3D Reconstruction and Segmentation in Laparoscopic Robotic Surgery

Laura Cruciani
Department of Electronics, Information and Bioengineering
Politecnico di Milano
Milan, Italy
laura.cruciani@mail.polimi.it

Ziyang Chen
Department of Electronics, Information and Bioengineering
Politecnico di Milano
Milan, Italy
ziyang.chen@mail.polimi.it

Matteo Fontana
Department of Urology, IRCSS Foundation
IEO, European Institute of Oncology
Milan, Italy
Matteo.Fontana@ieo.it

Gennaro Musi
Department of Urology, IRCSS Foundation
IEO, European Institute of Oncology
Milan, Italy

Ottavio de Cobelli
Department of Urology, IRCSS Foundation
IEO, European Institute of Oncology
Milan, Italy

Elena De Momi
Department of Electronics, Information and Bioengineering
Politecnico di Milano
Milan, Italy
elena.demomi@mail.polimi.it

***Abstract*—Laparoscopic robotic surgery has redefined complex surgical interventions offering enhanced precision and control. However, challenges persist due to the limited field of view and the lack of haptic feedback. To address these challenges, assistive technologies, such as Augmented Reality (AR), are gaining prominence, promising safer procedures, quicker recovery, and reduced hospitalization.**

To develop these technologies, the integration of computer vision tasks into robotic systems is a fundamental step, allowing robots to perceive and understand the surgical environment in real-time. In this context, Deep Learning (DL) techniques are key technologies in addressing complexities arising from intra-patient variability, lighting issues, occlusions, and texture limitations.

In this study, we present a comparative evaluation of various deep learning techniques to demonstrate their efficacy in reconstructing the surgical scene in 3D and accurately segmenting blood vessels with the aim to develop AR applications in the future. The accuracy and speed of these methods have been validated and compared using public or private laparoscopic datasets, providing valuable insights into the strengths and weaknesses of each approach.

The obtained results underscore the high potential of deep learning approaches, especially when leveraged on high-performance GPUs, to meet the demanding requirements in terms of accuracy and speed imposed by the surgical context. These findings represent a foundational step towards conducting more comprehensive comparisons in the future, ultimately paving the way for the advancement of augmented reality (AR) applications in surgical procedures.

***Index Terms*—3D reconstruction, Laparoscopic segmentation, Augmented Reality**

I. Introduction

Laparoscopic robotic surgery has revolutionized the approach to surgical procedures, offering surgeons increased precision and enhanced control during interventions. However, despite these advancements, surgeons still face significant challenges due to the reduced field of view and the lack of haptic feedback, factors that make the technique extremely dependent on the experience of the practitioner [3, 4]. A possible solution to face these challenges relies on the integration of assistive technologies, such as Augmented Reality (AR) [6], in order to reduce the risk of complications and increase surgical precision.

AR applications allow to overlay virtual information onto the real operative field, providing surgeons with additional detailed vision and critical information during the procedure. The development of these applications involves the integration of combined vision algorithms to create immersive and contextually relevant experiences for users and offer a wide range of functionalities, from object recognition and tracking to spatial mapping and gesture recognition. In the context of Robotic Assisted Minimally Invasive Surgery (RAMIS), AR can be leveraged to provide surgeons with enhanced visualization, information about surgical planning and navigation, as well as visual feedback related to instrument distance and anatomical structures. While assistive technologies remain an evolving research area, their integration into RAMIS is complicated by issues inherent to the surgical context. These challenges primarily stem from intra- and inter-patient variability, challenging lighting conditions, dynamic scenes, occlusions, and areas with limited texture. A promising approach to tackle these challenges involves using Deep Learning (DL) techniques that, compared to traditional methods, have demonstrated significant promise in handling vision tasks even under complex conditions. By learning patterns and feature representations from extensive datasets, these approaches enhance accuracy and robustness. Thus, Deep Learning holds the potential to address the quality of outcomes in the field of assistive technologies within RAMIS. For this reason, the aim of this study is to investigate the potential of deep learning techniques

in addressing the challenges of laparoscopic robotic surgery, paving the way for the advancement of augmented reality applications and ultimately enhancing the quality of surgical procedures.

II. Methodologies

In this work, we present a comparative evaluation of some state-of-the-art Deep Learning techniques. Our focus centers on three distinct DL approaches designed for computing the disparity map and performing 3D scene reconstruction through passive triangulation: HSM [9], CFNet [8], and RAFT [5]. Additionally, we explore two methods to perform the vessel segmentation: U-Net [7] and SETR [10]. The evaluation, in terms of accuracy and speed, was performed using Ubuntu server with an NVIDIA A100 GPU. The Wilcoxon rank-sum test is then performed to show significant differences in the performance.

A. Dataset

3D reconstruction models have been evaluated by testing the weights provided by the authors in a publicly available medical dataset (SCARED) [1], composed by images with a resolution of 1280x1024 pixels.
Segmentation models have been trained and tested using a private dataset provided by the Istituto Europeo di Oncologia: the dataset was composed by 686 images with a resolution of 1280x960 in which the common iliac artery was manually annotated using CVAT annotation Tool. The 80% of the data were used for training, tuning the hyperparamenters, and 20% of the data was used for testing.

B. Performance metrics

1) 3D reconstruction: The accuracy metrics were computed by calculating the error between the ground truth (GT) disparity map values and the predicted values. Every DL model was evaluated in terms of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), described by the following equations in which d'(x,y) is the value of GT in the pixel (x,y) and d(x,y) is the value of the predicted disparity map:

- **Mean Absolute Error (MAE)=**

$$\frac{1}{N}\sum_{i=1}^{N}|d_i(x,y)-d'_i(x,y)| \tag{1}$$

- **Root Mean Square Error (RMSE)=**

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i(x,y)-d'_i(x,y))^2} \tag{2}$$

2) Segmentation: The two proposed segmentation methods were evaluated in terms of speed and accuracy. The training was performed tuning the set of hyperparameters and then they were tested extracting performance metrics. These metrics need the values of true positive (TP), pixels correctly classified as part of the vessel, true negative (TN), pixels correctly classified as background, false positive (FP), pixels classified as vessel while contained in the background, and false negative (FN), pixels classified as background while belonging to the vessel. The employed metrics are described by the following equations:

- **Dice Similarity Coefficient**

$$DSC = \frac{2TP}{2TP+FP+FN} \tag{3}$$

- **Accuracy**

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

- **Sensitivity**

$$SENS = \frac{TP}{TP+FN} \tag{5}$$

- **Precision**

$$PREC = \frac{TP}{TP+FP} \tag{6}$$

III. Results

A. 3D reconstruction

The results of the 3D reconstruction models are summarized in Table I, which presents the MAE, RMSE, and inference time for each model. The three implemented methods demonstrated comparable accuracy in predicting the disparity map, without statistical significance. However, the HSM technique outperformed the others in terms of real-time performance, making it more suitable for practical applications. It's important to note that resolution of the images is highly correlated with latency, and resizing the image could potentially improve speed performance. However, further experiments are required to evaluate if image resizing could affect the reconstruction error.

	MAE	RMSE	Time[s]
HSM	2.64±1.64	5.47±1.46	**0.06±0.00**
CFNet	2.72±1.48	5.54±1.39	0.45±0.08
RAFT	2.63±1.65	5.47±1.49	2.15±0.01

TABLE I: Performance metrics of the 3D reconstruction models. The highligthed values represent the significant difference between the different architectures in terms of the Wilcoxon rank-sum test with $p<0.01$

B. Segmentation

Figure 1 reports the performance results obtained with clinical dataset while Table II shows the inference time. The results demonstrate that the U-Net architecture achieves significantly higher levels of accuracy compared to other architectures, although it comes with the trade-off of a higher inference time. However, meeting real-time requirements is still feasible by resizing the images or employing high-performance GPUs. It is worth noting that the dataset used in this study is characterized by several issues related to the clinical context, which considerably complicates the task and ultimately reduces the accuracy of the outcomes. Despite these challenges, the U-Net architecture shows great potential for improving accuracy, and with optimization strategies, real-time performance can still be achieved for practical applications.

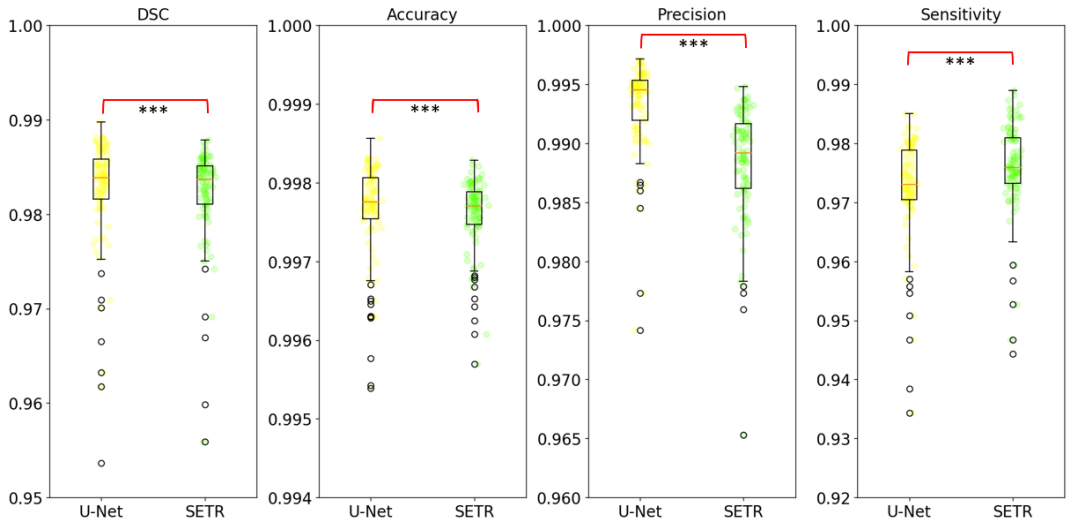

Fig. 1: Segmentation accuracy results. The asterisks represent the significant difference between the different architectures in terms of the Wilcoxon rank-sum test with ***p<0.001

	Inference time [s]
U-Net	0.20±0.03
SETR	**0.09±0.01**

TABLE II: Inference time of the segmentation models. The highligthed value represent the significant difference between the different architectures in terms of the Wilcoxon rank-sum test with p <0.01

IV. Conclusion

In conclusion, the integration of assistive technologies into RAMIS holds great promise for advancing surgical outcomes. In this work, we contribute by conducting a comprehensive comparative evaluation of various DL techniques for 3D reconstruction and segmentation in the context of laparoscopic robotic surgery. The evaluation of DL techniques for 3D reconstruction revealed that while different methods displayed similar accuracy in predicting disparity maps, the HSM technique demonstrated superior real-time performance. In the realm of segmentation, the U-Net architecture shows higher accuracy albeit with increased inference time, suggesting optimization avenues. Despite the challenges coming from clinical context, these outcomes underscore the potential of DL techniques in enhancing surgical precision and patient care. Further research and development in this area, emphasizing the integration of computer vision tasks into robots, could lead to even greater strides in the field of laparoscopic robotic surgery [2].
In the attached video, it's possible to observe the results of combining 3D reconstruction and vessel segmentation into a surgical scene. The visual presentation, in fact, showcases the 3D reconstruction of the surgical scene, with the common iliac artery highlighted in green. This demo highlights the practical benefits of our work, representing a foundation for the development of augmented reality (AR) applications in laparoscopic robotic surgery, with the aim of enhancing surgical procedures through AR technology, improving surgical precision and patient outcomes.

References

[1] Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021.

[2] Ziyang Chen, Aldo Marzullo, Davide Alberti, Elena Lievore, Matteo Fontana, Ottavio De Cobelli, Gennaro Musi, Giancarlo Ferrigno, and Elena De Momi. Frsr: Framework for real-time scene reconstruction in robot-assisted minimally invasive surgery. *Computers in Biology and Medicine*, page 107121, 2023.

[3] Ziyang Chen, Serenella Terlizzi, Tommaso Da Col, Aldo Marzullo, Michele Catellani, Giancarlo Ferrigno, and Elena De Momi. Robot-assisted ex vivo neobladder reconstruction: preliminary results of surgical skill evaluation. *International Journal of Computer Assisted Radiology and Surgery*, 17(12):2315–2323, 2022.

[4] Emanuele Colleoni, Sara Moccia, Xiaofei Du, Elena De Momi, and Danail Stoyanov. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. *IEEE Robotics and Automation Letters*, 4(3):2714–2721, 2019.

[5] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021.

[6] Veronica Penza, Sara Moccia, Elena De Momi, and Leonardo S Mattos. Enhanced vision to improve safety

in robotic surgery. In *Handbook of Robotic and Image-Guided Surgery*, pages 223–237. Elsevier, 2020.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[8] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2021.

[9] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019.

[10] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.