LC-SD: Realistic Endoscopic Image Generation with Limited Training Data

Joanna Kaleta [1,2*] **Diego Dall'Alba** [1,3] **Szymon Płotka** [1,4,5] **Przemysław Korzeniowski** [1]
[1]Sano Centre for Computational Medicine [2]Warsaw University of Technology [3]University of Verona
[4]University of Amsterdam [5]Amsterdam University Medical Center

Abstract

Computer-assisted surgical systems provide support information to the surgeon, which can improve the execution and overall outcome of the procedure. These systems are based on deep learning models that are trained on complex and challenging-to-annotate data. Generating synthetic data can overcome these limitations, but it is necessary to reduce the domain gap between real and synthetic data. We propose a method for image-to-image translation based on a Stable Diffusion model, which generates realistic images starting from synthetic data. Compared to previous works, the proposed method is better suited for clinical application as it requires a much smaller amount of input data and allows finer control over the generation of details by introducing different variants of supporting control networks. The proposed method is applied in the context of laparoscopic cholecystectomy, using synthetic and real data from public datasets. It achieves a mean Intersection over Union of 69.76%, significantly improving the baseline results (69.76% vs. 42.21%). The proposed method for translating synthetic images into images with realistic characteristics will enable the training of deep learning methods that can generalize optimally to real-world contexts, thereby improving computer-assisted intervention guidance systems.

1 Introduction

Computer-assisted intervention (CAI) is a research field focused on enhancing the safety, efficiency, and cost-effectiveness of medical procedures by minimizing errors and complications [15]. Within CAI, Laparoscopic Cholecystectomy (LC) has gained significant attention as a widely performed minimally invasive procedure for gallbladder removal [19]. However, LC presents technical challenges due to limited visibility and the use of laparoscopic instruments, leading to potential complications like bile duct injury (BDI) [31]. To address these complexities, CAI systems leveraging Deep Learning (DL) methods have been proposed. These systems aim to identify safe dissection zones, locate anatomical landmarks, and automatically assess critical safety criteria [14, 31]. DL techniques, including action-triplet recognition, temporal modeling, tools and anatomical structures segmentation, have been applied to LC [19, 35, 32, 7].

However, the availability of annotated data poses challenges for training DL models in this domain [22]. Limited datasets, primarily derived from the Cholec80 dataset, exist for LC, annotated with phases, tool presence, and action-triplets [32, 18, 19]. To overcome this limitation, generating synthetic data through virtual simulations along with rich annotations has been explored [4, 22]. Yet, DL models trained on synthetic data often struggle to perform well on real data due to the domain gap [22]. Image-to-image translation techniques based on Generative Adversarial Networks (GANs) have been proposed to mitigate this limitation [2]. However, these techniques still require a substantial amount of annotated data.

*Corresponding authors: joanna.kaleta.dokt@pw.edu.pl

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

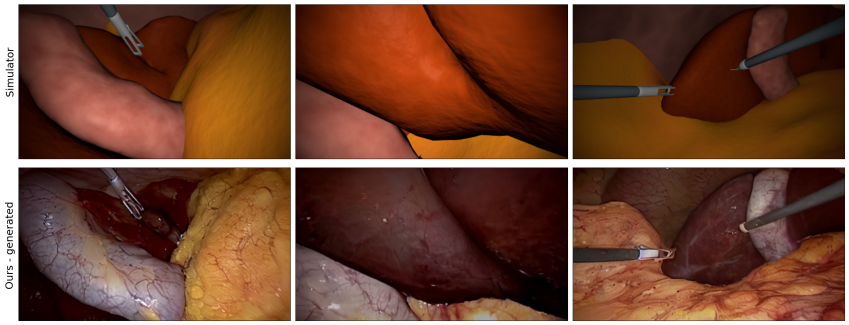

Figure 1: Examples of synthetic data translated with our fine-tuned model to the CholecT45 style. Three random frames from the simulator and their realistic translations are shown in the top and bottom rows, respectively.

Recently, Latent Diffusion Models (LDMs) have shown promise in generating highly detailed images while preserving semantic structure [9]. LDMs employ an iterative process involving noise addition and reverse learning to recover original data. In the medical field, LDMs have been utilized for tasks such as image translation, generation, preprocessing, segmentation, and classification [9]. Compared to other DL techniques like GANs, LDMs can be fine-tuned effectively with smaller datasets and combined with support methods for controlled generation. The widely used Stable Diffusion (SD) LDM model offers efficient conditioning of the generation process through text prompts [24].

In general, no existing work uses LDMs instead of GANs for the translation of synthetic images into realistic images with limited training data. Therefore, our contributions are as follows: (1) We introduce a novel application of the Stable Diffusion Model to generate synthetic surgical data in an unsupervised manner, addressing the issue of limited data availability in clinical environments. (2) We evaluate our approach using public datasets to demonstrate its effectiveness in generating realistic synthetic data. The results show that our approach outperforms the baseline method in preserving tissue integrity, achieving a mean Intersection over Union (IOU) of 69.76% compared to 42.21% for the CholecT80-styled baseline. Additionally, our method successfully captures the characteristic feature distribution of real surgical data, either comparable to or enhanced compared to the baseline dataset, (3) We provide public access to the code and our realistic rendering of the publicly available IRCAD dataset, which includes simulation frames, depth maps, segmentation maps, edges, and normals (`https://github.com/SanoScience/sim2real_with_Stable_Diffusion`).

2 Related work

Several approaches have been proposed for generating realistic synthetic data for surgical procedures, e.g., [10]. In [4], Unity3D was used to create a 3D liver and laparoscopic environment to generate training images for DL segmentation. Another study [16] used a GAN to generate images from segmentation maps, emphasizing instrument-anatomy differences. Synthetic images combined with real segmentation maps have trained GANs for tool segmentation using techniques like consistency losses and student-teacher learning [27, 26]. GANs have also been applied in cardiac intervention, colonoscopy, and sinus surgery [29, 20, 13]. Another relevant work [22] introduced image-to-image translation for 3D LC anatomy rendering from real endoscopic images. It used GANs trained in an unpaired manner, generating 100,000 annotated images. Though extended to video translation [23], it lacks tools and gallbladder-specific content, not representing LC procedures. While GANs show potential, they have limitations, including early discriminator convergence and unstable adversarial loss, leading to mode collapse and reduced diversity. LDMs are emerging as an alternative, excelling in computer vision tasks [3]. In the medical domain, LDMs find use in various applications, such as generating MRI sequences and histological images [9, 21, 17]. The Stable Diffusion (SD) model is notably used in similar medical tasks. Our work pioneers the use of LDMs for intra-operative endoscopic image generation, conditioning on text prompts and virtual simulator images.

3 Method

Our method involves adding a concept to the SD model and using it to generate realistic images from synthetic ones. We begin by fine-tuning SD based on Dreambooth (DB) [25]. Then, the fine-tuned Laparoscopic Cholecystectomy Stable Diffusion (LC-SD) model is employed to generate realistic

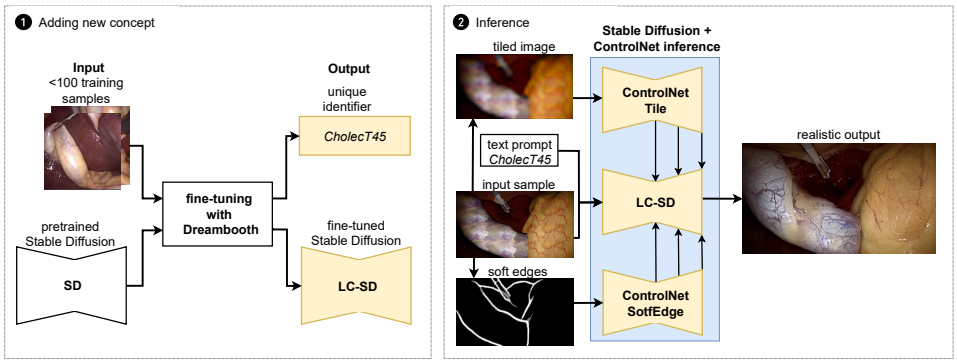

Figure 2: Overview of fine-tuning and inference: SD is fine-tuned using Dreambooth, which binds a unique text identifier with a newly incorporated CholecT45 style. In the inference stage, the fine-tuned LC-SD is conditioned with two ControlNet models. The textured input sample, along with the 'cholecT45' prompt, is passed to LC-SD. The Tile ControlNet accepts a tiled version of the input sample, while the SoftEdge ControlNet accepts edges detected by Pidinet from the input sample.

images. This is achieved by leveraging two versions of the ControlNet support architecture, namely Tile and SoftEdge control, to ensure consistency between label and generated images. An overview of the proposed method is depicted in Fig. 2.

Fine-tuning with Dreambooth. SD [24] is a LDM, uses a lower-dimensional latent space for denoising, with flexibility to condition on text or images via cross-attention. Notable progress in few-shot fine-tuning and personalized concept introduction for SD includes Textual-Inversion [5], Low Rank Adaptation [8], Custom Diffusion [11], and DB [25]. DB is selected for fine-tuning since it allows to add unconventional concepts using a small set of concept-specific images (3 to 5 for an object, 50 to 200 for a style). During training, a new concept is bound with an unique text identifier.

Inference with ControlNet. For realistic tissue generation, we use text-guided image-to-image inference, incorporating a unique text identifier bound to the CholecT45 style during DB training. We impose additional control over generated samples using ControlNet - an architecture designed for controlling pre-trained large LDMs by integrating conditions like sketches, key points, edges, and segmentation maps [36]. It is possible to combine multiple ControlNet models using the extended formula from [36], given in Appendix A. We explored various outputs contained in IRCAD dataset (depth maps, segmentation maps, normal vectors) as control inputs. However, preliminary tests showed unsatisfactory inference results. Instead, we utilized ControlNet models delivering more robust control for the CholecT45 style: SoftEdge and Tile. SoftEdge primarily preserves original edges and tissue folds, utilizing edges from Pidinet [30] or HED [34] models, while Tile ControlNet effectively adds tissue details and helps preserve accurate tissue colors. Selected control types are compared in Appendix B.

4 Experiments

To use the minimum amount of data while ensuring a sufficient variability of visual properties and the presence of all regions and instruments of interest, we trained three separate models, each based on two distinct videos from the CholecT45 dataset [19]. We carefully select pairs of videos that exhibit comparable visual characteristics and ensure that all classes are represented within each training set. We train each model with DB using a manually selected set of 85, 91, and 95 images, respectively. We further discuss the experimental setup in Appendix C. To evaluate the realism of the generated data, we employ established evaluation metrics [23, 9]: Frechet Inception Distance (FID) [6] and Kernel Inception Distance (KID) [1]. Following [33, 38, 12], we employ Learned Perceptual Image Patch Similarity (LPIPS) [37] to assess diversity of generated samples. Moreover, to evaluate the LC-SD models' ability to preserve labels, we fine-tuned a variant of U-Net with a pre-trained ResNet50 backbone using the CholecSeg8k dataset [7] for five classes present in both the CholecSeg8k and IRCAD datasets. We use mean Intersection over Union (mIoU) to calculate the average overlap between predicted and ground truth segmentation masks across multiple classes. Each metric was calculated for 10,000 samples. We exclude all training videos from CholecT45 dataset for evaluation.

Table 1: Quantitative results are presented for each style: the raw simulator, baseline data, and our generated data. We demonstrate this using mIoU, FID, KID, and LPIPS metrics. The best-performing methods are bolded.

Method	**FID** $\downarrow$	**KID** $\downarrow$	**LPIPS**$_{VGG}$ $\uparrow$	**mIoU [%]** $\uparrow$
Raw simulation images	305.00	.3739 $\pm$ .0041	.5820	24.73
[22] - Random	110.92	.1243 $\pm$ .0035	.5834	45.28
[22] - Cholec80	67.13	.0623 $\pm$ .0017	**.6407**	42.21
Ours - CholecT45 vid52 & vid56	68.35	.0658 $\pm$ .0015	.6245	66.85
Ours - CholecT45 vid25 & vid66	63.07	.0582 $\pm$ .0012	.6262	**69.76**
Ours - CholecT45 vid01 & vid49	57.47	.0513 $\pm$ .0011	.6175	67.20
Ours - Mixed styles	**54.57**	**.0473 $\pm$ .0011**	.6281	67.89

Table 2: The mIoU [%] values for different control types and improvement compared to no-control inference are presented.

Style	No control	Only SoftEgde	Only Tile	SoftEdge + Tile
CholecT45 vid52 & vid56	61.52	65.26 (+6.1%)	64.20 (+4.4%)	**66.85 (+8.7%)**
CholecT45 vid25 & vid66	63.35	67.16 (+6.0%)	68.01 (+7.4%)	**69.76 (+10.1%)**
CholecT45 vid01 & vid49	54.29	63.26 (+16.5%)	62.08 (+14.3%)	**67.20 (+23.8%)**

5 Results

Raw simulation images have an mIoU of 24.73%, FID of 305.00, KID of 0.3739 ± 0.0041, and LPIPS of 0.5820. Using the method from [28], we observe performance improvements across all metrics - mIOU of 42.21%, accompanied with FID of 67.13 and KID of 0.0623 ± 0.0017 and LPIPS of 0.6407 for Cholec80 style. Our method, denoted as "ours", showcases further enhancements. Across various CholecT45 styles, we achieve impressive mIoU scores, notably 69.76% for vid25 & vid66 style. These results are accompanied by competitive FID, KID, and LPIPS scores. Additional experiment demonstrates that mixing images from three styles further decreases FID and KID values. Overall, our method signifcantly outperforms the baseline method in terms of label preservation expressed through mIoU, while capturing the characteristic feature distribution on a comparable or better level. Additionally, Table 2 provides an overview of the mIoU values for different control types and their improvements compared to no-control inference. Across all CholecT45 styles, control mechanisms consistently boosted mIoU compared to no-control inference. The combination of SoftEdge and Tile controls delivered the most substantial improvement, demonstrating their effectiveness in enhancing model performance across diverse styles.

6 Discussion and Conclusions

In this work, we have proposed an SD-based approach to generate realistic surgical images from virtual simulator images and text prompts. The SD model was initially fine-tuned using DB and then used for inference, supported by Tile and SoftEdge ControlNets. The model can be trained using less than 100 real images without manual annotations and manages to generate realistic images that either outperform or are highly competitive to the baseline in all considered evaluation metrics. We consider this work to be a significant addition to the current foundation, offering researchers a valuable dataset to facilitate the development of machine learning solutions in image-guided and robotic surgery. This approach can produce fully labeled training data for supervised machine learning algorithms. Additionally, strict alignment of the created data with its ground truth annotations extends its potential for evaluation in various unsupervised and semi-supervised applications. Despite that, our method has some limitations, including the need for careful image selection due to a small training dataset, heavy reliance on input image features, and a lack of temporal consistency in simulated data. Overall, our proposed method represents a promising direction for generating realistic surgical images and has the potential to contribute to advancements in the field of image-guided and robotic surgery.

References

[1] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

[2] Y. Chen, X.-H. Yang, Z. Wei, A. A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, and Q. Guan. Generative adversarial networks in medical image augmentation: A review. *Computers in Biology and Medicine*, 144:105382, May 2022.

[3] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[4] T. Dowrick, B. Davidson, K. Gurusamy, and M. J. Clarkson. Large scale simulation of labeled intraoperative scenes in unity. *International Journal of Computer Assisted Radiology and Surgery*, 17(5):961–963, 2022.

[5] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, 2017.

[7] W.-Y. Hong, C.-L. Kao, Y.-H. Kuo, J.-R. Wang, W.-L. Chang, and C.-S. Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[9] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

[10] P. Korzeniowski, S. Płotka, R. Brawura-Biskupski-Samaha, and A. Sitek. Virtual reality simulator for fetoscopic spina bifida repair surgery. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 401–406. IEEE, 2022.

[11] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1941, June 2023.

[12] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[13] S. Lin, F. Qin, Y. Li, R. A. Bly, K. S. Moe, and B. Hannaford. Lc-gan: Image-to-image translation based on generative adversarial network for endoscopic images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2914–2920. IEEE, 2020.

[14] A. Madani, B. Namazi, M. S. Altieri, D. A. Hashimoto, A. M. Rivera, P. H. Pucher, A. Navarrete-Welton, G. Sankaranarayanan, L. M. Brunt, A. Okrainec, and A. Alseidi. Artificial intelligence for intraoperative guidance: Using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Annals of Surgery*, 276(2), 2022.

[15] L. Maier-Hein, M. Eisenmann, D. Sarikaya, K. März, T. Collins, A. Malpani, J. Fallert, H. Feussner, S. Giannarou, P. Mascagni, et al. Surgical data science–from concepts toward clinical translation. *Medical image analysis*, 76:102306, 2022.

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

[17] P. A. Moghadam, S. Van Dalen, K. C. Martin, J. Lennerz, S. Yip, H. Farahani, and A. Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2000–2009, 2023.

[18] C. I. Nwoye, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 364–374, 2020.

[19] C. I. Nwoye, T. Yu, C. Gonzalez, B. Seeliger, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

[20] M. Oda, K. Tanaka, H. Takabatake, M. Mori, H. Natori, and K. Mori. Realistic endoscopic image generation method using virtual-to-real image-domain translation. *Healthcare Technology Letters*, 6(6):214–219, 2019.

[21] M. Özbey, O. Dalmaz, S. U. Dar, H. A. Bedel, Ş. Özturk, A. Güngör, and T. Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023.

[22] M. Pfeiffer, I. Funke, M. R. Robu, S. Bodenstedt, L. Strenger, S. Engelhardt, T. Roß, M. J. Clarkson, K. Gurusamy, B. R. Davidson, et al. 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*, pages 119–127. Springer, 2019.

[23] D. Rivoir, M. Pfeiffer, R. Docea, F. Kolbinger, C. Riediger, J. Weitz, and S. Speidel. Long-term temporally consistent unpaired video translation from simulated surgical 3d data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3323–3333, 2021.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[25] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, June 2023.

[26] M. Sahu, A. Mukhopadhyay, and S. Zachow. Simulation-to-real domain adaptation with teacher–student learning for endoscopic instrument segmentation. *International journal of computer assisted radiology and surgery*, 16(5):849–859, 2021.

[27] M. Sahu, R. Strömsdörfer, A. Mukhopadhyay, and S. Zachow. Endo-sim2real: Consistency learning-based domain adaptation for instrument segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 784–794. Springer, 2020.

[28] P. M. Scheikl, E. Tagliabue, B. Gyenes, M. Wagner, D. Dall'Alba, P. Fiorini, and F. Mathis-Ullrich. Sim-to-real transfer for visual reinforcement learning of deformable object manipulation for robot-assisted surgery. *IEEE Robotics and Automation Letters*, 8(2):560–567, 2022.

[29] L. Sharan, G. Romano, S. Koehler, H. Kelm, M. Karck, R. De Simone, and S. Engelhardt. Mutually improved endoscopic image synthesis and landmark detection in unpaired image-to-image translation. *IEEE Journal of Biomedical and Health Informatics*, 26(1):127–138, 2021.

[30] Z. Su, W. Liu, Z. Yu, D. Hu, Q. Liao, Q. Tian, M. Pietikainen, and L. Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5117–5127, 2021.

[31] T. Tokuyasu, Y. Iwashita, Y. Matsunobu, T. Kamiyama, M. Ishikake, S. Sakaguchi, K. Ebe, K. Tada, Y. Endo, T. Etoh, et al. Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. *Surgical endoscopy*, 35:1651–1658, 2021.

[32] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[33] Z. Wang, L. Zhao, H. Chen, L. Qiu, Q. Mo, S. Lin, W. Xing, and D. Lu. Diversified arbitrary style transfer via deep feature perturbation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7786–7795, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[34] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[35] B. Zhang, B. Goel, M. H. Sarhan, V. K. Goel, R. Abukhalil, B. Kalesan, N. Stottler, and S. Petculescu. Surgical workflow recognition with temporal convolution and transformer for action segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 18(4):785–794, 2023.

[36] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[37] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[38] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 465–476, Red Hook, NY, USA, 2017. Curran Associates Inc.

A ControlNet

ControlNet, designed for controlling pre-trained large DMs, maintains two sets of UNet weights: a locked copy preserving original weights and a trainable copy fine-tuned using task-specific data. Neural network blocks of pre-trained DM and ControlNet use trainable "zero convolution" layers, and detailed functionality and application are explained in [36].

It is possible to directly apply ControlNet trained on original SD to LC-SD, and even to combine multiple ControlNet models (each with desired strength) to impose diversified control. To combine single LC_SD block with corresponding blocks of N ControlNets we present the extended formula from [36] as:

$$\boldsymbol{y}_{\mathrm{c}} = \mathcal{F}(\boldsymbol{x}; \boldsymbol{\theta}_{LC\_SD}) + \sum_{i=1}^{N} w_i \mathcal{Z}\left(\mathcal{F}\left(\boldsymbol{x} + \mathcal{Z}\left(\boldsymbol{c}_i; \boldsymbol{\theta}_{Z1,i}\right); \boldsymbol{\theta}_{C,i}\right); \boldsymbol{\theta}_{Z2,i}\right) \tag{1}$$

where $\boldsymbol{x}$ is an input feature map to the LC-SD block, $\boldsymbol{c}_i$ is a conditioning input feature map to the corresponding block of the i-th trained ControlNet and $\boldsymbol{y}_{\mathrm{c}}$ is a conditioned output feature map from the LC-SD block. We denote the weights of the LC-SD block as $\boldsymbol{\theta}_{\boldsymbol{LC\_SD}}$ and the trainable weights for the block of the i-th ControlNet as $\boldsymbol{\theta}_{\boldsymbol{C,i}}$. The function denoted as $\mathcal{F}(\cdot;\cdot)$ transforms the input feature map into the output feature map given a set of parameters. We denote the "zero convolution" operation as $\mathcal{Z}(\cdot;\cdot)$. Within the block of the i-th ControlNet two "zero convolution" operations are performed with optimized parameters $\{\boldsymbol{\theta}_{Z1,i}, \boldsymbol{\theta}_{Z2,i}\}$, respectively. w_i is the strength the i-th ControlNet is applied with. The first term on the right side of Eq.1 represents the result of applying LC-SD, while the second term relates to the contribution of the different ControlNets.

B ControlNet Influence

For inference we use two ControlNet models which influence is shown in Fig.3. The SoftEdge control utilizes edges generated with Pidinet [30] or HED [34] models. It primarily preserves original edges and tissue folds. On the other hand, Tile ControlNet exhibits conceptual similarities with tile-based super-resolution models but offers broader applications. It operates in two modes: generating new details while ignoring existing ones, and ignoring global prompts when local tile semantics and prompts do not align, guiding the diffusion process with local context. In the context of endoscopic image generation, Tile ControlNet effectively adds tissue details and helps preserve accurate tissue colors.

C Dataset and Implementation details

We train our models on manually selected small subsets from the CholecT45 dataset [19]. Despite the limited number of images, it is crucial to choose representative and consistent samples that cover various procedure stages and tissues present in the synthetic dataset. Furthermore, to prevent the models from introducing tool artifacts in each frame, it is highly important to include images both with surgical tools and with minimal or no presence of them. All models are based on Stable Diffusion v1.5 and we train them with DB using a learning rate of 1×10^{-6} and a batch size of 4 for 2,000 steps.

In the inference stage, we utilize fully labeled synthetic data from the IRCAD 3D CT liver dataset. The dataset contains 20,000 synthetic images rendered from 3D scenes obtained from the CT data of 10 different patients including models of the liver, gallbladder (only for 6 patients), insufflated abdominal wall, fat, and connective tissue. In addition, tools, light sources, and endoscopic cameras have been added in random positions in the scene. In the image-to-image approach, the prior information significantly influences the resulting image. However, the IRCAD dataset presents simplified anatomy, and as a result, plain structures and distorted colors can lead to unrealistic results. To address this issue, we enhance the raw simulation images by incorporating texture information from example samples. We extract small texture samples for each tissue from the corresponding training set and blend them with the raw simulation scenes, guided by segmentation maps, as shown in Fig. 5.

For inference, we adjust the model checkpoint, denoising strength, classifier-free guidance scale (CFG), noise scheduler, and ControlNet v1.1. strengths for each LC-SD model separately. Although

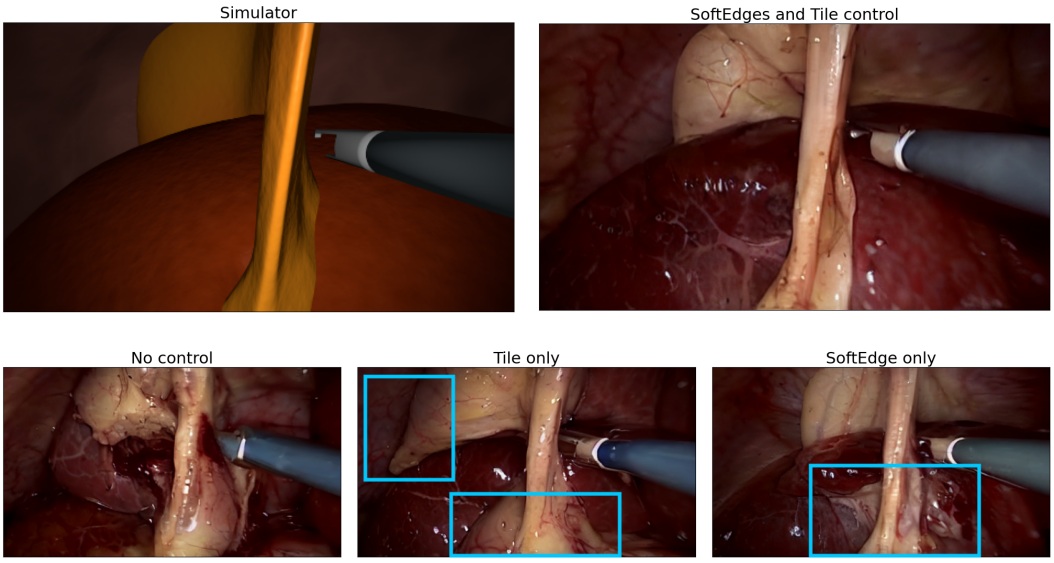

Figure 3: Visual comparison on example image generated with different control types applied. Without any type of control, the overall image consistency is degraded. With Tile control details are clearly rendered, but tissue shapes do not correspond to the given labels. With only SoftEdge control color artifacts appeared.

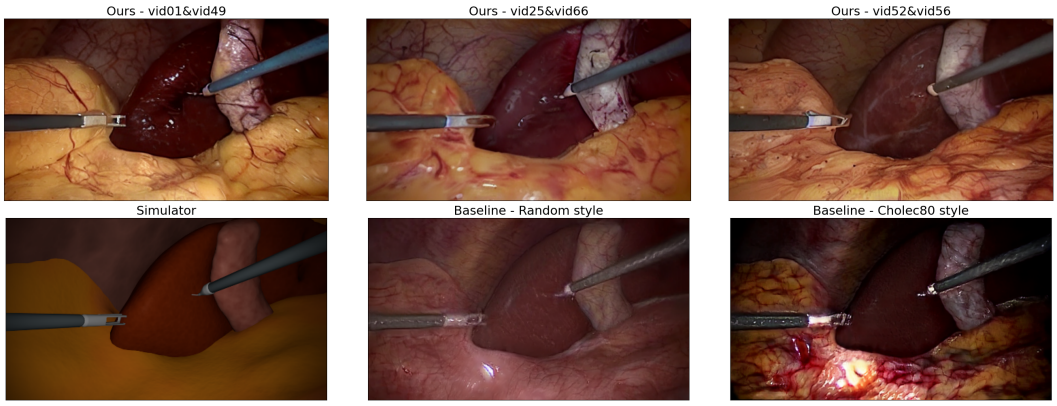

Figure 4: A visual comparison is made between the data processed using our fine-tuned Stable Diffusion model (first row), the raw image from the simulator, and the data generated in [22] for random and Cholec80 styles (second row).

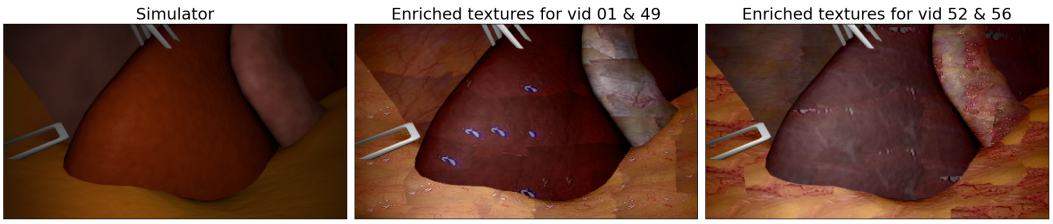

Figure 5: Visual comparison between the raw image from the simulator (first from left) and image with enriched textures for two styles (second and third from left).

all models are trained with the same parameters, variations in the complexity and diversity of the training sets resulted in differences in denoising capabilities across the models. We carefully balance the ControlNet strengths for each model separately. In addition to tissue placement, we also consider overall image realism and details, such as tissue folds. The lack of tissue folds would not necessarily

Table 3: Selected inference parameter values for each model: denoising strength, CFG, noise scheduler, SoftEdge, and Tile control strength. All the models use noise scheduler DPM++ 2M Karras.

Style	Denoising	CFG	SoftEdge	Tile
CholecT45 vid52 & vid56	0.45	4.5	0.5	0.3
CholecT45 vid25 & vid66	0.45	5.0	0.4	0.3
CholecT45 vid01 & vid49	0.5	5.0	0.55	0.3

degrade mIoU. To achieve the desired balance, we use a stronger SoftEdge control in combination with a weaker Tile control. Using only SoftEdge with high control strength could compromise image quality by erasing valuable details. To prevent Tile control from introducing excessive detail based on the input sample, we use a smaller strength. To generate data at a large scale while maintaining reasonable inference time and acceptable image quality, we limit the denoising steps to 20. The selected parameter values are shown in Table 3.

The data generation process is carried out on a single NVIDIA A100 GPU. The generation time takes up to 3 seconds per image, depending on the number of ControlNets utilized.