



RARPL6: Development of a clinical dataset for surgical workflow recognition from robot-assisted radical prostatectomy with lymphadenectomy

Ziyang Chen^{*†}
Politecnico di Milano, Milan, Italy
ziyang.chen@polimi.it

Alice Pierini[†]
Politecnico di Milano, Milan, Italy
alice.pierini@mail.polimi.it

Eleonora Pollini[†]
Politecnico di Milano, Milan, Italy
eleonora.pollini@mail.polimi.it

Raffaella Salama[†]
Politecnico di Milano, Milan, Italy
raffaella.salama@mail.polimi.it

Théo Pauvel
Télécom Physique
Strasbourg, Strasbourg, France
theo.pauvel@etu.unistra.fr

Elena Lievore
European Institute of Oncology,
Milan, Italy
elena.lievore@ieo.it

Giancarlo Ferrigno
Politecnico di Milano, Milan, Italy
giancarlo.ferrigno@polimi.it

Elena De Momi
Politecnico di Milano, Milan, Italy;
and European Institute of Oncology,
Milan, Italy
elena.demomi@polimi.it

ABSTRACT

Surgical workflow recognition has attracted widespread attention in robot-assisted surgery since it can provide surgical context information automatically, which releases the cognitive burden of the surgeons and allows more appropriate surgical decisions. One major dilemma in this community is the limitation of clinical datasets with annotated ground truth, because it requires experienced surgeons to provide specific recognition information during the annotation progress. In this paper, we developed a clinical dataset with annotated workflow information, and we provided a potential baseline for the evaluation of this dataset by predicting different surgical steps. Specifically, our dataset was captured from the robot-assisted radical prostatectomy with lymphadenectomy performed on six patients, using the da Vinci Xi robot at European Institute of Oncology, Milan, Italy, and all annotated outputs concerning various surgical information were obtained under the supervision of an experienced surgeon. Furthermore, an advanced neural network was adopted to predict surgical steps based on this dataset by using two different training strategies (i.e., the entire dataset and the downsampled one for the balance of class), and it presented a potential baseline (0.7825 DICE and 0.7918 DICE, respectively). It is expected that this dataset could promote the development of surgical workflow recognition in the medical image community, and this dataset is now accessible at the link: <https://zenodo.org/record/7644037>.

^{*}Corresponding author

[†]These authors contributed equally to this work



This work is licensed under a Creative Commons Attribution International 4.0 License.

ICBIP 2023, July 21–23, 2023, Chengdu, China
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0769-8/23/07.
<https://doi.org/10.1145/3613307.3613325>

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Computer vision; Computer vision tasks; Activity recognition and understanding.

KEYWORDS

Surgical workflow recognition, Robot-assisted surgery, Clinical dataset, Neural network

ACM Reference Format:

Ziyang Chen, Alice Pierini, Eleonora Pollini, Raffaella Salama, Théo Pauvel, Elena Lievore, Giancarlo Ferrigno, and Elena De Momi. 2023. RARPL6: Development of a clinical dataset for surgical workflow recognition from robot-assisted radical prostatectomy with lymphadenectomy. In *2023 8th International Conference on Biomedical Signal and Image Processing (ICBIP 2023)*, July 21–23, 2023, Chengdu, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3613307.3613325>

1 INTRODUCTION

Robots provide great benefits in the surgery field and the large number of robot-assisted procedures performed nowadays highlights them, such as less intraoperative bleeding and fewer adverse events during both the surgery and the recovery. Moreover, patients who undergo robot-assisted surgery have a better chance to remove positive surgical margins, leading to reduced pain, almost no need for transfusions and shorter hospital stays [1-3]. Advances in artificial intelligence also offer broad promise for robotic surgery, and one of the most representative applications is automatic surgical workflow recognition. It releases the cognitive overload of surgeons and allows better decision making and surgical planning by providing focused information, which could enhance the safety of surgery [4-5]. Particularly, various computer-assisted technologies present increasing demands for the recognition of surgical workflow, so that they could be applied more appropriately in different surgical contexts. For instance, novice surgeons perform repeated exercises

to master the surgery’s technique, involving activities that are common across different surgical tasks, and feedback from high-quality surgical motion data could help them to improve their skills. To understand which phase or step should be paid more attention to, it is needed to determine what surgical activities are taking place and when they are taking place, i.e., recognize the surgical workflow [6].

One of the challenges in automatic surgical workflow recognition is the limitation of the public datasets today. Annotating a clinical dataset always relies on medical background, which is more difficult than natural scene recognition [7]. Furthermore, the context and workflow information always significantly vary in different surgeries, promoting the demand to expand datasets of different surgical procedures. There are limited public datasets in the medical field, such as Cholec80 [8] captured from cholecystectomy surgeries, and Nephrec9 taken from robot-assisted partial nephrectomy [5]. Similarly, Robot-Assisted Radical Prostatectomy with Lymphadenectomy (RARPL) could be regarded as one of the most common procedures in robot-assisted surgery today, in which the entire prostate gland is removed; since cancer probably spreads in metastasis, it is necessary to excise both the prostate gland and the lymph nodes attached [9-10]. The Da Vinci Surgical System (DVSS) has been widely utilized in this kind of minimally invasive surgery. It was reported around three hundred procedures of RARPL per year with the Da Vinci robot are carried out at European Institute of Oncology, Milan, Italy, which shows the significance of this procedure and motivates us to capture the videos of RARPL. Although some of the actions performed during this surgical procedure differ according to the patient and the evolution of the tumor, surgeons have demonstrated how the sequence of steps is very consistent and almost always the same, which shows the generalization and potential to make this dataset. Hence, this paper presents two contributions:

- 1) A clinical dataset captured from RARPL procedures was made and annotated with different surgical workflow and context information.
- 2) A potential baseline for the evaluation of this dataset was provided using two different training strategies.

The rest of this paper is organized as follows: Section 2 presents the details of making this dataset; Section 3 gives the evaluation results using different metrics based on an advanced neural network; and finally, Section 4 summarizes this work and the next research.

2 DATASET DETAILS

Six complete surgical videos with a resolution of 1920×1080 and 25 Frames Per Second (FPS) were captured from patients who accepted the procedure of RARPL using the da Vinci Xi robot at European Institute of Oncology. Each procedure was recorded using a 3D HD video recorder (HVO-3300MT, SONY, Tokyo) with a length of approximately 55 minutes each, for a total of approximately 990 minutes. Annotation related to the specific surgical workflow was achieved under the guidance of an experienced surgeon. Specifically, 6 surgical steps of the procedure, representing the surgeons’ different actions, were defined as ‘Dissection’, ‘Traction’, ‘Clip of the vases (Clips)’, ‘Suction’, ‘Irrigation’ and ‘Suturing’. It could be seen that surgical steps contain important contextual information,

but the recognition is always full of challenges due to the alternating action and time continuity. In addition, 4 surgical phases were also annotated, including ‘Collapse of the peritoneum’, ‘Prostate removal’, ‘Lymphadenectomy’ and ‘Anastomosis’, as well as the types of instruments that appeared, including monopolar scissors, Maryland bipolar forceps, Cadieere forceps, suction tube, clip applicator and, only for the suturing part, the large needle driver. Figure 1 shows the recognition details of this procedure, providing the example frames, the number of frames collected, and the tools mainly used for each step and phase.

After the recognition of the surgical workflow, the original videos were divided into smaller videos manually following different steps using iMovie, hence six videos containing the six steps for each patient were obtained. Clips with heavy motion blur, bleeding or smoke caused by the cauterization during the dissection, which obstructed the view of the endoscope, were manually removed. Then, ANVIL, a free video annotation tool that offers multi-layered annotation based on a user-defined coding scheme [11], was adopted. On ANVIL it is possible to create some specifications that contain different attributes to annotate the video, so the step was set as the primary track, and another track, a subdivision of the primary one, was also highlighted, representing the phase. In the subdivision track, another attribute was also defined, related to the instruments. Once obtained all the frames related to the specific step, they were further divided into different phases, giving the dataset the property of being double-sided, as one can focus on the division of frames into steps or on the division into phases. This procedure was manually carried out, selecting the frames related to a specific phase, based on the ground truth labels obtained with ANVIL. The manual annotation took around 100 hours in our case. Figure 2 gives the distribution of frames in different steps. It could be seen that the dataset is not homogeneously distributed concerning steps. This result is consistent with what the surgeons have shown us: a large part of the RARPL operation is dedicated to dissection, while other steps are shorter and less recognizable.

Considering that RARPL is a surgical procedure with a reduced inter-variability between patients, the workflow of the whole procedure is quite always the same, as also confirmed by the surgeons of IEO hospital. Hence, a transition diagram is useful to understand the probabilities of going from one step to another. In general, a transition diagram of a Markov chain describes the probabilities associated with state changes, called transition probabilities, depending on the previous state. In our case, state changes are the transitions from one step to another, as shown in Figure 3.

3 CLASSIFICATION EVALUATION

3.1 Evaluation Approach

Considering that the surgical steps contain rich context information for the recognition of surgeon’s behaviors, we conducted the prediction of steps in this section. A state-of-the-art deep learning based approach named Temporal Memory Relation Network (TMRNet) [12] was chosen to conduct the evaluation. It adopts dual branches with different temporal inputs to incorporate context information, as shown in Figure 4. More specifically, an external long-range memory bank generated by ResNet and LSTM [13-15]

INSTRUMENTS LEGEND:
● Monopolar scissors ● Maryland bipolar forceps ● Cadiere forceps
● Suction tube ● Clip applicator ● Large needle driver

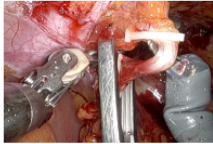
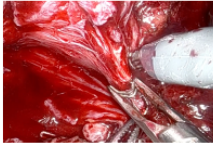
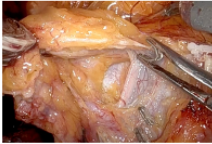
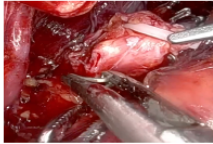

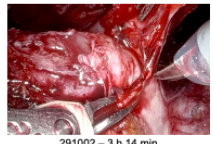


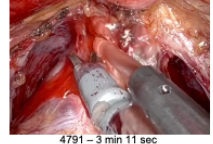
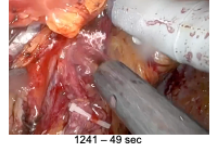
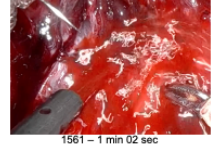
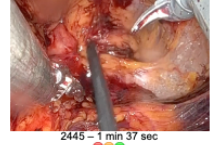
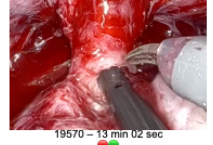
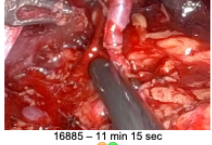

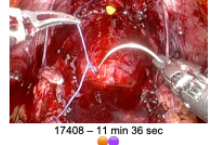


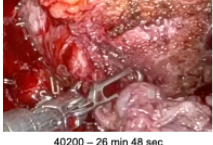

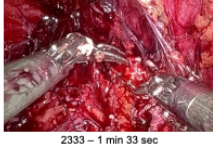
Step/Phase	Collapse of the peritoneum	Prostate removal	Lymphadenectomy	Anastomosis
Clips	 5045 - 3 min 21 sec ● ● ●	 26210 - 17 min 28 sec ● ● ●	 14662 - 9 min 46 sec ●	 8021 - 5 min 20 sec ● ● ● ●
Dissection	 89545 - 59 min 41 sec ● ●	 291002 - 3 h 14 min ● ●	 174355 - 1 h 56 min ●	 5826 - 3 min 53 sec ● ● ●
Irrigation	/	 4791 - 3 min 11 sec ● ●	 1241 - 49 sec ●	 1561 - 1 min 02 sec ● ●
Suction	 2445 - 1 min 37 sec ● ● ●	 19570 - 13 min 02 sec ● ● ●	 16885 - 11 min 15 sec ● ● ●	 14377 - 9 min 35 sec ● ● ●
Suturing	/	 17408 - 11 min 36 sec ● ● ●	/	 267201 - 2 h 58 min ● ● ●
Traction	 17298 - 11 min 31 sec ● ●	 40200 - 26 min 48 sec ●	 22768 - 15 min 10 sec ●	 2333 - 1 min 33 sec ● ● ●

Figure 1: The recognition details of our dataset. The data at the bottom of the picture represents the specific frame number and the entire duration. The three blank boxes mean that the surgeons didn't perform the corresponding steps at that phase.

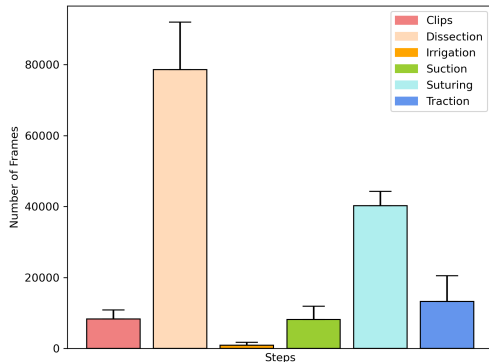
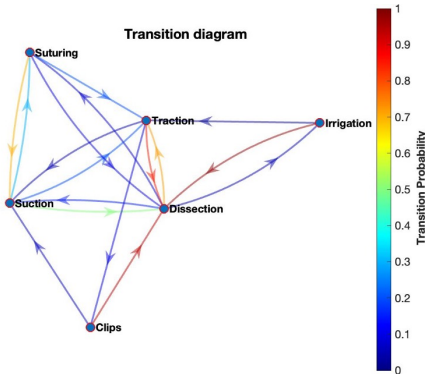
was first created to store the global surgical information by injecting the long video VC_l , and then multi-scale convolutions were adopted to enhance the representation of temporal features f_t . On another branch, a short video clip VC_s extracted from the long video was injected into the network using the same architecture consisting of ResNet and LSTM to generate the current features f_c . Then the feature clips f_t and f_c from dual branches were aggregated by implementing the non-local operator [12, 16] to generate the

attention features f_a . Finally, features f_c from the short clip were concatenated with f_a , and forwarded to two fully connected layers for the prediction.

To train the network, we divided the whole dataset into five videos as training data, and one video left for evaluation. The resolution of the original images is 1920×1080 which consumes lots of computing resources, so we resized the images into 250×250 to save memory. We also kept the same hyperparameters for the

Table 1: Recognition result of six surgical steps using all frames.

Recognized Steps	DICE	Precision	Recall	Specificity	Accuracy
Clips	0.6716	0.5922	0.7756	0.9833	0.9771
Dissection	0.8658	0.9732	0.7798	0.9588	0.8412
Irrigation	0.2014	0.2024	0.2004	0.9938	0.9877
Suction	0.4216	0.2970	0.7262	0.9412	0.9341
Suturing	0.8988	0.9833	0.8277	0.9954	0.9540
Traction	0.2064	0.1219	0.6722	0.8762	0.8711

**Figure 2: The intra-class variability of the number of frames in different steps.****Figure 3: Transition probabilities between different steps in the whole procedure.**

model training recommended in [12]. All experiments were conducted on a local computer with an Nvidia GeForce GTX 980 Ti. We noticed that the distribution of different steps is imbalanced, so two different training strategies were adopted in our evaluation: 1) We kept the original frames to train the model. 2) We downsampled the dataset to narrow the difference in the distribution. Specifically, we used the images of ‘Dissection’ and ‘Suturing’ every 25 frames, the images of ‘Clips’, ‘Suction’ and ‘Traction’ every 10 frames and we kept the original frames of ‘Irrigation’. Five common evaluation metrics were calculated for the dataset evaluation, including DICE Coefficient, precision, recall, specificity, and accuracy [17-18].

3.2 Evaluation Results

The normalized confusion matrices based on the entire dataset and the downsampled one were shown in Figure 5. We could notice that the predicted labels are more accurate after balancing the number distribution of steps than the original distribution since the true positive rates are increasing. However, the predicted accuracy of ‘Irrigation’ is still low (0.20 and 0.21 respectively). It can be seen that the prediction among ‘Irrigation’, ‘Dissection’ and ‘Suction’ was difficult to achieve when the true step is ‘Irrigation’, since the prediction probability of other two steps was even higher than the correct label. The step of ‘irrigation’ occupies the minimum number of frames in the entire dataset, while we augmented the proportion of this step in the downsampled dataset. Nevertheless, the prediction performance of this step is still unsatisfactory, which can be given by the fact that the discrimination among those three steps was inconspicuous and challenging after our observation. Table 1 and Table 2 also showed the prediction values in each step using different metrics. On the one hand, the performance keeps better when adopting downsampled dataset than the original one, and the prediction of ‘Clips’ got the most improvement. On the other hand, we found the step of ‘Traction’ performs badly since its precision keeps low (0.1219 and 0.3394 respectively). It means that the false positive rates are high in this step, i.e., other steps (specifically, ‘Clips’, ‘Dissection’, ‘Irrigation’ and ‘Suction’) were prone to be predicted as ‘Traction’, which can also be observed in Figure 5. Finally, we calculated the DICE by adding all six steps together, and the values were 0.7825 and 0.7918 using the entire dataset and the downsampled dataset, respectively. It could be regarded as a baseline for the evaluation of more advanced prediction approaches in the future.

4 CONCLUSION

This paper proposed a public clinical dataset for automatic surgical workflow recognition for safer robotic surgery. It was captured using da Vinci Xi surgical system based on robot-assisted radical prostatectomy with lymphadenectomy on six patients. We also provided the annotated surgical workflow and context information under the supervision of an experienced surgeon. An advanced method was adopted to evaluate our dataset using two different training strategies. Given the comprehensive results calculated by different metrics, our dataset could provide a potential baseline in the surgical workflow recognition community. Future work will continue the surgical workflow recognition using 3D surgical scenes. We intuitively think that adding extra depth information using 3D surgical scene reconstruction [19-22] could enhance the

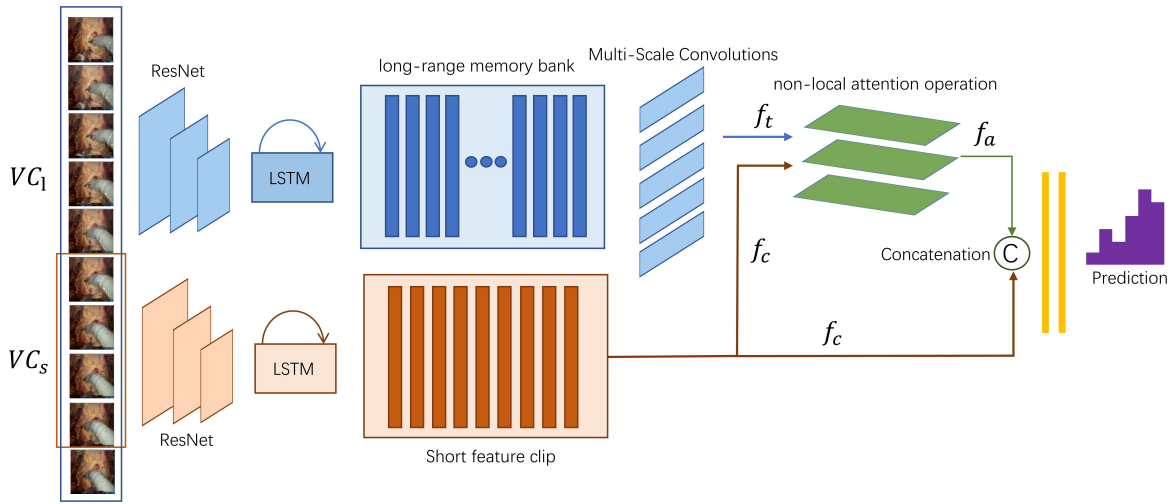


Figure 4: The architecture of the adopted temporal memory relation network

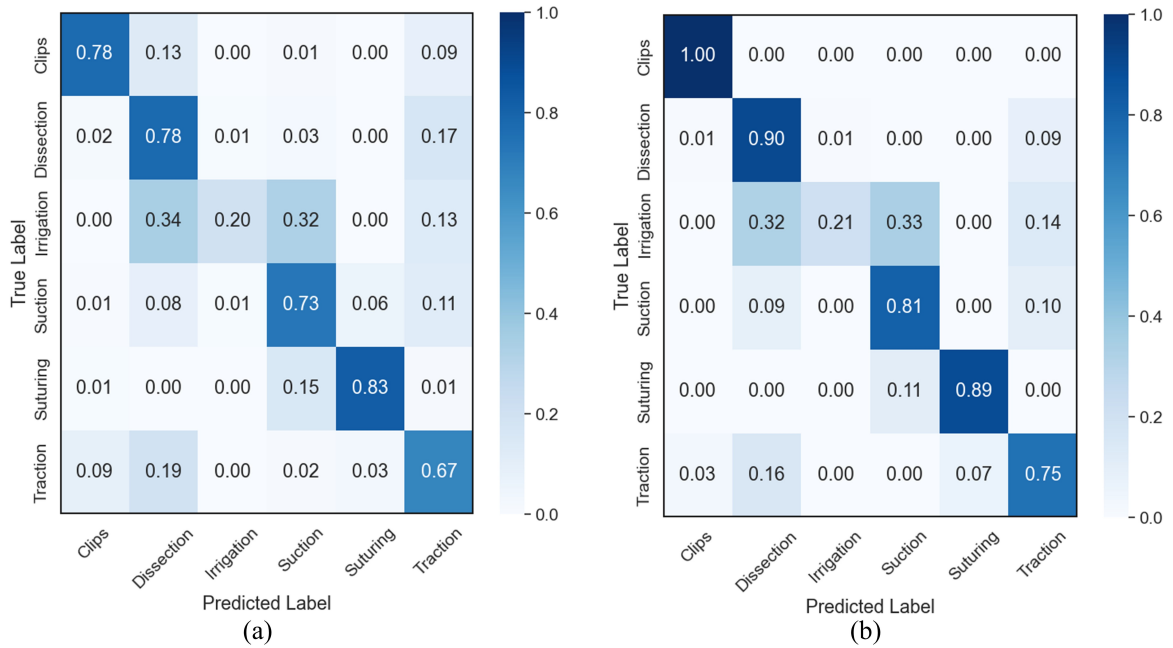


Figure 5: The normalized confusion matrices with two different training strategies. (a) is the result using the complete dataset, while (b) represents the downsampled one.

Table 2: Recognition result of six surgical steps using the downsampled dataset.

Recognized Steps	DICE	Precision	Recall	Specificity	Accuracy
Clips	0.9581	0.9218	0.9974	0.9948	0.9950
Dissection	0.8930	0.8826	0.9037	0.8806	0.8921
Irrigation	0.3401	0.9244	0.2084	0.9971	0.8813
Suction	0.5576	0.4251	0.8103	0.9265	0.9192
Suturing	0.9354	0.9819	0.8932	0.9962	0.9769
Traction	0.4671	0.3394	0.7484	0.9276	0.9191

recognition performance since different soft tissues and tools are easier to be recognized in 3D space, which will be verified in the future.

REFERENCES

- [1] Sestini Luca, Rosa Benoit, De Momi Elena, *et al.* FUN-SIS: A fully unsupervised approach for surgical instrument segmentation[J]. *Medical Image Analysis*, 2023: 102751.
- [2] Da Col Tommaso, Mariani Andrea, Deguet Anton, *et al.* Scan: System for camera autonomous navigation in robotic-assisted surgery[C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 2996-3002.
- [3] Chen Ziyang, Terlizzi Serenella, Da Col Tommaso, *et al.* Robot-assisted ex vivo neobladder reconstruction: preliminary results of surgical skill evaluation[J]. *International journal of computer assisted radiology and surgery*, 2022, 17(12): 2315-2323.
- [4] Zhang Bokai, Goel Bharti, Sarhan MohammadHasan, *et al.* Surgical workflow recognition with temporal convolution and transformer for action segmentation[J]. *International Journal of Computer Assisted Radiology and Surgery*, 2022: 1-10.
- [5] Nakawala Hirenkumar, Bianchi Roberto, Pescatori LauraErica, *et al.* “Deep-Onto” network for surgical workflow and context recognition[J]. *International journal of computer assisted radiology and surgery*, 2019, 14: 685-696.
- [6] DiPietro Robert, Ahmidi Narges, Malpani Anand, *et al.* Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks[J]. *International journal of computer assisted radiology and surgery*, 2019, 14(11): 2005-2020.
- [7] Iodice Francesco, De Momi Elena, Ajoudani Arash. Hri30: An action recognition dataset for industrial human-robot interaction[C]. 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022: 4941-4947.
- [8] Twinanda Andru P, Shehata Sherif, Mutter Didier, *et al.* Endonet: a deep architecture for recognition tasks on laparoscopic videos[J]. *IEEE transactions on medical imaging*, 2016, 36(1): 86-97.
- [9] Student Jr Vladimir, Tudos Zbynek, Studentova Zuzana, *et al.* Effect of Peritoneal Fixation (PerFix) on Lymphocele Formation in Robot-assisted Radical Prostatectomy with Pelvic Lymphadenectomy: Results of a Randomized Prospective Trial[J]. *European Urology*, 2023, 83(2): 154-162.
- [10] Deutsch Sebastian, Hadaschik Boris, Lebentrau Steffen, *et al.* Clinical importance of a peritoneal interposition flap to prevent symptomatic lymphoceles after robot-assisted radical prostatectomy and pelvic lymph node dissection: a systematic review and meta-analysis[J]. *Urologia Internationalis*, 2022, 106(1): 28-34.
- [11] Kipp Michael. Anvil-a generic annotation tool for multimodal dialogue[C]. Seventh European conference on speech communication and technology. 2001.
- [12] Jin Yueming, Long Yonghao, Chen Cheng, *et al.* Temporal memory relation network for workflow recognition from surgical video[J]. *IEEE Transactions on Medical Imaging*, 2021, 40(7): 1911-1923.
- [13] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [14] Shi Xingjian, Chen Zhourong, Wang Hao, *et al.* Convolutional LSTM network: A machine learning approach for precipitation nowcasting[J]. *Advances in neural information processing systems*, 2015, 28.
- [15] Jin Yueming, Dou Qi, Chen Hao, *et al.* SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network[J]. *IEEE transactions on medical imaging*, 2017, 37(5): 1114-1126.
- [16] Wang Xiaolong, Girshick Ross, Gupta Abhinav, *et al.* Non-local neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.
- [17] Casella Alessandro, Moccia Sara, Paladini Dario, *et al.* A shape-constraint adversarial framework with instance-normalized spatio-temporal features for inter-fetal membrane segmentation[J]. *Medical Image Analysis*, 2021, 70: 102008.
- [18] Moccia Sara, De Momi Elena, El Hadji Sara, *et al.* Blood vessel segmentation algorithms—review of methods, datasets and evaluation metrics[J]. *Computer methods and programs in biomedicine*, 2018, 158: 71-91.
- [19] Yang Gengshan, Sun Deqing, Jampani Varun, *et al.* Viser: Video-specific surface embeddings for articulated 3d shape reconstruction[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 19326-19338.
- [20] Wang Yuehao, Long Yonghao, Fan SiuHin, *et al.* Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery[C]. *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII*. Cham: Springer Nature Switzerland, 2022: 431-441.
- [21] Sharma Kirti Shankar, Manivannan P V. Stereo Image Partitioning Based Fuzzy Logic[J]. *International Journal of Mechanical Engineering and Robotics Research*, 2020, 9(8).
- [22] Chen Ziyang, Marzullo Aldo, Alberti Davide, *et al.* FRSR: Framework for real-time scene reconstruction in robot-assisted minimally invasive surgery[J]. *Computers in Biology and Medicine*, 2023: 107121.