Semantic Image Synthesis for Realistic Image Generation in Robotic Assisted Partial Nephrectomy

Stefano Mazzocchetti[1], Laura Cercenelli[1], Lorenzo Bianchi[2,3], Riccardo Schiavina[2,3] and Emanuela Marcelli[1]

[1]*eDIMES Lab, Laboratory of Bioengineering, Department of Medical and Surgical Sciences, University of Bologna, Via Massarenti, 9, 40138 Bologna, Italy*

[2]*Division of Urology, IRCCS Azienda Ospedaliero, Universitaria di Bologna, Via Massarenti, 9, 40138 Bologna, Italy*

[3]*Department of Medical and Surgical Sciences, University of Bologna, Via Massarenti, 9, 40138 Bologna, Italy*

Keywords: Minimally Invasive Surgery, Robotic Surgery, Semantic Image Synthesis, Deep Learning, GAN, Computer Vision.

Abstract: With the continuous evolution of robotic-assisted surgery, the integration of advanced technologies into the field becomes pivotal for improving surgical outcomes. The lack of labelled surgical datasets limits the range of possible applications of deep learning techniques in the surgical field. As a matter of fact, the annotation process to label datasets is time consuming. This paper introduces an approach for realistic image generation in the context of Robotic Assisted Partial Nephrectomy (RAPN) using the Semantic Image Synthesis (SIS) technique. Leveraging descriptive semantic maps, our method aims to bridge the gap between abstract scene representation and visually compelling laparoscopic images. It is shown that our approach can effectively generate photo-realistic Minimally Invasive Surgery (MIS) synthetic images starting from a sparse set of annotated real images. Furthermore, we demonstrate that synthetic data can be used to train a semantic segmentation network that generalizes on real data reducing the annotation time needed.

1 INTRODUCTION

The transition from traditional open surgeries to minimally invasive procedures, such as laparoscopy, has significantly reduced patient trauma and recovery times. Concurrently, the integration of computer vision and image synthesis techniques into the surgical domain has shown great potential for enhancing surgical planning, training, and intraoperative decision-making. As the field of robotic-assisted surgery continues to evolve, there is an increasing demand for advanced technologies that enhance both preoperative planning and intraoperative decision-making. Data-driven methods can develop solutions for Computer Assisted Interventions (CAI) to support surgeons during the procedure (Vercauteren et al., 2019). This paper presents an approach to generate photo-realistic laparoscopic images in the context of Robotic Assisted Partial Nephrectomy (RAPN). Our primary objective is to demonstrate that descriptive semantic maps can serve as a bridge between abstract scene representation and visually compelling, anatomically accurate images. The key novelty of this paper lies in the application of semantic image synthesis (SIS) specifically to RAPN, showcasing that semantic maps can effectively guide the generation of images with high anatomical fidelity. Our work serves as an initial step towards a comprehensive framework for computer-assisted interventions through augmented reality. Furthermore, we establish a foundation for incorporating knowledge about object positioning into downstream tasks, such as 6-degree-of-freedom (6-DoF) pose estimation.

In the subsequent sections of this paper, we will delve into the related works in the field of image generation for Minimally Invasive Surgery (MIS), the dataset used for training and the methodology employed for semantic image synthesis in the context of laparoscopic surgery. Additionally, we will discuss the assessment methodology and details of the experiments undertaken, followed by concluding discussions.

Mazzocchetti, S., Cercenelli, L., Bianchi, L., Schiavina, R. and Marcelli, E.
Semantic Image Synthesis for Realistic Image Generation in Robotic Assisted Partial Nephrectomy.
DOI: 10.5220/0012611200003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 4: VISAPP*, pages 645-652
ISBN: 978-989-758-679-8; ISSN: 2184-4321

2 RELATED WORK

Image-to-image translation is a computer vision task that involves converting an input image from one domain to an output image in a different domain while preserving relevant structures and features. The goal is to learn a mapping between the two domains (Zhu et al., 2017; Isola et al., 2017), allowing the transformation of images from, for example, grayscale to color, or from satellite imagery to maps. This task is often approached using generative models, such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) or Variational Autoencoders (VAEs) (Kingma et al., 2019), to learn the complex relationships between the input and output domains. Image-to-image translation finds applications in various fields and recently has also been exploit in the medical domain.

(Pfeiffer et al., 2019) proposed a method based on unpaired image to image translation (Zhu et al., 2017) to translate simulated images taken from a 3D software to real laparoscopic images. Those methods relies on an unpaired dataset, i.e. where there is not a one-to-one correspondence between an image in domain A and an image in domain B. The authors exploited the network proposed by (Huang et al., 2018) with an additional structural similarity loss to preserve image content. (Rivoir et al., 2021) combined unpaired image-to-image translation and neural rendering in order to transfer simulated to photorealistic surgical abdominal scenes with a long-term consistency in the video. However, those methods were proposed for liver segmentation and require a 3D scene setup from real patient-specific 3D mesh obtained from medical imaging like Computed Tomography (CT). Moreover, (Ozawa et al., 2021) employ the cycle GANs (Zhu et al., 2017) to generate realistic synthetic data for surgical instrument segmentation.

Another generative method is the Semantic Image Synthesis (SIS) task which generates realistic images starting from a semantic map. It was first introduced by (Isola et al., 2017). Usually, it requires a paired dataset, consisting of the coupling of the real image to the associated semantic map. Most of the works rely on the conditional GANs (Isola et al., 2017). In order to augment the labelled training data for deep learning algorithms, (Rau et al., 2019) translated endoscopic images into depth maps applying Image-to-image translation (Isola et al., 2017). Recently, (Yoon et al., 2022) released a dataset composed by real labelled data and synthetic images generated by semantic image synthesis. They combined real data segmented manually and a virtual surgery environment created from the 3D organ meshes obtained from the CT with different surgical instruments. They used the SIS to minimize the semantic gap between real and synthetic data. They showed that synthetic data are no longer helpful when the models already achieve high performance with the real data.

Another work that exploits SIS to generate data is (Marzullo et al., 2021). Starting from the EndoVis 2017 surgical instrument segmentation task dataset (Allan et al., 2019), they added coarse segmentation of fat and organ tissue to perform SIS with (Isola et al., 2017). Different from all previous works, this approach is used to generate photorealistic images for the the (RAPN) surgical procedure. A dataset was collected from six surgical procedures and labelled with more semantic information with respect to (Marzullo et al., 2021).Even with a limited training data availability, exploiting the SPatially-Adaptive (DE)normalization SPADE (Park et al., 2019) architecture, a semantic segmentation network has been trained on the generated data achieving good quantitative and qualitative results. Adding more semantic features to the image let a better mapping between semantic information and realistic images.

3 MATERIALS AND METHODS

In this section, the dataset and the network used for the experiments are described.

3.1 Data

The data exploited for this work consist of 2D in-vivo images from (RAPN) surgical procedures performed with the da Vinci Xi robot at Division of Urology - IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna. Six clinical cases of patients with clinical diagnoses of T1 renal mass extracted from data acquired for a previous work (Bianchi et al., 2020) have been used. Participants signed a written informed consent document. The study was approved by our Institutional Ethics Committee (IRB approval 3386/2018). A total number of 318 frames were obtained from the procedures where the kidney is fully or partially visible. The number of frames for each clinical case (C) can be seen in Figure 1.

In order to perform the Semantic Image Synthesis task, each frame has been carefully labelled with a software tool (Wada,). Each pixel can be identified as one of the following 12 classes: background, kidney, Monopolar Curved Scissors, Fenestrated Bipolar, Bipolar Forceps, fat, abdomen tissue, liver, blood, gauze, renal vein and others (a class con-

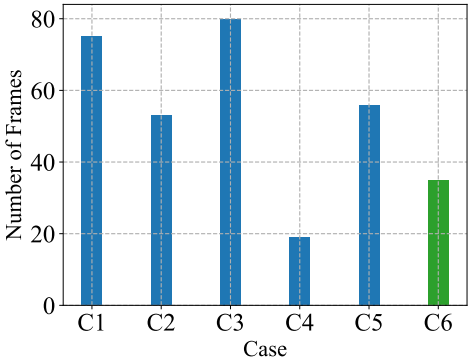

Figure 1: Number of frames extracted for each case. The case C6 has been used as an additional test set.

taining rare surgical tools that can be seen in the procedure). Figure 2 shows two samples and the corresponding semantic map. It can be noticed that differently from (Allan et al., 2019) or others Minimally Invasive Surgery (MIS) datasets, a part segmentation of the surgical tools is not present. The semantic mask is accurate only for the surgical tools, kidney and liver (when present). In a different way from (Marzullo et al., 2021), were the authors added only two semantic information beside the manipulators to the EndoVis (Allan et al., 2019) dataset, more information is present in the semantic map which can help the generative network to perform a better mapping between the mask and the real scene. All the extracted

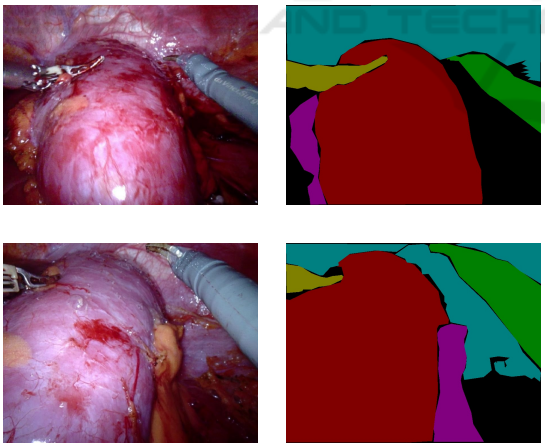

Figure 2: Training data samples and corresponding semantic segmentation map.

frames have been reshaped to a fixed size of 256x256 for memory restrictions. Five clinical cases (C1-C5) have been used to generate the train (168 : 60%), validation (26 : 10%) and test(89 : 30%) sets. Moreover, in order to test the ability to generalize on unseen data of the networks, the sixth clinical case (C6) has been used as an additional test set and is composed of 35 frames. The number of training set data has been augmented to a total of 1448 frames by random vertical or horizontal flip, elastic and affine transformation or a combination of those.

3.2 Semantic Image Synthesis Network

Generative Adversarial Network (GAN) (Goodfellow et al., 2020) are generative networks that the learn probability distribution of the dataset by learning a mapping between a random noise vector to an image. It is achieved by a min-max optimization between a Generator G and a Discriminator D. Additionally, conditional GANs (Isola et al., 2017) leverages on adversarial training to let a generator G learn a mapping between a condition c and a random noise vector z to an output image y:

$$G : \{c, z\} \longrightarrow y \tag{1}$$

In contrast, the discriminator D is trained to distinguish between real and fake images generated by G. Indeed, the generator G and discriminator D are trained simultaneously and they compete to maximize their own payoff. The supervision is achieved since the dataset is composed by pairs $\{(m_i, x_i)\}$, where for each image x_i exists the semantic map associated m_i. Given a segmentation mask $m_i \in \mathbf{L}^{H \times W}$ with image height H, width W and where $\mathbf{L}$ is a set of integers denoting the semantic labels, SIS networks aim to learn a mapping function that converts input segmentation masks to photo-realistic images. In this case the segmentation mask m acts as the condition for the generative model.

In this work, the SPADE (Park et al., 2019) has been employed to perform the image synthesis task. The authors built the architecture starting from a previous work (Wang et al., 2018), where they added a SPatially-Adaptive (DE)normalization. In previous methods, the semantic map was passed directly as input to the network and processed by stacks of convolution, normalization and activation layers. In particular, the normalization layers tend to take out the semantic information. For this reason, (Park et al., 2019) introduced a new conditional normalization method. After every batch normalization, the spatially adaptive modulation parameters are generated directly from the condition (semantic map) by projecting c to an embedding space and then convolved to produce two spatial modulation parameters that are added and multiplied in an element-wise manner with the batch normalization output. With this setting, the SPADE Residual Block is composed by a stack of two SPADE, ReLU and convolution layers. The generator is composed of several SPADE residual blocks

with upsampling layers where the semantic map is downsampled to the right dimension at each stage of the SPADE residual block. In particular, differently from (Wang et al., 2018), the generator G is composed of only the decoding block since there is no need to encode the segmentation map. In this work, the deterministic version of SPADE has been used, where the generator G starts with processing a downsampled version of the semantic map c. On the other hand, the discriminator D is a multi-scale patch-based fully convolutional network (Long et al., 2015) and takes as inputs the concatenation of the semantic map and the generated image and the concatenation of the semantic map and the real image. The average prediction of all patches is used to classify the whole image as real or fake. Conditional GANs are trained with the adversarial setting trying to model the conditional distribution of the real image given the semantic map to solve the adversarial min-max problem:

$$\min_G \max_D L_{GAN}(G,D) \tag{2}$$

In particular, the loss is composed by 3 terms, the GAN loss, discriminator-based feature matching loss and VGG perceptual loss (Wang et al., 2018). For further details please refer to the original paper (Park et al., 2019).

4 EXPERIMENTS

In the following section the quantitative and qualitative results are presented. All experiments were performed with PyTorch on a NVIDIA GeForce RTX 3070 Laptop GPU. The SPADE (Park et al., 2019) network was trained for 50 epochs and Adam as optimizer with a learning rate of $2e^{-4}$ with a batch size of 2.

4.1 Evaluation Protocol

The evaluation of synthesized images is a challenging problem. Following previous works (Isola et al., 2017; Park et al., 2019; Wang et al., 2018), a segmentation network has been trained on the real images and tested on the synthetic data generated from the test set and the one generated from the additional test set. This network is a U-Net (Ronneberger et al., 2015) architecture with a ResNet18 (He et al., 2016) backbone pretrained on ImageNet (Deng et al., 2009). Indeed, if the generated images reflect the distribution of the trained data, the segmentation network should generalize well on synthetic data.

Additionally, to demonstrate that the generated data can be exploited to train deep learning networks, a segmentation network with only 3 output classes (background, kidney and surgical tool) has been trained on synthetic generated data and tested on the real test set and the additional one. Both the segmentation networks have been trained for 60 epochs, Adam optimizer with learning rate of $1e^{-4}$ and batch size of 8. In this case, the validation set has been used to select the best epoch. The loss function is a weighted sum of the cross-entropy loss and the multi-class dice loss. Semantic segmentation results are evaluated quantitatively in terms of Dice Score, Intersection over Union (IoU), Precision (P) and Recall (R):

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

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

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

Where, given the confusion matrix the True Positive (TP), False Positive (FP), True Negatives (TN) and False Negatives (FN) are defined. The Dice Score is defined as two times the area of intersection divided by the sum of the areas of the prediction and ground truth mask. The IoU (or Jaccard index) is defined as the intersection between the predicted segmentation with respect to the ground truth divided by the area of the union. The Precision (P) gives a measure of how many pixels are labelled correctly over the overall prediction. It is a measure of the quality of the prediction. Having high precision tells that the network can accurately predict the correct pixel label with low FP labels. On the other hand, the Recall (R) measures the completeness of the prediction performed against all the relevant ground truth pixels.

Table 1: Mean intersection over union (mIoU), Dice score (Dice), precision (P) and recall (R) of the U-Net architecture pre-trained on the real train set computed on the synthetic test set (STS) and synthetic additional test set (SATS).

		mIoU	Dice	P	R
STS	Mean	0.78	0.82	0.85	0.90
	Std	0.13	0.12	0.10	0.06
SATS	Mean	0.73	0.78	0.83	0.86
	Std	0.13	0.12	0.10	0.08

4.2 Results

In Table 1 there are the results for the pre-trained segmentation network on real data evaluated on the syn-

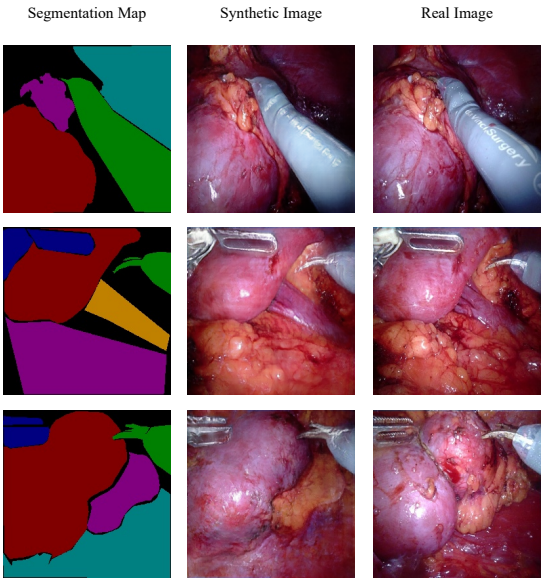

Figure 3: Some examples of synthetic images (centre) generated from the semantic segmentation map (left) in comparison with the corresponding real image (right) for the test set.

thetic test set (STS), i.e. the synthetic images generated with the SIS starting from the real test set, and the synthetic additional test set (SATS), i.e. the translated images from the real additional test set. The semantic segmentation network has been trained to label each pixel in one of the 12 classes mentioned in 3.1. A mean IoU over all classes of 0.78 and 0.73 for the STS and the SATS, respectively, exhibits that the SPADE (Park et al., 2019) can generate realistic MIS images.

Moreover, in Figure 3 there are some synthetic images generated from the ground truth segmentation map of the test set. The conditioned generative network can effectively produce realistic laparoscopic images starting from the semantic map. In particular, even if some fine-grained details on the Monopolar Curved Scissors are missing (like the da Vinci surgery text) in the first row of Figure 3, the generated image (central column in Figure 3) has a quality comparable with the real one. In addition, in Figure 4 there are samples generated starting from the semantic map of the additional test set.

Furthermore, some label maps were generated using a painting software in order to test the network in different conditions. As can be seen in Figure 5, the model can generate texture maps from uniform segmentation maps (first two rows) or with a more elaborate map (last row). Even if in the majority of the training data there are more complex, and yet more informative semantic maps, the network can produce a plausible mapping from semantic segmentation label to texture.

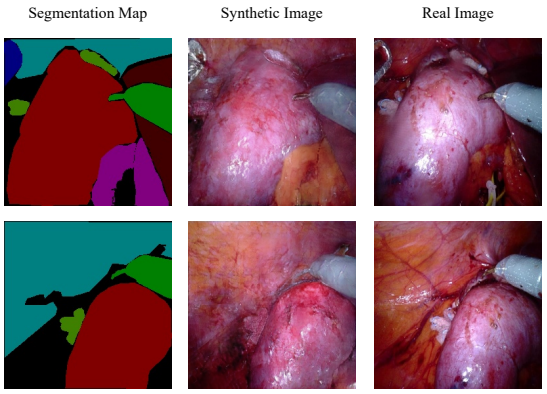

Figure 4: Some examples of synthetic images (centre) generated from the semantic segmentation map (left) in comparison with the corresponding real image (right) for the additional test set.

As said before, a U-Net architecture has been trained on synthetic images and then tested on the real test set (RTS) and the real additional test set (RATS). The network was trained with only two segmentation classes : kidney and surgical tools. In Table 2 there are the quantitative results for this experiment for the RTS and the RATS in terms of Dice Score mean IoU, Precision and Recall. In particular, the metrics are computed as mean over all the classes (O), only for the class associated with the kidney (K) and for the surgical tools (ST). Overall the synthetic images provide images with informative content that lets the model to generalize on real, previously unseen data.

In terms of mean IoU, the segmentation network reaches an overall value of 0.83 for the RTS and 0.82 for the RATS. Even if all the training was performed

Table 2: Quantitative evaluation of the U-Net semantic segmentation architecture trained on synthetic data over the real test set (RTS) and real additional test set (RATS). There are the average metrics and standard deviation computed over all classes (O), only for the kidney (K) and for the surgical tools (ST).

			mIoU	Dice	P	R
RTS	O	Mean	0.83	0.89	0.92	0.89
		Std	0.12	0.10	0.09	0.10
	K	Mean	0.84	0.90	0.91	0.91
		Std	0.19	0.16	0.15	0.16
	ST	Mean	0.81	0.88	0.92	0.87
		Std	0.16	0.12	0.10	0.15
RATS	O	Mean	0.82	0.89	0.95	0.85
		Std	0.11	0.08	0.05	0.11
	K	Mean	0.90	0.94	0.98	0.92
		Std	0.08	0.05	0.02	0.09
	ST	Mean	0.74	0.84	0.92	0.79
		Std	0.17	0.13	0.08	0.18

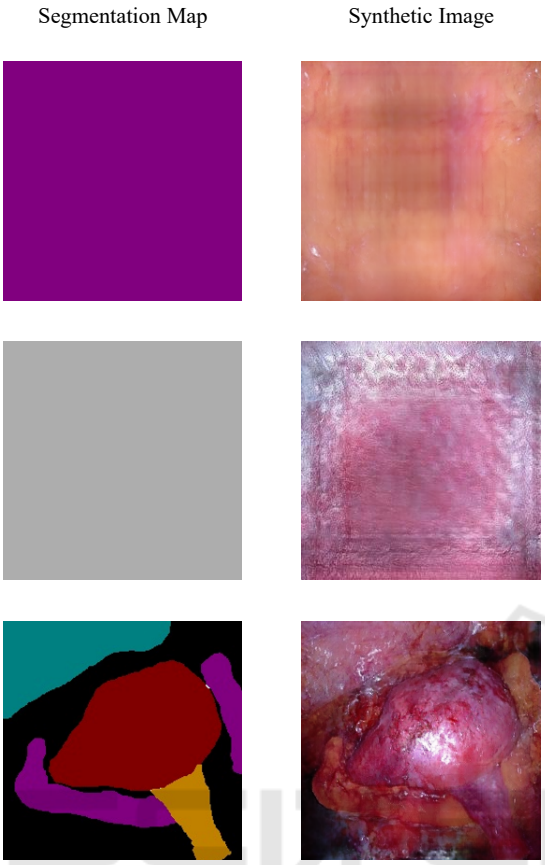

Figure 5: Synthetic images (right) generated from uniform semantic maps (first two rows) or more complex semantic maps (last row).

on synthetic images the network generalizes over real ones, as can be seen from qualitative results shown in Figures 6 and 7 where there are the input image (left) the predicted semantic map (centre) and the ground truth map (right). In red is shown the kidney while the surgical tools are in blue. Moreover, the inference time is about 9*ms*, therefore the segmentation network can be used for real-time applications. For example, to mask an instrument out so that the augmented reality overlay does not occlude the surgeon's view (Allan et al., 2019).

4.3 Discussion and Limitations

Providing the SIS network with a rich semantic map leads to a generation of realistic laparoscopic surgery images. Once the network is trained, it can be used to generate synthetic data that could be useful for the training of other networks, as shown by our experiment. Some networks rely on supervision signals that are difficult to obtain for minimally invasive surgeries, like the relative position of the organ

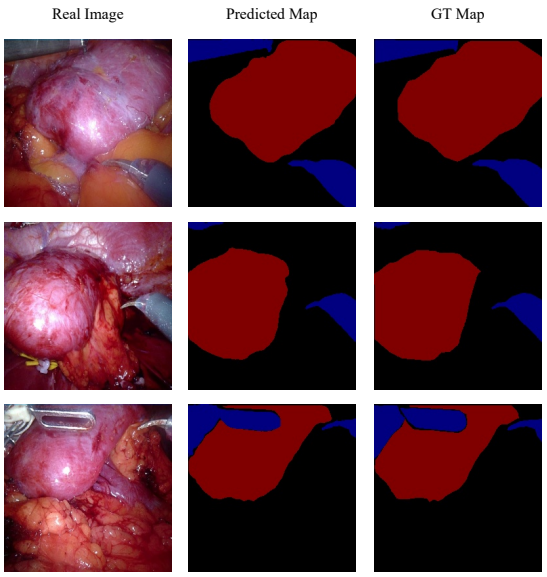

Figure 6: Qualitative results for the U-Net semantic segmentation architecture trained on synthetic data to predict the kidney (red) and surgical tools (blue) semantic map for the real test set (RTS). On the left is there is the real image, on the centre there is the predicted semantic map and on the right there is the ground truth (GT) semantic map.

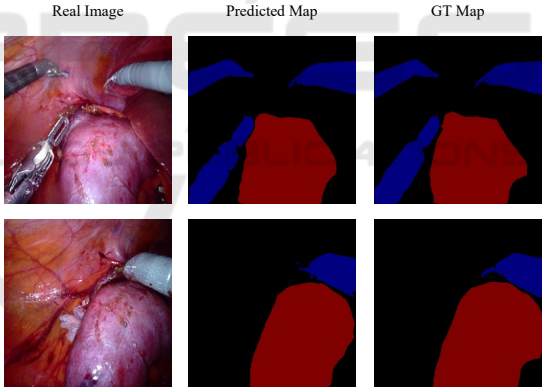

Figure 7: Qualitative results for the U-Net semantic segmentation architecture trained on synthetic data to predict the kidney (red) and surgical tools (blue) semantic map for the real additional test set (RATS). On the left is there is the real image, on the centre there is the predicted semantic map and on the right there is the ground truth (GT) semantic map.

with respect to the endoscopic camera. By leveraging on the semantic image translation, segmentation maps can be generated artificially from the patient-specific 3D anatomical mesh in order to virtually generate a semantic mask in a given position. Moreover, the surgical tool segmentation maps can be added later together with other semantic information, such as the presence of fat or abdominal tissue, in order to generate a realistic image where the 6-DoF pose

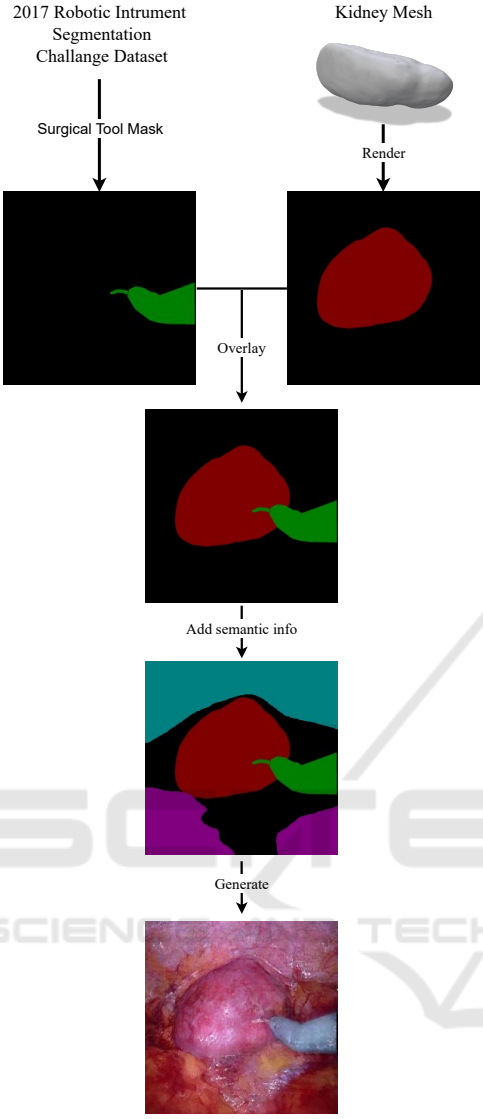

Figure 8: Synthetic data generation pipeline.

of the organ is known. An example can be seen in Figure 8, where the segmentation of the kidney has been generated from a patient-specific 3D kidney model. Then, the segmentation mask associated with the Monopolar Curved Scissors extracted from the 2017 robotic instrument segmentation challenge dataset (Allan et al., 2019) has been attached together with a rough semantic map of the fat and abdominal tissue. The generated data can be used to train networks that can retrieve the position of the organ and be used to improve Augmented Reality (AR) guided interventions in robotic-assisted urologic surgery (Bianchi et al., 2021; Schiavina et al., 2021b; Schiavina et al., 2021a; Tartarini et al., 2023).

The generative approach proposed for Minimally Invasive Surgery (MIS) images confronts several limitations. First of all, usually the images captured during robotic surgeries have higher resolution. Moreover, the dynamic nature of MIS environments introduces complexities such as smoke and motion blur, further complicating the generation process. To address these issues comprehensively, the dataset used to train the generative network should encompass diverse and challenging scenarios, including instances of smoke, varying degrees of motion blur, and other intricacies characteristic of MIS procedures.

5 CONCLUSIONS

A semantic image synthesis (SIS) method has been exploited to generate realistic Robotic Assisted Partial Nephrectomy (RAPN) surgical images. Starting from rich semantic maps we achieved highly realistic synthetic images that can be used to train neural networks. As future works, we aim to include the powerful multi-modal generation ability of Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) for the conditional image generation from semantic maps (Wang et al., 2022) in robotic surgery. Moreover, we intend to generate a synthetic dataset including 6-DoF pose estimation (Xiang et al., 2017) starting from patient-specific organ meshes and the presented SIS approach.

REFERENCES

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.-H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al. (2019). 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*.

Bianchi, L., Barbaresi, U., Cercenelli, L., Bortolani, B., Gaudiano, C., Chessa, F., Angiolini, A., Lodi, S., Porreca, A., Bianchi, F. M., et al. (2020). The impact of 3d digital reconstruction on the surgical planning of partial nephrectomy: a case-control study. still time for a novel surgical trend? *Clinical Genitourinary Cancer*, 18(6):e669–e678.

Bianchi, L., Chessa, F., Angiolini, A., Cercenelli, L., Lodi, S., Bortolani, B., Molinaroli, E., Casablanca, C., Droghetti, M., Gaudiano, C., et al. (2021). The use of augmented reality to guide the intraoperative frozen section during robot-assisted radical prostatectomy. *European Urology*, 80(4):480–488.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Marzullo, A., Moccia, S., Catellani, M., Calimeri, F., and De Momi, E. (2021). Towards realistic laparoscopic image generation using image-domain translation. *Computer Methods and Programs in Biomedicine*, 200:105834.

Ozawa, T., Hayashi, Y., Oda, H., Oda, M., Kitasaka, T., Takeshita, N., Ito, M., and Mori, K. (2021). Synthetic laparoscopic video generation for machine learning-based surgical instrument segmentation from real laparoscopic video and virtual surgical instruments. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 9(3):225–232.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346.

Pfeiffer, M., Funke, I., Robu, M. R., Bodenstedt, S., Strenger, L., Engelhardt, S., Roß, T., Clarkson, M. J., Gurusamy, K., Davidson, B. R., et al. (2019). Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*, pages 119–127. Springer.

Rau, A., Edwards, P. E., Ahmad, O. F., Riordan, P., Janatka, M., Lovat, L. B., and Stoyanov, D. (2019). Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *International journal of computer assisted radiology and surgery*, 14:1167–1176.

Rivoir, D., Pfeiffer, M., Docea, R., Kolbinger, F., Riediger, C., Weitz, J., and Speidel, S. (2021). Long-term temporally consistent unpaired video translation from simulated surgical 3d data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3343–3353.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. arxiv 2015. *arXiv preprint arXiv:1505.04597*.

Schiavina, R., Bianchi, L., Chessa, F., Barbaresi, U., Cercenelli, L., Lodi, S., Gaudiano, C., Bortolani, B., Angiolini, A., Bianchi, F. M., et al. (2021a). Augmented reality to guide selective clamping and tumor dissection during robot-assisted partial nephrectomy: a preliminary experience. *Clinical genitourinary cancer*, 19(3):e149–e155.

Schiavina, R., Bianchi, L., Lodi, S., Cercenelli, L., Chessa, F., Bortolani, B., Gaudiano, C., Casablanca, C., Droghetti, M., Porreca, A., et al. (2021b). Real-time augmented reality three-dimensional guided robotic radical prostatectomy: preliminary experience and evaluation of the impact on surgical planning. *European Urology Focus*, 7(6):1260–1267.

Tartarini, L., Riccardo, S., Bianchi, L., Lodi, S., Gaudiano, C., Bortolani, B., Cercenelli, L., Brunocilla, E., and Marcelli, E. (2023). Stereoscopic augmented reality for intraoperative guidance in robotic surgery. *Journal of Mechanics in Medicine and Biology*, page 2340040.

Vercauteren, T., Unberath, M., Padoy, N., and Navab, N. (2019). Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions. *Proceedings of the IEEE*, 108(1):198–214.

Wada, K. Labelme: Image Polygonal Annotation with Python.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807.

Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., and Li, H. (2022). Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*.

Xiang, Y., Schmidt, T., Narayanan, V., and Fox, D. (2017). Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*.

Yoon, J., Hong, S., Hong, S., Lee, J., Shin, S., Park, B., Sung, N., Yu, H., Kim, S., Park, S., et al. (2022). Surgical scene segmentation using semantic image synthesis with a virtual surgery environment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 551–561. Springer.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.