Detecting the Sensing Area of a Laparoscopic Probe in Minimally Invasive Cancer Surgery

Baoru Huang[1,2(✉)], Yicheng Hu[1,2], Anh Nguyen[3], Stamatia Giannarou[1,2], and Daniel S. Elson[1,2]

[1] The Hamlyn Centre for Robotic Surgery, Imperial College London, London, UK
Baoru.Huang18@imperial.ac.uk
[2] Department of Surgery & Cancer, Imperial College London, London, UK
[3] Department of Computer Science, University of Liverpool, Liverpool, UK

Abstract. In surgical oncology, it is challenging for surgeons to identify lymph nodes and completely resect cancer even with pre-operative imaging systems like PET and CT, because of the lack of reliable intraoperative visualization tools. Endoscopic radio-guided cancer detection and resection has recently been evaluated whereby a novel tethered laparoscopic gamma detector is used to localize a preoperatively injected radiotracer. This can both enhance the endoscopic imaging and complement preoperative nuclear imaging data. However, gamma activity visualization is challenging to present to the operator because the probe is non-imaging and it does not visibly indicate the activity origination on the tissue surface. Initial failed attempts used segmentation or geometric methods, but led to the discovery that it could be resolved by leveraging high-dimensional image features and probe position information. To demonstrate the effectiveness of this solution, we designed and implemented a simple regression network that successfully addressed the problem. To further validate the proposed solution, we acquired and publicly released two datasets captured using a custom-designed, portable stereo laparoscope system. Through intensive experimentation, we demonstrated that our method can successfully and effectively detect the sensing area, establishing a new performance benchmark. Code and data are available at https://github.com/br0202/Sensing_area_detection.git.

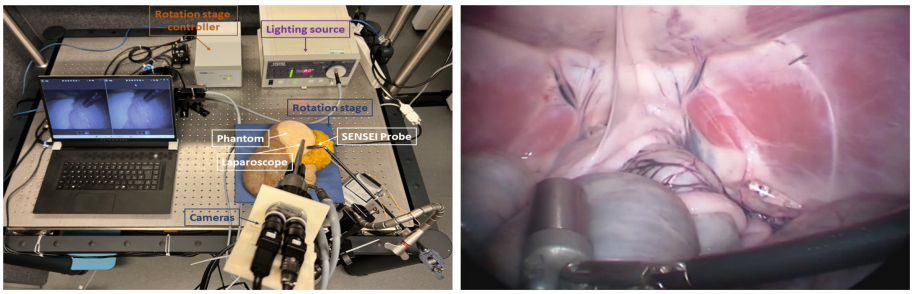
Keywords: Laparoscopic Image-guided Intervention · Minimally Invasive Surgery · Detection of Sensing Area

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_25.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 260–270, 2023.
https://doi.org/10.1007/978-3-031-43996-4_25

1 Introduction

Cancer remains a significant public health challenge worldwide, with a new diagnosis occurring every two minutes in the UK (Cancer Research UK[1]). Surgery is one of the main curative treatment options for cancer. However, despite substantial advances in pre-operative imaging such as CT, MRI, or PET/SPECT to aid diagnosis, surgeons still rely on the sense of touch and naked eye to detect cancerous tissues and disease metastases intra-operatively due to the lack of reliable intraoperative visualization tools. In practice, imprecise intraoperative cancer tissue detection and visualization results in missed cancer or the unnecessary removal of healthy tissues, which leads to increased costs and potential harm to the patient. There is a pressing need for more reliable and accurate intraoperative visualization tools for minimally invasive surgery (MIS) to improve surgical outcomes and enhance patient care.



(a) (b)

Fig. 1. (a) Hardware set-up for experiments, including a customized portable stereo laparoscope system and the 'SENSEI' probe, a rotation stage, a laparoscopic lighting source, and a phantom; (b) An example of the use of the 'SENSEI' probe in MIS.

A recent miniaturized cancer detection probe (i.e., 'SENSEI®' developed by Lightpoint Medical Ltd.) leverages the cancer-targeting ability of nuclear agents typically used in nuclear imaging to more accurately identify cancer intraoperatively from the emitted gamma signal (see Fig. 1b)[6]. However, the use of this probe presents a visualization challenge as the probe is non-imaging and is air-gapped from the tissue, making it challenging for the surgeon to locate the probe-sensing area on the tissue surface.

It is crucial to accurately determine the sensing area, with positive signal potentially indicating cancer or affected lymph nodes. Geometrically, the sensing area is defined as the intersection point between the gamma probe axis and the tissue surface in 3D space, but projected onto the 2D laparoscopic image. However, it is not trivial to determine this using traditional methods due to

[1] https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk.

poor textural definition of tissues and lack of per-pixel ground truth depth data. Similarly, it is also challenging to acquire the probe pose during the surgery.

Problem Redefinition. In this study, in order to provide sensing area visualization ground truth, we modified a non-functional 'SENSEI' probe by adding a miniaturized laser module to clearly optically indicate the sensing area on the laparoscopic images - i.e. the 'probe axis-surface intersection'. Our system consists of four main components: a customized stereo laparoscope system for capturing stereo images, a rotation stage for automatic phantom movement, a shutter for illumination control, and a DAQ-controlled switchable laser module (see Fig. 1a). With this setup, we aim to transform the sensing area localization problem from a geometrical issue to a high-level content inference problem in 2D. It is noteworthy that this remains a challenging task, as ultimately we need to infer the probe axis-surface intersection without the aid of the laser module to realistically simulate the use of the 'SENSEI' probe.

2 Related Work

Laparoscopic images play an important role in computer-assisted surgery and have been used in several problems such as object detection [9], image segmentation [23], depth estimation [20] or 3D reconstruction [13]. Recently, supervised or unsupervised depth estimation methods have been introduced [14]. Ye *et al.* [22] proposed a deep learning framework for surgical scene depth estimation in self-supervised mode and achieved scalable data acquisition by incorporating a differentiable spatial transformer and an autoencoder into their framework. A 3D displacement module was explored in [21] and 3D geometric consistency was utilized in [8] for self-supervised monocular depth estimation. Tao *et al.* [19] presented a spatiotemporal vision transformer-based method and a self-supervised generative adversarial network was introduced in [7] for depth estimation of stereo laparoscopic images. Recently, fully supervised methods were summarized in [1] for depth estimation. However, acquiring per-pixel ground truth depth data is challenging, especially for laparoscopic images, which makes it difficult for large-scale supervised training [8].

Laparoscopic segmentation is another important task in computer-assisted surgery as it allows for accurate and efficient identification of instrument position, anatomical structures, and pathological tissue. For instance, a unified framework for depth estimation and surgical tool segmentation in laparoscopic images was proposed in [5], with simultaneous depth estimation and segmentation map generation. In [12], self-supervised depth estimation was utilized to regularize the semantic segmentation in knee arthroscopy. Marullo *et al.* [16] introduced a multi-task convolutional neural network for event detection and semantic segmentation in laparoscopic surgery. The dual swin transformer U-Net was proposed in [11] to enhance the medical image segmentation performance, which leveraged the hierarchical swin transformer into both the encoder and the decoder of the standard U-shaped architecture, benefiting from the self-attention computation in swin transformer as well as the dual-scale encoding design.

Although the intermediate depth information was not our final aim and can be bypassed, the 3D surface information was necessary in the intersection point inference. ResNet [3] has been commonly used as the encoder to extract the image features and geometric information of the scene. In particular, in [21], concatenated stereo image pairs were used as inputs to achieve better results, and such stereo image types are also typical in robot-assisted minimally invasive surgery with stereo laparoscopes. Hence, stereo image data was also adopted in this paper.

If the problem of inferring the intersection point is treated as a geometric problem, both data collection and intra-operative registration would be difficult, which inspired us to approach this problem differently. In practice, we utilize the laser module to collect the ground truth of the intersection points when the laser is on. We note that the standard illumination image from the laparoscopic probe is also captured with the same setup when the laser module is on. Therefore, we can establish a dataset with an image pair (RGB image and laser image) that shares the same intersection point ground truth with the laser image (see Fig. 2a and Fig. 2b). The assumptions made are that the probe's 3D pose when projected into the two 2D images is the observed 2D pose, and that the intersection point is located on its axis. Hence, we input these axes to the network as another branch and randomly sampled points along them to represent the probe.

3 Dataset

To validate our proposed solution for the newly formulated problem, we acquired and publicly released two new datasets. In this section, we introduce the hardware and software design that was used to achieve our final goal, while Fig. 2 shows a sample from our dataset.

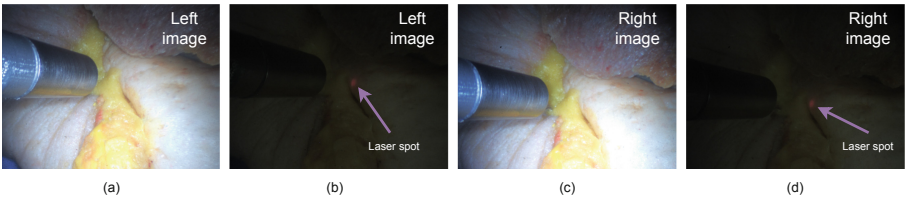


Fig. 2. Example data. (a) Standard illumination left RGB image; (b) left image with laser on and laparoscopic light off; same for (c) and (d) but for right images.

Data Collection. Two miniaturized, high-resolution cameras were coupled onto a stereo laparoscope using a custom-designed connector. The accompanying API allowed for automatic image acquisition, exposure time adjustment, and white balancing. An electrically controllable shutter was incorporated into the standard laparoscopic illumination path. To indicate the probe axis-surface intersection,

we incorporated a DAQ controlled cylindrical miniature laser module into a 'SENSEI' probe shell so that the adapted tool was visually identical to the real probe. The laser module emitted a red laser beam (wavelength 650 nm) that was visible as a red spot on the tissue surface.

We acquired the dataset on a silicone tissue phantom which was $30 \times 21 \times 8$ cm and was rendered with tissue color manually by hand to be visually realistic. The phantom was placed on a rotation stage that stepped 10 times per revolution to provide views separated by a 36-degree angle. At each position, stereo RGB images were captured *i)* under normal laparoscopic illumination with the laser off; *ii)* with the laparoscopic light blocked and the laser on; and *iii)* with the laparoscopic light blocked and the laser off. Subtraction of the images with laser on and off readily allowed segmentation of the laser area and calculation of its central point, i.e. the ground truth probe axis-surface intersection.

All data acquisition and devices were controlled by Python and LABVIEW programs, and complete data sets of the above images were collected on visually realistic phantoms for multiple probe and laparoscope positions. This provided 10 tissue surface profiles for a specific camera-probe pose, repeated for 120 different camera-probe poses, mimicking how the probe may be used in practice. Therefore, our first newly acquired dataset, named **Jerry**, contains 1200 sets of images. Since it is important to report errors in 3D and in millimeters, we recorded another dataset similar to **Jerry** but also including ground truth depth map for all frames by using structured-lighting system [8]—namely the **Coffbee** dataset.

These datasets have multiple uses such as:

- Intersection point detection: detecting intersection points is an important problem that can bring accurate surgical cancer visualization. We believe this is an under-investigated problem in surgical vision.
- Depth estimation: corresponding ground truth will be released.
- Tool segmentation: corresponding ground truth will be released.

4 Probe Axis-Surface Intersection Detection

4.1 Overview

The problem of detecting the intersection point is trivial when the laser is on and can be solved by training a deep segmentation network. However, segmentation requires images with a laser spot as input, while the real gamma probe produces no visible mark and therefore this approach produces inferior results.

An alternative approach to detect the intersection point is to reconstruct the 3D tissue surface and estimate the pose of the probe in real time. A tracking and pose estimation method for the gamma probe [6] involved attaching a dual-pattern marker to the probe to improve detection accuracy. This enabled the derivation of a 6D pose, comprising a rotation matrix and translation matrix with respect to the laparoscope camera coordinate. To obtain the intersection point, the authors used the Structure From Motion (SFM) method to compute the

3D tissue surface, combining it with the estimated pose of the probe, all within the laparoscope coordinate system. However, marker-based tracking and pose estimation methods have sterilization implications for the instrument, and the SFM method requires the surgeon to constantly move the laparoscope, reducing the practicality of these methods for surgery.

In this work, we propose a simple, yet effective regression approach to address this problem. Our approach relies solely on the 2D information and works well without the need for the laser module after training. Furthermore, this simple methodology facilitated an average inference time of 50 frames per second, enabling real-time sensing area map generation for intraoperative surgery.

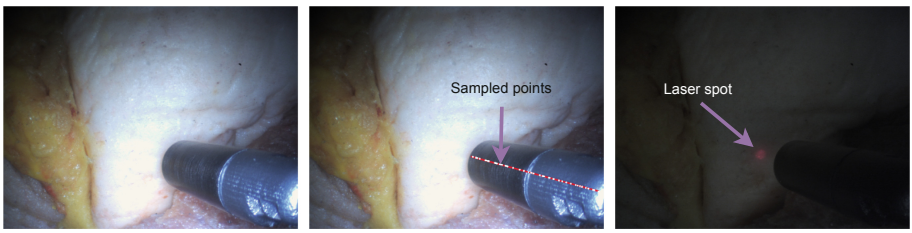


Fig. 3. Sensing area detection. (a) The input RGB image, (b) The estimated line using PCA for obtaining principal points, (c) The image with laser on that we used to detect the intersection ground truth.

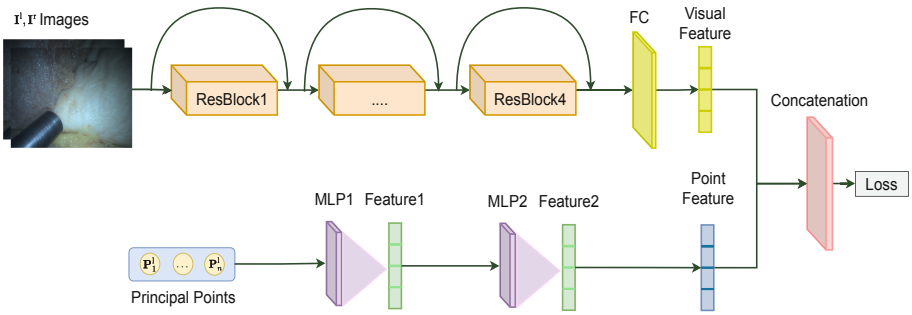


Fig. 4. An overview of our approach using ResNet and MLP.

4.2 Intersection Detection as Segmentation

We utilized different deep segmentation networks as a first attempt to address our problem [10,18]. Please refer to the Supplementary Material for the implementation details of the networks. We observed that when we do not use images with the laser, the network was not able to make any good predictions. This is

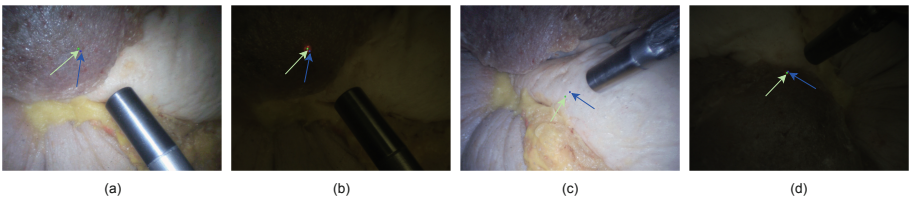


Fig. 5. Qualitative results. (a) and (c) are standard illumination images and (b) and (d) are images with laser on and laparoscopic light off. The predicted intersection point is shown in blue and the green point indicates the ground truth, which are further indicated by arrows for clarity. (Color figure online)

understandable as the red laser spot provides the key information for the segmentation. Therefore the network does not have any visual information to make predictions from images of the gamma probe. We note that to enable real-world applications, we need to estimate the intersection point using the images when the laser module is turned off.

4.3 Intersection Detection as Regression

Problem Formulation. Formally, given a pair of stereo images $\mathbf{I}^{\mathrm{l}}, \mathbf{I}^{\mathrm{r}}$, n points $\{\mathbf{P}^{\mathrm{l}}_1, \mathbf{P}^{\mathrm{l}}_2, ..., \mathbf{P}^{\mathrm{l}}_n\}$ were sampled along the principal axis of the probe, $\mathbf{P}^{\mathrm{l}}_i \in \mathbb{R}^2$ from the left image. The same process was repeated for the right image. The goal was to predict the intersection point $\mathbf{P}_{\mathrm{intersect}}$ on the surface of the tissue. During the training, the ground truth intersection point position was provided by the laser source, while during testing the intersection was estimated solely based on visual information without laser guidance (see Fig. 3).

Network Architecture. Unlike the segmentation approach, the intersection point was directly predicted using a regression network. The images fed to the network were 'laser off' stereo RGB, but crucially, the intersection point for these images was known *a priori* from the paired 'laser on' images. The raw image resolution was 4896×3680 but these were binned to 896×896. Principal Component Analysis (PCA) [15] was used to extract the central axis of the probe and 50 points were sampled along this axis as an extra input dimension. A network was designed with two branches, one branch for extracting visual features from the image and one branch for learning the features from the sequence of principal points using ResNet [3] and Vision Transformer (ViT) [2] as two backbones. The principal points were learned through a multi-layer perception (MLP) or a long short-term memory (LSTM) network [4]. The features from both branches were concatenated and used for regressing the intersection point (see Fig. 4). Finally, the whole network is trained end-to-end using the mean square error loss.

4.4 Implementation

Evaluation Metrics. To evaluate sensing area location errors, Euclidean distance was adopted to measure the error between the predicted intersection points and the ground truth laser points. We reported the mean absolute error, the standard derivation, and the median in pixel units.

Implementation Details. The networks were implemented in PyTorch [17], with an input resolution of 896×896 and a batch size of 12. We partitioned the **Jerry** dataset into three subsets, the training, validation, and test set, consisting of 800, 200, and 200 images, respectively, and the same for the **Coffbee** dataset. The learning rate was set to 10^{-5} for the first 300 epochs, then halved until epoch 400, and quartered until the end of the training. The model was trained for 700 epochs using the Adam optimizer on two NVIDIA 2080 Ti GPUs, taking approximately 4 h to complete.

Table 1. Results using ResNet50. Grey color denotes the Jerry dataset and Blue color is for Coffbee dataset (2D errors are in pixels and 3D errors are in mm).

ResNet	✓	✓	✓	✓	✓
MLP		✓			✓
LSTM			✓		
Stereo	✓	✓	✓		
Mono				✓	✓
2D Mean E.	73.5	70.5	73.7	75.6	76.7
2D Std.	65.1	56.8	62.1	62.9	64.4
2D Median	57.5	59.8	56.9	58.8	68.4
2D Mean E.	63.2	52.9	62.0	55.8	60.2
2D Std.	71.4	42.9	63.4	55.3	42.1
2D Median	44.9	44.6	43.4	42.5	52.3
R2 Score	0.55	0.82	0.63	0.73	0.78
3D Mean E.	8.5	7.4	6.5	6.4	11.2
3D Std.	15.7	6.7	6.8	7.1	18.2
3D Median	4.5	4.6	4.0	4.3	5.4

Table 2. Results using ViT. Grey color denotes the Jerry dataset and Blue color is for Coffbee dataset (2D errors are in pixels and 3D errors are in mm).

ViTNet	✓	✓	✓	✓	✓
MLP		✓			✓
LSTM			✓		
Stereo	✓	✓	✓		
Mono				✓	✓
2D Mean E.	77.9	92.3	80.9	87.7	112.1
2D Std.	69.1	71.0	67.4	68.6	84.2
2D Median	59.0	75.0	64.8	74.9	90.0
2D Mean E.	76.3	75.0	88.0	56.5	82.7
2D Std.	69.8	60.6	83.3	75.8	63.9
2D Median	59.9	59.6	68.3	34.5	69.1
R2 Score	0.58	0.66	0.33	0.65	0.60
3D Mean E.	7.9	9.1	11.4	11.6	7.7
3D Std.	6.9	8.2	16.7	21.3	7.0
3D Median	6.0	5.9	7.1	5.3	6.2

5 Results

Quantitative results on the released datasets are shown in Table 1 and Table 2 with different backbones for extracting image features, ResNet and ViT. For the 2D error on two datasets, among the different settings, the combination of ResNet and MLP gave the best performance with a mean error of 70.5 pixels

and a standard deviation of 56.8. The median error of this setting was 59.8 pixels while the R2 score was 0.82 (higher is better for R2 score). Comparing the Table 1 and Table 2, we found that the ResNet backbone was better than the ViT backbone in the image processing task, while MLP was better than LSTM in probe pose representation. ResNet processed the input images as a whole, which was better suited for utilizing the global context of a unified scene composed of the tissue and the probe, compared to the ViT scheme, which treated the whole scene as several patches. Similarly, the sampled 50 principal points on the probe axis were better processed using the simple MLP rather than using a recurrent procedure LSTM. It is worth noting that the results from stereo inputs exceeded those from mono inputs, which can be attributed to the essential 3D information included in the stereo image pairs.

For the 3D error, the ResNet backbone still gave generally better performance than the ViT backbone while under the ResNet backbone, LSTM and MLP gave competitive results and they are all in sub-milimeter level. We note that the 3D error subjected to the quality of the acquired ground truth depth maps, which had limited resolution and non-uniformly distributed valid data due to hardware constraints. Hence, we used the median depth value of a square area of 5 pixels around the points where depth value was not available.

Figure 5 shows visualization results of our method using ResNet and MLP. This figure illustrates that our proposed method successfully detected the intersection point using solely standard RGB laparoscopic images as the input. Furthermore, based on the simple design, our method achieved the inference time of 50 frames per second, making it well-suitable for intraoperative surgery.

6 Conclusion

In this work, a new framework for using a laparoscopic drop-in gamma detector in manual or robotic-assisted minimally invasive cancer surgery was presented, where a laser module mock probe was utilized to provide training guidance and the problem of detecting the probe axis-tissue intersection point was transformed to laser point position inference. Both the hardware and software design of the proposed solution were illustrated and two newly acquired datasets were publicly released. Extensive experiments were conducted on various backbones and the best results were achieved using a simple network design, enabling real time inference of the sensing area. We believe that our problem reformulation and dataset release, together with the initial experimental results, will establish a new benchmark for the surgical vision community.

References

1. Allan, M., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv:2101.01133 (2021)
2. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
5. Huang, B., et al.: Simultaneous depth estimation and surgical tool segmentation in laparoscopic images. IEEE Trans. Med. Robot. Bionics **4**(2), 335–338 (2022)
6. Huang, B., et al.: Tracking and visualization of the sensing area for a tethered laparoscopic gamma probe. Int. J. Comput. Assist. Radiol. Surg. **15**(8), 1389–1397 (2020). https://doi.org/10.1007/s11548-020-02205-z
7. Huang, B., et al.: Self-supervised generative adversarial network for depth estimation in laparoscopic images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 227–237. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_22
8. Huang, B., et al.: Self-supervised depth estimation in laparoscopic image using 3d geometric consistency. In: Medical Image Computing and Computer Assisted Intervention (2022)
9. Jo, K., Choi, Y., Choi, J., Chung, J.W.: Robust real-time detection of laparoscopic instruments in robot surgery using convolutional neural networks with motion vector prediction. Appl. Sci. **9**(14), 2865 (2019)
10. Koch, G., Zemel, R., Salakhutdinov, R., et al.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop, vol. 2. Lille (2015)
11. Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., Zhang, D.: DS-TransUNet: dual Swin transformer u-net for medical image segmentation. IEEE Trans. Instrum. Meas. **71**, 1–15 (2022)
12. Liu, F., Jonmohamadi, Y., Maicas, G., Pandey, A.K., Carneiro, G.: Self-supervised depth estimation to regularise semantic segmentation in knee arthroscopy. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 594–603. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_58
13. Liu, X., Li, Z., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Sage: slam with appearance and geometry prior for endoscopy. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 5587–5593. IEEE (2022)
14. Liu, X., et al.: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE Trans. Med. Imaging **39**(5), 1438–1447 (2019)
15. Maćkiewicz, A., Ratajczak, W.: Principal components analysis (PCA). Comput. Geosci. **19**(3), 303–342 (1993)
16. Marullo, G., Tanzi, L., Ulrich, L., Porpiglia, F., Vezzetti, E.: A multi-task convolutional neural network for semantic segmentation and event detection in laparoscopic surgery. J. Personalized Med. **13**(3), 413 (2023)
17. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Tao, R., Huang, B., Zou, X., Zheng, G.: SVT-SDE: spatiotemporal vision transformers-based self-supervised depth estimation in stereoscopic surgical videos. IEEE Trans. Med. Robot. Bionics **5**, 42–53 (2023)
20. Tukra, S., Giannarou, S.: Stereo depth estimation via self-supervised contrastive representation learning. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13437, pp. 604–614. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_58

21. Xu, C., Huang, B., Elson, D.S.: Self-supervised monocular depth estimation with 3-D displacement module for laparoscopic images. IEEE Trans. Med. Robot. Bionics **4**(2), 331–334 (2022)
22. Ye, M., Johns, E., Handa, A., Zhang, L., Pratt, P., Yang, G.Z.: Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. arXiv preprint arXiv:1705.08260 (2017)
23. Yoon, J., et al.: Surgical scene segmentation using semantic image synthesis with a virtual surgery environment. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13437, pp. 551–561. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_53