerend G.A. Gerats[1,2] (✉), Jelmer M. Wolterink[3,4], Seb P. Mol[1,4], and Ivo A.M.J. Broeders[1,2]

[1] AI & Data Science Center, Meander Medical Center, The Netherlands
bga.gerats@meandermc.nl
[2] Robotics and Mechatronics, University of Twente, The Netherlands
[3] Department of Applied Mathematics, University of Twente, The Netherlands
[4] Technical Medical Center, University of Twente, The Netherlands

Abstract. Laparoscopic video tracking primarily focuses on two target types: surgical instruments and anatomy. The former could be used for skill assessment, while the latter is necessary for the projection of virtual overlays. Where instrument and anatomy tracking have often been considered two separate problems, in this paper, we propose a method for joint tracking of all structures simultaneously. Based on a single 2D monocular video clip, we train a neural field to represent a continuous spatiotemporal scene, used to create 3D tracks of all surfaces visible in at least one frame. Due to the small size of instruments, they generally cover a small part of the image only, resulting in decreased tracking accuracy. Therefore, we propose enhanced class weighting to improve the instrument tracks. We evaluate tracking on video clips from laparoscopic cholecystectomies, where we find mean tracking accuracies of 92.4% for anatomical structures and 87.4% for instruments. Additionally, we assess the quality of depth maps obtained from the method's scene reconstructions. We show that these pseudo-depths have comparable quality to a state-of-the-art pre-trained depth estimator. On laparoscopic videos in the SCARED dataset, the method predicts depth with an MAE of 2.9 mm and a relative error of 9.2%. These results show the feasibility of using neural fields for monocular 3D reconstruction of laparoscopic scenes.

Keywords: Neural Fields · Scene Reconstruction · Surgical Instrument Tracking · Tissue Deformation · Laparoscopic Videos.

1 Introduction

Optical tracking of elements in laparoscopic videos provides valuable information for computer-assisted surgery. Tracking efforts primarily focus on two target types: surgical instruments and anatomy. E.g., the former is used for skill assessment [6,12], while the latter is necessary for the projection of virtual overlays [17].

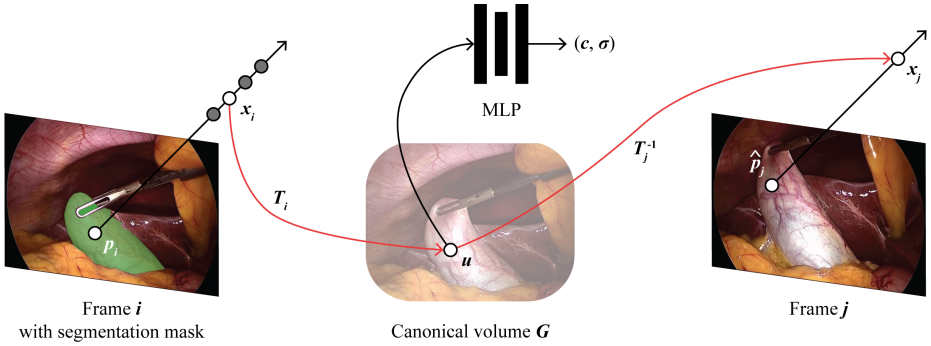

Fig. 1. Overview of our method. A pixel p_i is selected in frame i, from where a ray is cast through the virtual scene. A sampled location x_i is mapped to location u in canonical volume G. This location is used to predict color c and material density σ. To predict the translation of this point to frame j, the location is mapped via an inverse transform and reprojected to the predicted pixel location $\hat{p}_j$.

Several 2D instrument trackers have been proposed [5,11,20], but these trackers generally lack depth perception and fail to track points adequately in 3D. This might be overcome by the inclusion of markers [3] or electromagnetic tracking [14]. However, such approaches require adaptations to the instruments or surgical workflow and are not widely available. 3D pose trackers designed for correcting position errors, e.g. in robotic systems, require a known pose at the start of tracking [4]. Other methods use camera parameters to calculate 3D instrument locations [28], which are typically unknown for regular surgery. In anatomical tracking, some approaches use an additional depth sensor [7]. However, these are difficult to implement in-vivo. Alternatively, depth could be estimated from stereoscopic [15] or monocular video [2,21]. Although the depth estimates are highly informative, these methods do not model tissue deformation. Methods for 3D scene reconstruction based on simultaneous localization and mapping, do model for anatomical deformations [13,15,22], but are limited in handling complex deformations and color changes.

While instrument and anatomy tracking have often been considered two separate problems, here, we propose a method for joint tracking of *all* elements in a laparoscopic video. Based on a 2D video clip, we track the 3D location of all surfaces that are visible in at least one frame of the video. Key to our approach is the use of neural fields, which represent continuous spatio-(temporal) scenes in the weights of neural networks. Neural fields have previously been used in laparoscopic scene representation [25,26,27]. A limitation of these prior works is the need for stereoscopic vision, which limits the applicability of these methods in standard minimally invasive procedures. Instead, we propose to leverage the recently proposed OmniMotion [24] approach, in which monocular videos are used to reconstruct a 3D + time scene representation, without the need for camera calibration, stereoscopic vision, or depth maps. We evaluate the tracking

performance on video clips from laparoscopic cholecystectomies for various tissue and instrument types. We show that instrument tracks can be obtained using enhanced class weighting. Last, we assess the quality of the depth maps obtained from the method's scene reconstructions. We show that these pseudo-depths have comparable quality to a state-of-the-art pre-trained depth estimator.

2 Methods

We propose to use neural fields as representations of surgical scenes visualized in monocular laparoscopic video clips. The neural fields are optimized using OmniMotion [24], with additional class weighting to improve surgical instrument tracking. An overview of the method is given in Figure 1.

2.1 Neural Fields for Pixel Tracking

We use OmniMotion [24] for the reconstruction of 3D + time scene representations from 2D videos. The method captures camera movement and scene deformations in monocular video by mapping through a canonical volume G. Camera rays are cast through the scene at frame i, along which 3D locations x_i are sampled. Using bijective mapping $u = T_i(x_i)$, the 3D locations are translated to location u in the canonical volume. The mappings are invertible such that the new position of this point in frame j can be calculated with $x_j = T_j^{-1}(T_i(x_i))$. A canonical location u is used to query a multi-layer perception (MLP) that returns material density σ and color c for this point. When all sampled points along a camera ray are mapped to canonical space and queried to the MLP to obtain their densities and colors, the quadrature rule is used to find the predicted color $\hat{C}_i$ for the corresponding pixel p_i [16]. The predicted flow of a pixel between two frames $\hat{f}_{i\to j} = \hat{p}_j - p_i$ can be calculated similarly, using the material density values of the sampled points along the ray. In this way, the model can account for occlusions. The predicted flows are supervised with flows from a pre-trained optical flow model (RAFT) [23] (Equation 1), while the predicted pixel colors are supervised with the training images (Equation 2).

$$L_{\text{flow}} = \sum_{f_{i\to j}\in\Omega_p} \|\hat{f}_{i\to j} - f_{i\to j}\|_1 \tag{1}$$

$$L_{\text{color}} = \sum_{(i,p)\in\Omega_p} \|\hat{C}_i(p) - C_i(p)\|_2^2 \tag{2}$$

Although the method is supervised with 2D flows, the method's representation of scene deformation is in 3D. This makes it possible to track locations on anatomical structures and to follow surgical instruments over the depth axis. Because of the flow regularization, the method does not require camera parameters, stereoscopic video, or pre-computed depth images. Additionally, as camera movement is compensated for implicitly by the method, camera locations are not needed as well. This makes the method applicable to video clips from a regular laparoscope. Note that for each scene, we optimize an individual model.

2.2 Segmentation Masks

Since OmniMotion simultaneously tracks all pixels in the video, it is not possible to distinguish instruments from anatomy pixels. Therefore, we propose to combine the method with segmentation masks that cover the elements of interest. The segmentation masks can be manually annotated or found by pre-trained segmentation models [8]. Note that the masks are not necessary for training and can be applied afterward unless class weighting is used (see next section). Only a single mask at the start of the video is required to track these pixels throughout the rest of the video. An example mask can be seen in Figure 1.

2.3 Class Weighting

One of our goals is to track surgical instruments through the video. However, due to their size, instruments generally cover a small part of the image only. In the regular training approach of neural radiance fields, there is no incentive to represent a pixel more accurately than others. This means that reconstructions of irrelevant background pixels are reconstructed with similar importance as the surgical instruments. Therefore, we propose to use weighting in the loss function, where the weight depends on the class of the pixel, i.e. "instrument" or "background":

$$L = w_{\text{class}}(L_{\text{flow}} + \lambda L_{\text{color}}) + L_{\text{other}} \tag{3}$$

where $w_{\text{class}} = 1$ if the pixel is in the background and a larger value otherwise, λ is a loss weighting factor, and L_{other} is a combination of regularizing losses. The pixel classes could be obtained by manual or automatic segmentation masks.

2.4 Data

We evaluate our model on two tasks. First, we assess the method's performance in 2D tracking of anatomical structures and surgical instruments. We train OmniMotion on all 80-frame video clips in the CholecSeg8k dataset [9]. The set includes 101 short videos of 25fps that display laparoscopic cholecystectomies. Each video frame is accompanied by manually annotated segmentation masks of tissues and instruments. We pre-process the videos by cropping to a rectangle, to remove the circular border of the scope as much as possible, and resizing to 256×256 pixels. After pre-processing, we find 84 clips that contain a segmentation in the first frame for the "gallbladder" class. For "liver" and "gastrointestinal tract" we find 101 and 46 clips, respectively. For the instruments "grasper" and "l-hook electrocautery" there are 65 and 24 clips available, respectively. In total, we track 231 anatomical structures and 89 surgical instruments.

Second, we explore the accuracy of the third axis, by comparing predicted depths with ground-truth depth measurements. For the evaluation of the depth predictions, we use laparoscopic videos of the SCARED dataset [1]. The videos display a static surgical scene, recorded by various moving cameras. We select

Fig. 2. Examples of tracked pixels (top: gallbladder, bottom: grasper) through 80-frame video clips, with the first, middle and last frame displayed here. The full videos are available via: https://vimeo.com/920225544.

the first camera of each scene, as we evaluate the method for the monocular setting. The videos in the dataset are attributed with ground-truth depth values, measured via a structured light pattern projected on the field of view. We train OmniMotion on all nine videos individually, where the video length ranges between 88 and 728 frames.

2.5 Evaluation Metrics

For 2D tracking, we calculate the mean accuracy of pixels in a segmentation mask that are tracked throughout the video. For each pixel p_1 in the segmentation mask of the first frame, we get a predicted pixel location $\hat{p}_j$ at consecutive frames $j \in [2, 3, ..., 80]$. Predictions that are located within the segmentation mask of that frame are considered accurate, while those that fall outside the mask are inaccurate. Predicted locations that are outside the image boundaries are excluded from evaluation. In this way, we calculate the per-frame accuracy. When no segmentation mask is available, e.g. when an instrument is temporarily out of view, the frame is discarded. Finally, we calculate the mean accuracy by averaging the per-frame accuracies over all frames first, and over all videos afterwards.

For the evaluation of the predicted depth maps, we follow the general approach for evaluating pseudo-depths [19], meaning that we shift and scale the depth predictions towards the ground-truth values before calculating the metrics.

3 Experiments and Results

We train the method in batches of 256 correspondences from 8 image pairs. Along each camera ray, 32 points are sampled. Learning rates 3×10^{-4} and 1×10^{-4} are used for color and flow loss, respectively. Using an NVIDIA A40 GPU, it takes $\pm$10 minutes for RAFT pre-processing an 80-frame video clip. Training OmniMotion for 10K iterations requires an additional $\pm$45 minutes.

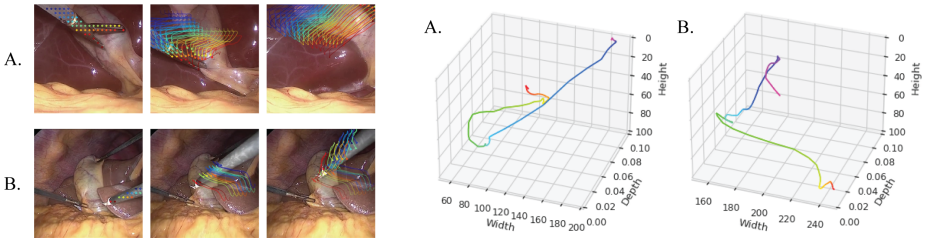

Fig. 3. Tracks of surgical instruments. Left: first, middle and last frame of the video, with trails following the instrument. Right: instrument tracks visualized in 3D.

3.1 Evaluation of Anatomy and Instrument Tracking

A selection of qualitative results is given in Figure 2. The top row displays two videos where gallbladder pixels are tracked. In the left example, the tracks accurately follow the gallbladder deformation under retraction with a grasper. In the right example, the camera zooms out while the gallbladder is released from the grasper grip. The pixel tracks accurately follow the anatomical structure. In the bottom row, the grasper is tracked. Left, the instrument leaves the field of view during the middle section of the video. It can be seen that the method accurately re-identifies the grasper when re-entering the field of view. In the right example, the grasper is heavily displaced to the upper-right corner in the last frame, where only a small part of the instrument remains available. Also in this scenario, the tracks follow the target pixels accurately. Instrument movements are visualized in Figure 3 for two example clips. In the 3D plots, we show the median location of the instrument over time. It can be seen that the method captures movements over three axes.

Table 1 provides the quantitative measurements in tracking performance of anatomies and instruments. With a mean accuracy of 92.4%, the method is able to track a large portion of the anatomy pixels accurately. Although most instrument pixels are properly tracked as well, the tracking performance is lower. The drop in performance can be explained by the small image area that the instru-

Table 1. Mean accuracy (%) of 2D tracking of anatomical structures and surgical instruments in CholecSeg8k. Instruments are tracked with different class weights w_{class}. Note that w_{class}=1 means that no class weighting is used.

	w_{class}=1	w_{class}=5	w_{class}=10	w_{class}=20
Gallbladder	91.7 (± 11.8)	-	-	-
Liver	93.6 (± 8.7)	-	-	-
Gastroint. tract	90.9 (± 15.0)	-	-	-
Grasper	83.6 (± 23.0)	84.3 (± 21.3)	85.1 (± 19.8)	85.9 (± 20.0)
L-hook	83.6 (± 28.8)	92.0 (± 21.7)	92.5 (± 20.6)	91.5 (± 20.3)
All anatomies	92.4 (± 11.4)	-	-	-
All instruments	83.6 (± 24.7)	86.3 (± 21.7)	87.1 (± 20.3)	87.4 (± 20.3)

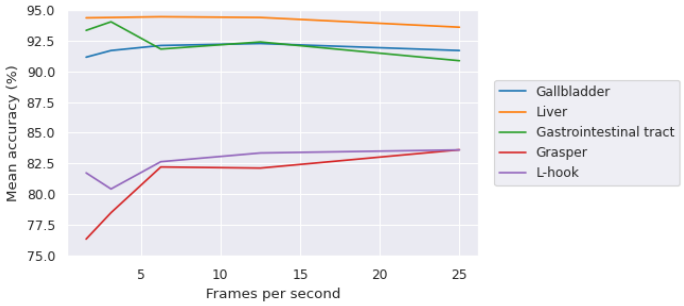

Fig. 4. Performance in 2D tracking for various temporal resolutions.

ments typically cover. The effect can be counterbalanced with class weighting, where instrument tracking performance gradually increases from 83.6% to 87.4% for larger weights.

Tracking accuracy for different frame rates is given in Figure 4. Here, we decrease the temporal resolution of 25fps gradually to 50, 25, 12.5, and 6.25 percent of the original. Where the tracking performance for the anatomical structures remains largely unaffected, the quality of the instrument tracks decreases with fewer frames per second. This effect could be explained by the more drastic movements that instruments generally make in comparison to the anatomy.

3.2 Evaluation of Reconstructed Pseudo-Depth

To assess the quality of the third pixel-tracking axis, we compare the estimated depth values with ground-truth measurements. Table 2 shows the error in depth estimation for all nine videos in the SCARED dataset [1]. On average, the estimates have a mean average error of 2.9 mm and a relative error of 9.2% to the

Table 2. Error of pseudo-depth estimation on surgical scenes in SCARED dataset, in comparison with a pre-trained dense depth estimator (DPT) [18].

Scene	MAE (mm) ↓		AbsRel (%)↓		$\delta < 1.25$ (%)↑	
no.	OmniMotion	DPT	OmniMotion	DPT	OmniMotion	DPT
1	2.9	2.2	7.8	6.2	95.5	99.0
2	1.4	1.5	4.8	4.9	97.8	99.8
3	5.3	5.0	14.6	13.0	77.0	85.4
4	1.7	2.0	7.4	8.1	97.1	96.3
5	2.1	4.2	7.8	17.3	95.5	77.8
6	2.5	2.2	7.2	6.4	96.2	98.8
7	5.1	4.5	13.9	13.0	80.0	84.6
8	2.2	2.1	7.5	7.3	96.7	96.0
9	2.7	1.6	12.0	6.8	90.3	96.5
Average	2.9 (± 1.3)	2.8 (± 1.3)	9.2 (± 3.2)	9.2 (± 4.0)	91.8 (± 7.4)	92.7 (± 7.5)

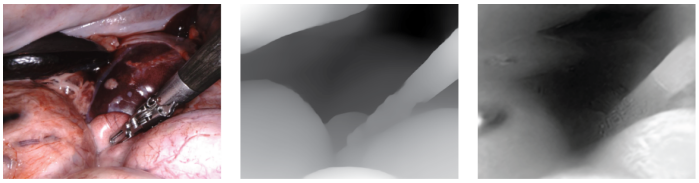

Fig. 5. An example frame from the SCARED dataset, with disparity estimated by DPT [18] and reconstructed with OmniMotion.

actual depth. This performance is comparable to depth estimates by DPT [18], a state-of-the-art pre-trained transformer for monocular depth estimation. The comparison shows that the depth axis of the pixel tracks is relatively accurate as well. Figure 5 shows an example frame, with depth estimates by DPT and OmniMotion. Although the method's depth values are less smooth, they represent an accurate geometry of the underlying scene.

4 Discussion & Conclusion

We combined tissue deformation modeling and surgical instrument tracking by tracking all pixels in a laparoscopic video at once. We used neural fields for 3D scene reconstruction, which could be trained without the need for camera calibration, camera locations, stereoscopic vision, or depth maps. In contrast to previously proposed methods for 3D scene reconstructions, this makes the method applicable as-is to standard minimally invasive procedures. In our experiments, we show that the combination of OmniMotion [24] and segmentation masks can be used for constructing accurate 3D anatomy and instrument tracks. The quality of the instrument tracks can be improved by using class weighting, where instrument pixels are weighted more heavily during optimization.

A limitation of the used method is the relatively slow training speed. Ideally, the scene reconstructions are built in real-time, such that the resulting tracks are available during surgery. Yang et al. [26] have shown that drastic speed-ups, over a 100× faster, are possible by leveraging multi-plane fields. Moreover, the use of Gaussian splatting [10] could potentially result in even faster training and real-time rendering.

A direction for future work is a deeper exploration of uses for neural fields from surgical videos. For example, when scene reconstructions are accurate enough, anatomical structures in pre-operative scans could potentially be mapped to the neural representation. Because neural fields can model tissue deformation, virtual overlays could be rendered that follow the underlying structures. Another example is the use of 3D anatomical tracks for the calculation of stress applied to the tissue, which could be visualized or used for haptic feedback. In our work, we have shown the feasibility of using neural fields for 3D monocular scene reconstructions, which could form the basis for these further explorations.

Declarations

Research at the Meander AI & Data Science Center was sponsored by Johnson & Johnson MedTech. Jelmer M. Wolterink was supported by NWO domain Applied and Engineering Sciences VENI grant (18192).

References

1. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)
2. Beilei, C., Mobarakol, I., Long, B., Hongliang, R.: Surgical-dino: Adapter learning of foundation model for depth estimation in endoscopic surgery. arXiv preprint arXiv:2401.06013 (2024)
3. Cartucho, J., Wang, C., Huang, B., S. Elson, D., Darzi, A., Giannarou, S.: An enhanced marker pattern that achieves improved accuracy in surgical tool tracking. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **10**(4), 400–408 (2022)
4. Du, X., Allan, M., Dore, A., Ourselin, S., Hawkes, D., Kelly, J.D., Stoyanov, D.: Combined 2d and 3d tracking of surgical instruments for minimally invasive and robotic-assisted surgery. International journal of computer assisted radiology and surgery **11**, 1109–1119 (2016)
5. Du, X., Kurmann, T., Chang, P.L., Allan, M., Ourselin, S., Sznitman, R., Kelly, J.D., Stoyanov, D.: Articulated multi-instrument 2-d pose estimation using fully convolutional networks. IEEE transactions on medical imaging **37**(5), 1276–1287 (2018)
6. Fathollahi, M., Sarhan, M.H., Pena, R., DiMonte, L., Gupta, A., Ataliwala, A., Barker, J.: Video-based surgical skills assessment using long term tool tracking. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 541–550. Springer (2022)
7. Golse, N., Petit, A., Lewin, M., Vibert, E., Cotin, S.: Augmented reality during open liver surgery using a markerless non-rigid registration system. Journal of Gastrointestinal Surgery **25**, 662–671 (2021)
8. Grammatikopoulou, M., Sanchez-Matilla, R., Bragman, F., Owen, D., Culshaw, L., Kerr, K., Stoyanov, D., Luengo, I.: A spatio-temporal network for video semantic segmentation in surgical videos. International Journal of Computer Assisted Radiology and Surgery **19**(2), 375–382 (2024)
9. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)
10. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
11. Kurmann, T., Marquez Neila, P., Du, X., Fua, P., Stoyanov, D., Wolf, S., Sznitman, R.: Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20. pp. 505–513. Springer (2017)
12. Lavanchy, J.L., Zindel, J., Kirtac, K., Twick, I., Hosgor, E., Candinas, D., Beldi, G.: Automation of surgical skill assessment using a three-stage machine learning algorithm. Scientific reports **11**(1), 5197 (2021)

13. Li, Y., Richter, F., Lu, J., Funk, E.K., Orosco, R.K., Zhu, J., Yip, M.C.: Super: A surgical perception framework for endoscopic tissue manipulation with surgical robotics. IEEE Robotics and Automation Letters **5**(2), 2294–2301 (2020)
14. Liu, X., Kang, S., Plishker, W., Zaki, G., Kane, T.D., Shekhar, R.: Laparoscopic stereoscopic augmented reality: toward a clinically viable electromagnetic tracking solution. Journal of Medical Imaging **3**(4), 045001 (2016)
15. Long, Y., Li, Z., Yee, C.H., Ng, C.F., Taylor, R.H., Unberath, M., Dou, Q.: E-dssr: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 415–425. Springer (2021)
16. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
17. Pelanis, E., Teatini, A., Eigl, B., Regensburger, A., Alzaga, A., Kumar, R.P., Rudolph, T., Aghayan, D.L., Riediger, C., Kvarnström, N., et al.: Evaluation of a novel navigation platform for laparoscopic liver surgery with organ deformation compensation using injected fiducials. Medical image analysis **69**, 101946 (2021)
18. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
19. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)
20. Robu, M., Kadkhodamohammadi, A., Luengo, I., Stoyanov, D.: Towards real-time multiple surgical tool tracking. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **9**(3), 279–285 (2021)
21. Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., Zhang, B.: Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. Medical image analysis **77**, 102338 (2022)
22. Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G.: Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. IEEE Robotics and Automation Letters **3**(1), 155–162 (2017)
23. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020)
24. Wang, Q., Chang, Y.Y., Cai, R., Li, Z., Hariharan, B., Holynski, A., Snavely, N.: Tracking everything everywhere all at once. In: International Conference on Computer Vision (2023)
25. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 431–441. Springer (2022)
26. Yang, C., Wang, K., Wang, Y., Yang, X., Shen, W.: Neural lerplane representations for fast 4d reconstruction of deformable tissues. arXiv preprint arXiv:2305.19906 (2023)
27. Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z.: Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In: International

Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–23. Springer (2023)
28. Zhao, Z., Voros, S., Weng, Y., Chang, F., Li, R.: Tracking-by-detection of surgical instruments in minimally invasive surgery via the convolutional neural network deep learning-based method. Computer Assisted Surgery **22**(sup1), 26–35 (2017)