Supplementary material: Data Augmentation for Surgical Scene Segmentation with Anatomy-Aware Diffusion Models

1. Diffusion training details

We use the Stable Diffusion v1.5 inpainiting model as our baseline diffusion model. The training parameters for the CholecSeg8K dataset are shown in Tab. 1. The dataset consists of 5080 images for training, 2000 images for testing and 1000 images as validation set. We resized all the images to 512x512. Depending upon our initial experiments, we noticed for the abdominal wall that the generated images either suffered from creating the correct texture or having semantic leakage i.e., texture of other organs being replaced. One reason could be the variance in the lighting conditions. To account for this, be opted for the v-prediction loss from [7]. This has shown to improve image quality in low lighting. For the HeiSurf dataset, we used the parameters as the CholecSeg8K dataset. Similarly, we used the diffusion model intialized from the CholecSeg8K dataset to fine-tune for the HeiSurf dataset. We opted for this strategy as the number of images were low in HeiSurf dataset and theer also existed overlap of the different anatomies.

All the diffusion models were trained 1500 steps. The generated images were evaluated at every 500 steps using different metrics and visual examination. The pre-trained soft edge ControlNet was used to control the anatomy shape during the inference process. We sampled images from the inpainted model with pre-trained CN with DDIM [8] scheduler using 30 steps. A similar process was used for the DSAD dataset where a guidance scale of 5.5 was used for all the organs. For this dataset, we did not use the v-prediction method for training any specific organ. To reduce the overhead in the inference pipeline, we used the fast MultiStep [11] sampler in Stage-4 with SDEdit for image enchancement. We opted for $5 - 8$ steps for inconsistency removal process. Since the HeiSurf dataset had smaller image of images, we trained only the ControlNet-SE and T2i-Adapter-CY models on this dataset.

Text prompts For our approach, we used simple text prompts like *an image of abdominal wall in cholec* for the abdominal wall images in the CholecSeg8K dataset. Similarly, for the other organs and datasets we exchanged the organ name and the dataset name accordingly. For the SD model used in Stage-4, the text prompt was chosen as *an image in cholec* for an image in CholecSeg8K.

Organ	Pred-type	Gd. scale
Abdominal wall	v-prediction	0.6
Fat	ϵ-prediction	5.0
Liver	ϵ-prediction	6.0
Gall bladder	ϵ-prediction	5.5
Ligament	ϵ-prediction	5.0

Table 1. The parameters used for training and sampling from the CholecSeg8K dataset.

To train the ControlNet and T2i-Adapters pre-trained SD model is needed. We experimented with different text prompts. Intially, we used the same prompts from our method like *an image of cholec surgery* to train the SD model. This model was then used to train the ControlNet and T2i models. For their training, we again used the same prompts as the SD model. We noticed that the generated images lacked quality and did not correspond well to the conditioning masks. Hence, we used the segmentation masks to extract the classes present and constructed the prompt like *an image of cholec surgery with abdominal wall, liver and gall bladder with a hook*. We train the the SD model with such prompts and use similar prompts for training the ControlNet and T2i models. We found the best results with such expressive prompts rather than just mentioning a prompt like *an image of cholec surgery*. We hypothesize such prompts are necessary to make the model explicitly understand the different organs present in the scene. It is to noted that extra effort in constructing such prompts were necessary to train the baseline models in comparison our model which works on simpler text prompts. As we had limits on our training infrastructure, we did not train the text encoder of these models. As a future work we intend to train the train encoder along with the diffusion models to scope their performance on image quality.

Training scheme	Unet++ [13]		
	Dice(↑)	IOU (↑)	HD (↓)
No-aug	$0.74_{\pm 0.03}$	$0.65_{\pm 0.01}$	$126.08_{\pm 3.02}$
Color-aug	$0.76_{\pm 0.01}$	$0.66_{\pm 0.02}$	$118.98_{\pm 1.32}$
Color+spatial-aug	$0.79_{\pm 0.01}$	$0.69_{\pm 0.01}$	$88.86_{\pm 9.93}$
Implicit label [2]	$0.22_{\pm 0.05}$	$0.11_{\pm 0.01}$	$347.33_{\pm 9.12}$
Implicit label + Real	$0.77_{\pm 0.04}$	$0.67_{\pm 0.03}$	$97.44_{\pm 2.75}$
Only *Syn*	$0.44_{\pm 0.03}$	$0.34_{\pm 0.01}$	$132.63_{\pm 4.16}$
Syn + Implicit label + Real	$0.75_{\pm 0.02}$	$0.65_{\pm 0.03}$	$93.82_{\pm 1.87}$
SS-Syn + Real	$0.80_{\pm 0.03}$	$0.70_{\pm 0.02}$	$102.01_{\pm 3.34}$
Syn + Real	$\mathbf{0.82_{\pm 0.01}}$	$\mathbf{0.72_{\pm 0.01}}$	$\mathbf{85.27_{\pm 1.04}}$

Table 2. The segmentation scores on the DSAD dataset. The best scores are indicated in bold.

2. Segmentation training details

To train the baselines on different augmentation schemes, we collected and experimented with multiple methods. We curated a set of color and spatial augmentations based on prior works that focussed on medical (surgical) domain [1–3]. We used the following augmentations:grid/elastic distortion, perspective change, RGB channel shift, ColorJitter, blur, hue, contrast & brightness, maskdropout. We tuned the hyperparameters included in each of these methods to attain the best scores. Similarly, for spatial transformations we used perspective change, grid distortion, rotation, random flipping. For the combined (color+spatial) augmentations, we chose the best combination via experimentation with different combinations of methods mentioned before. To find the best combinations of methods and hyperparameters, we conducted experiments on each dataset separately.

3. Additional results

The segmentation results on the DSAD dataset with Unet++ architecture is shown in Tab. 2. Similarly, for the CholecSeg8K dataset, we also trained the UperNet-small model. The results are shown in Tab. 3. The seg. scores from different image synthesis models for the HeiSurf dataset is shown in Tab. 4.

Auxillary surgical task. We used the generated datasets to train models for another surgical task: *surgical target prediction*. We used the CholecT50 [5] as it forms a part of CholecSeg8K and DSAD datasets to show the capability of our *Syn* dataset in multi-class and multi-label classification tasks. The training and test splits were maintained throughout to avoid any data leakage. Tab. 5 shows that our *Syn* datasets proves useful beyond segmentation tasks.

The results in Tab. 6 shows using UniPc [12] scheduler with 20 sampling steps. This leads to the inference time of 4.07s in comparison to 5.25s. We also notice that the downstream performance of the generated images matches that of the DDIM scheduler.

Training scheme	UperNet-small		
	Dice(↑)	IOU (↑)	HD (↓)
No-aug	$0.55_{\pm 0.01}$	$0.47_{\pm 0.02}$	$118.37_{\pm 6.62}$
Color-aug	$0.57_{\pm 0.01}$	$0.45_{\pm 0.04}$	$115.80_{\pm 1.38}$
Color+spatial-aug	$0.62_{\pm 0.02}$	$0.51_{\pm 0.01}$	$108.63_{\pm 1.51}$
Only *Syn*	$0.56_{\pm 0.01}$	$0.45_{\pm 0.01}$	$111.71_{\pm 0.62}$
SS-Syn + Real	$\mathbf{0.69_{\pm 0.01}}$	$\mathbf{0.53_{\pm 0.02}}$	$\mathbf{95.76_{\pm 2.49}}$
Syn + Real	$0.67_{\pm 0.01}$	$\mathbf{0.53_{\pm 0.01}}$	$105.90_{\pm 5.28}$

Table 3. The segmentation scores on the CholecSeg8K dataset. The best scores are indicated in bold.

Method	Dice (↑)	IOU (↑)	HD(↓)
SPADE [6]	$0.39\pm_{0.01}$	$0.27_{\pm 0.01}$	$252.70_{\pm 7.17}$
SPADE-vae [6]	$0.39_{\pm 0.01}$	$0.28_{\pm 0.02}$	$234.89_{\pm 6.15}$
Pix2Pix-HD [9]	$0.39_{\pm 0.03}$	$0.27_{\pm 0.02}$	$236.35_{\pm 8.73}$
ControlNet-SE [10]	$0.40_{\pm 0.02}$	$0.27_{\pm 0.03}$	$224.28_{\pm 5.02}$
T2I-adapter-CY [4]	$0.38_{\pm 0.01}$	$0.26_{\pm 0.01}$	$234.97_{\pm 8.32}$
Ours-*SS-Syn*	$0.47_{\pm 0.01}$	$0.33_{\pm 0.01}$	$170.63_{\pm 2.19}$
Ours-*Syn*	$\mathbf{0.49_{\pm 0.01}}$	$\mathbf{0.36_{\pm 0.01}}$	$\mathbf{165.42_{\pm 3.04}}$

Table 4. **Segmentation eval. on HeiSurf dataset.** Our synthetic datasets outperforms other models.

Training method	CholecT50		DSAD	
	F1(↑)	Accuracy(↑)	F1(↑)	Accuracy(↑)
Real with cl+sp aug.	$0.50_{\pm 0.05}$	$0.52_{\pm 0.04}$	$0.42_{\pm 0.002}$	$0.83_{\pm 0.001}$
Ours-*Syn*+Real	$\mathbf{0.64_{\pm 0.01}}$	$\mathbf{0.63_{\pm 0.02}}$	$\mathbf{0.45_{\pm 0.001}}$	$\mathbf{0.86_{\pm 0.001}}$

Table 5. Surgical target prediction results on two datasets.

Training method	Scheduler		CholecSeg8K		HeiSurf		DSAD	
	DDIM	UniPC	Dice(↑)	IOU(↑)	Dice(↑)	IOU(↑)	Dice(↑)	IOU(↑)
Only *Syn*	✓		$0.53_{\pm 0.01}$	$0.41_{\pm 0.02}$	$0.35_{\pm 0.02}$	$0.24_{\pm 0.01}$	$0.60_{\pm 0.03}$	$0.51_{\pm 0.01}$
		✓	$0.51_{\pm 0.02}$	$0.39_{\pm 0.01}$	$0.34_{\pm 0.01}$	$0.24_{\pm 0.03}$	$0.58_{\pm 0.01}$	$0.50_{\pm 0.01}$
Ours-*Syn* + Real	✓		$0.68_{\pm 0.01}$	$0.56_{\pm 0.01}$	$0.49_{\pm 0.01}$	$0.36_{\pm 0.01}$	$0.83_{\pm 0.01}$	$0.74_{\pm 0.01}$
		✓	$0.66_{\pm 0.02}$	$0.55_{\pm 0.03}$	$0.49_{\pm 0.01}$	$0.37_{\pm 0.01}$	$0.82_{\pm 0.01}$	$0.74_{\pm 0.02}$

Table 6. Inf. time comparison with different schedulers.

The additional qualitative results from our method on CholecSeg8K, HeiSurf and DSAD datasets are shown in Figs. 1 to 3 respectively. Fig. 4 and Fig. 5 show the comparison of images with and without the image enhancement stage. For conditioning the ControlNet we use edge images extracted from the segmentation mask. A comparison is shown in Fig. 6. The per-organ evaluation scores on the three datasets are shown in Figs. 7 to 9. We see consistent improvement across different organs when combing our generated datasets with real images.

References

[1] Evgin Goceri. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56(11):12561–12605, 2023. 2

[2] Alexander C. Jenke, Sebastian Bodenstedt, Fiona R. Kol-

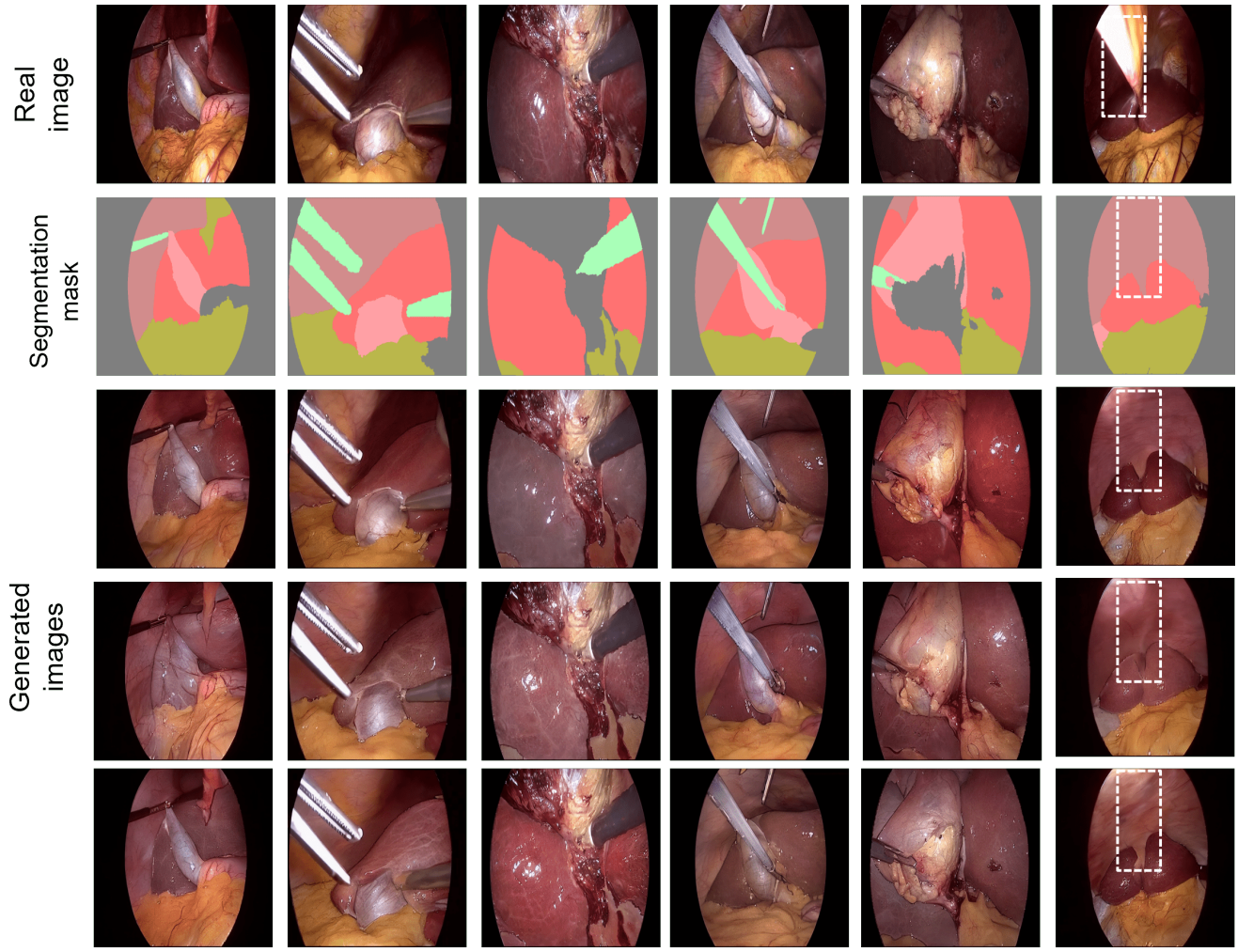


Figure 1. The generated images from the CholecSeg8K dataset. The different organs generated by our method followed by fusion creates images looking similar to real images. The diversity in the texture of the generated organs are quite visible in these images(zoom into the images to see the texture difference). Column $3, 4, 5$ clearly shows the difference in the liver and gall bladder textures. In the 6^{th} column we see that the generated images differ (indicated in white box) to the real image. This is because the ligament (bright yellow organ) in the real image was not labeled due to the camera angle and the light source and rather a common label of abdominal wall was indicated. Since our approach uses the segmentation masks for generating the organs, the ligament is not generated in the images, which does not affect downstream performance as the generated image still corresponds to the label. We see this an avenue for future work rather than a limitation. Using surgical simulations, either ligament or new organs can be generated using our diffusion approach, wherein only one model needs to be trained on that specific organ.

binger, Marius Distler, Jürgen Weitz, and Stefanie Speidel. One model to use them all: Training a segmentation model with complementary datasets, 2024. 2

[3] Fiona R Kolbinger, Franziska M Rinner, Alexander C Jenke, Matthias Carstens, Stefanie Krell, Stefan Leger, Marius Distler, Jürgen Weitz, Stefanie Speidel, and Sebastian Bodenstedt. Anatomy segmentation in laparoscopic surgery: comparison of machine learning and human expertise–an experimental study. *International Journal of Surgery*, 109(10):2962–2974, 2023. 2

[4] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. 2

[5] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022. 2

[6] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive nor-

Figure 2. The generated images from the HeiSurf dataset. The real images are shown as representative example, as there exists many different textures of the organs. Our approach is capable of generating different textures for each organ while maintaining the spatial consistency. Our method is capable of generating gall bladders (green color in segmentation mask) which is not completely covered within the fat tissue (1^{st} and 2^{nd} column). One failure case is indicated in white box. The texture of the generated liver tissue differs slightly from the real images. The real image in the 6^{th} column contains blood on the liver. Our generated images do not synthesize blood pools and could serve effectively as an augmentation method to improve segmentation. Adjusting the CFG scale would be method to rectify this case.

malization, 2019. 2

[7] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 1

[8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 1

[9] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 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, 2018. 2

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[11] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models, 2023. 1

[12] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[13] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation, 2018. 2

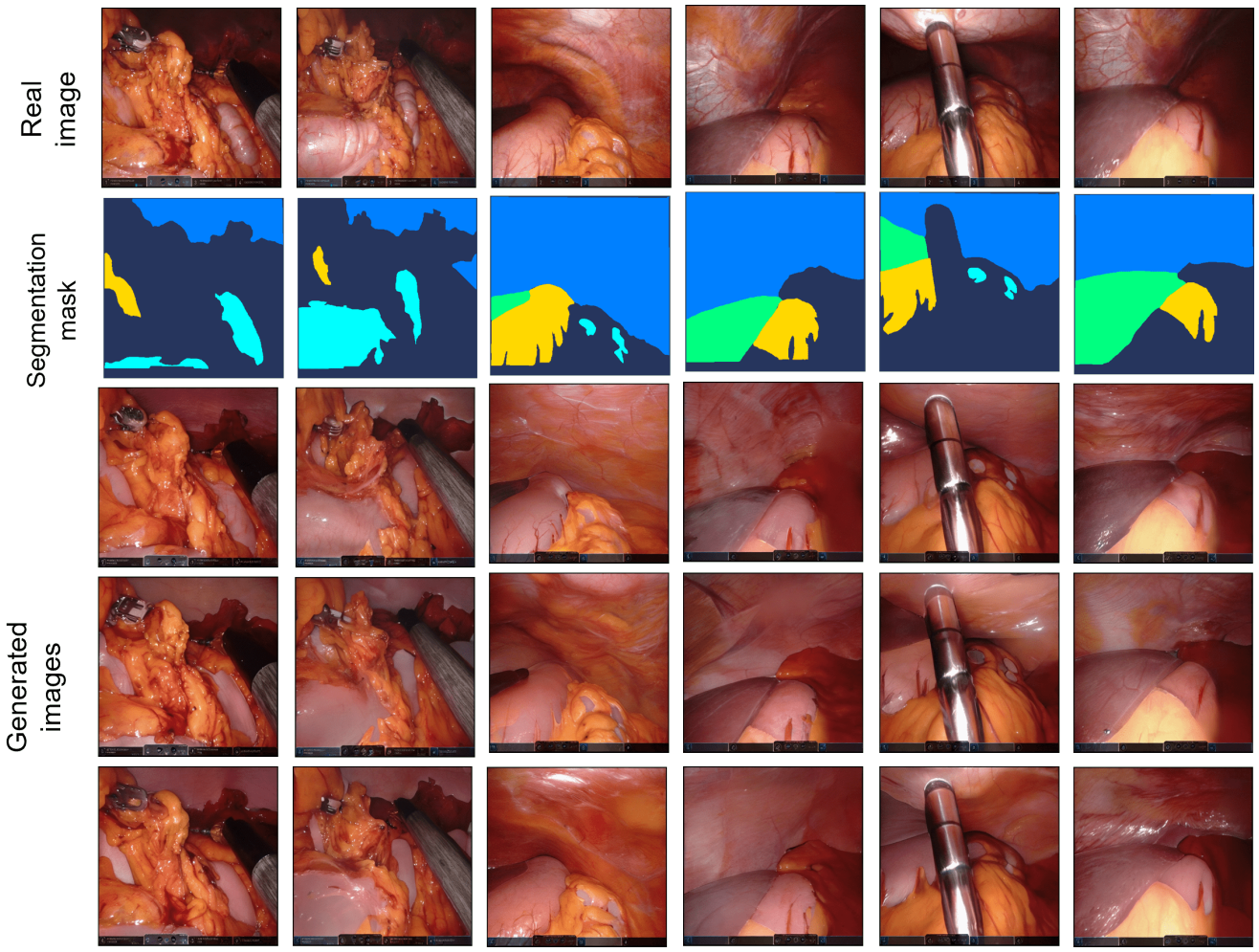


Figure 3. The generated images from the DSAD dataset. The generated images resemble the texture characteristics of the real images. Especially for the liver and stomach (indicated in green and gold color in segmentation mask), we see that the generated images maintain the texture well and adds finer details like vessels (column 3). It is to be noted that that binary datasets were utilized to generate the organs in this case. This results shows the particular importance of our approach that only real binary datasets can be used to generate multi-class datasets.

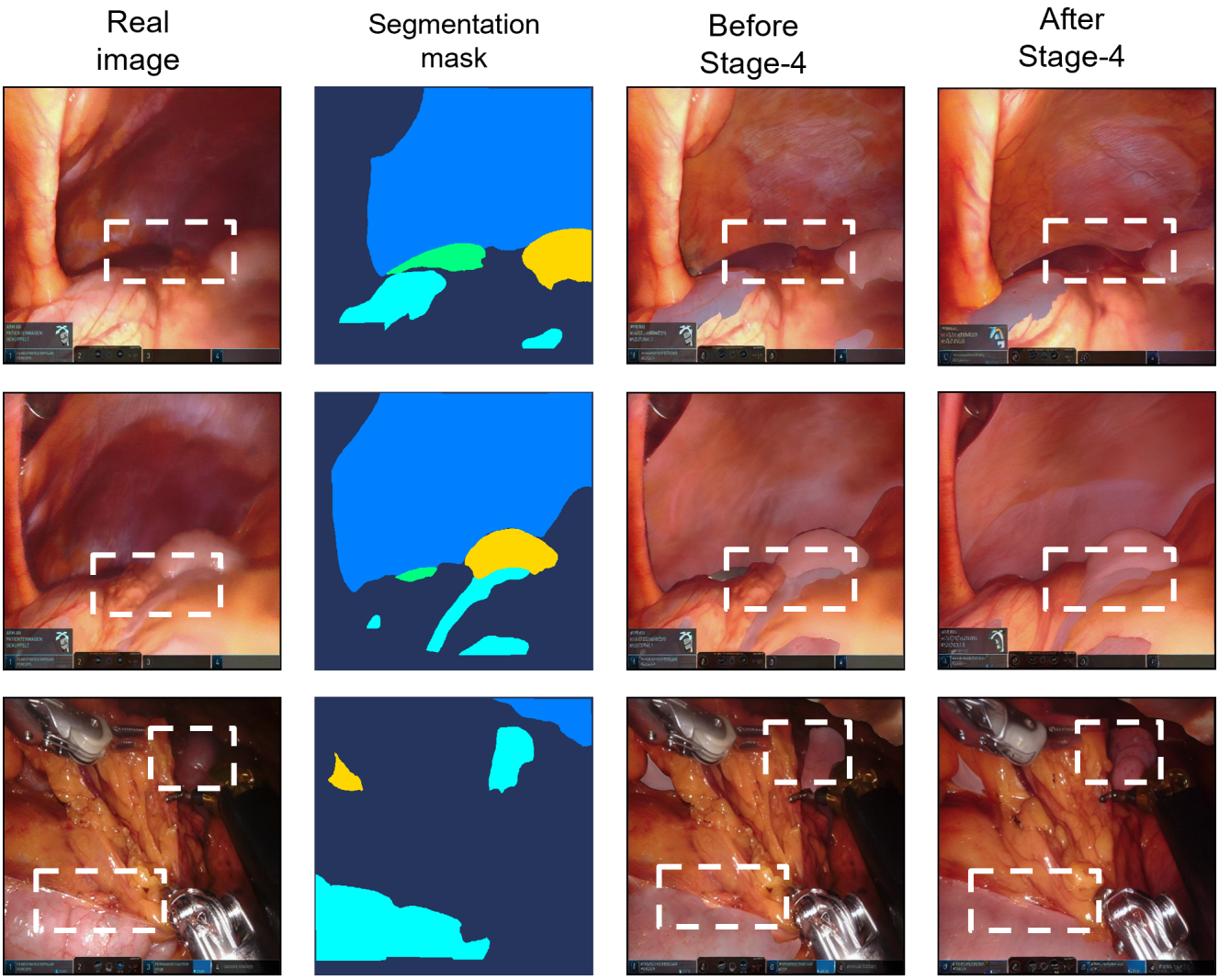


Figure 4. The images generated using the DSAD dataset with and without the Stage-4 in our pipeline. The Stage-4 is an image enhancement process that removes the inconsistencies from the image fusion stage. The white boxes indicate the regions comparing the difference between the real image, image after Stage-3 (3^{rd} column) and image generated after Stage-4. Clearly, fusing the images creates a junction between the different organs. There also exists a slight difference in the background lighting of the generated images from Stage-3 (3^{rd} column). To remove them these inconsistencies, the images are processed via a SDEdit method combined with SD model. We use the SD because the model is already aware of the texture of such surgical images. In process leads to a smoother junction between the organs which resembles like the real images.

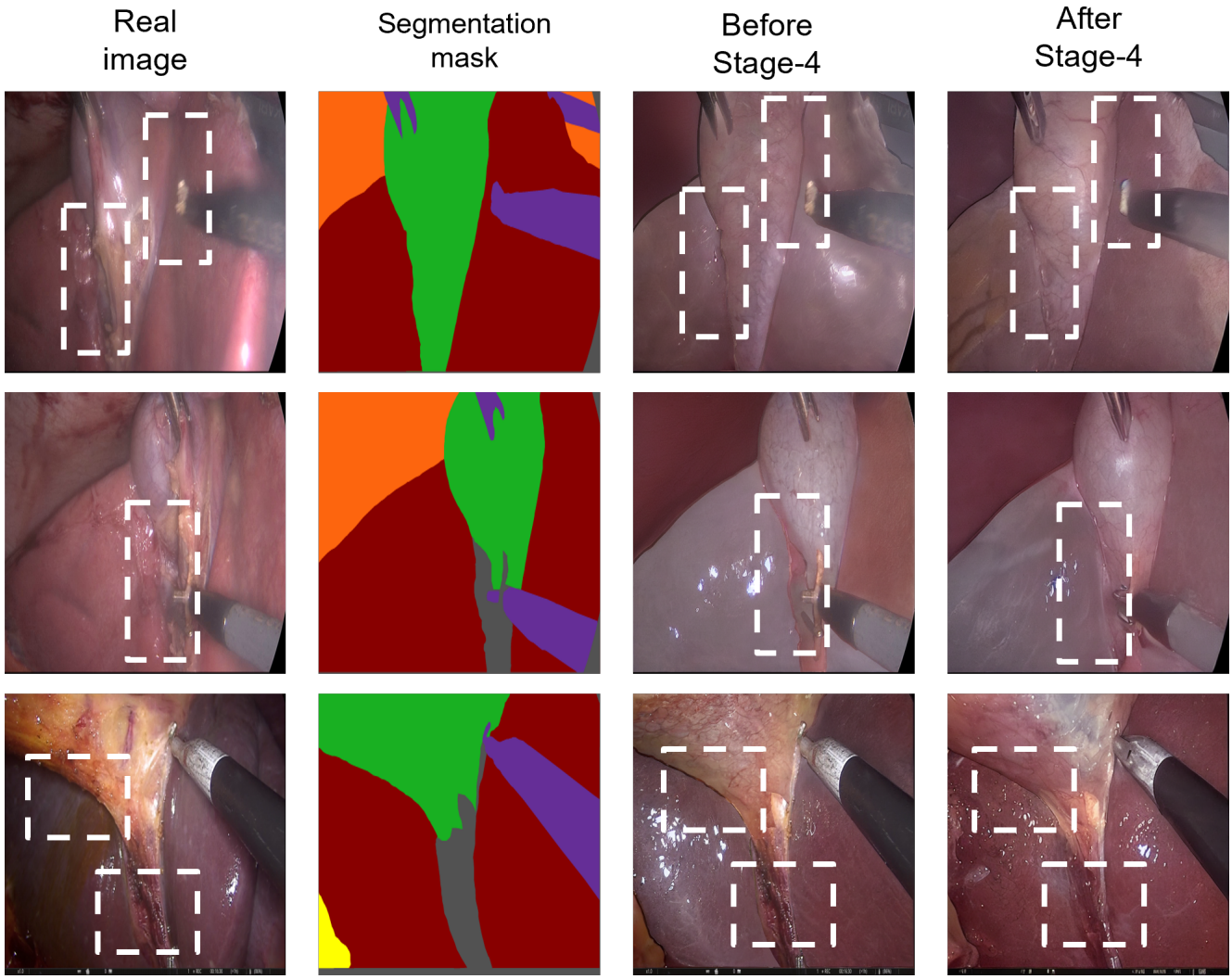


Figure 5. The generated images from the HeiSurf dataset before and after Stage-3. In the 1^{st} column, we noticed that the images ater Stage-4 had finer details like vessels added to the gall bladder. This is advantageous as it enhances the real texture of organs. Additionally, the edges between the organs are smoothened, which is similar to the real images.

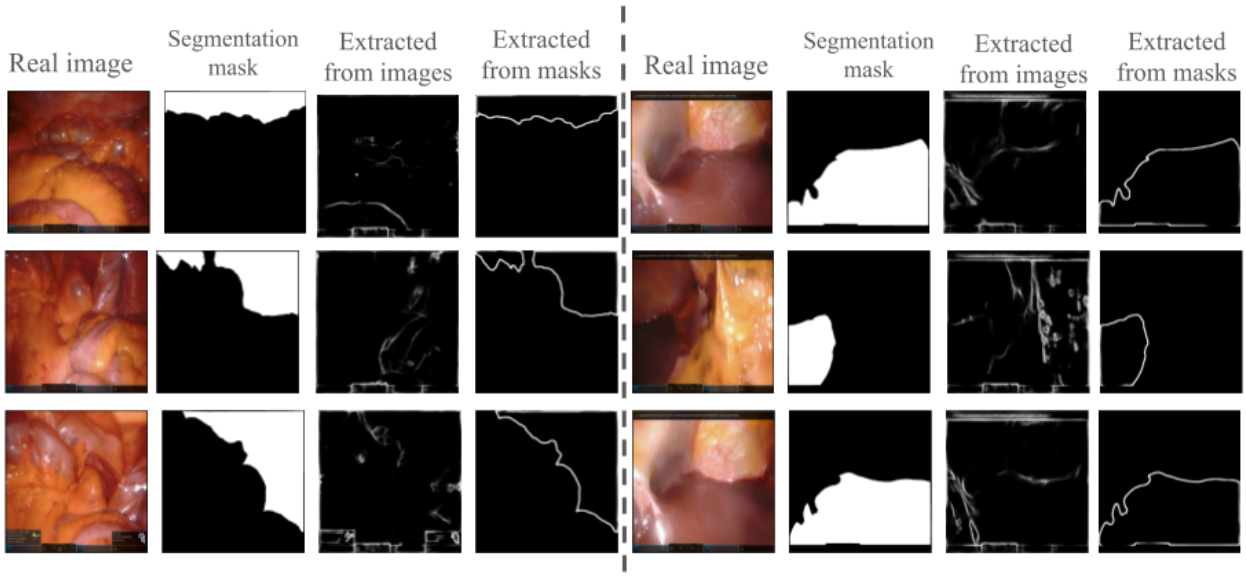
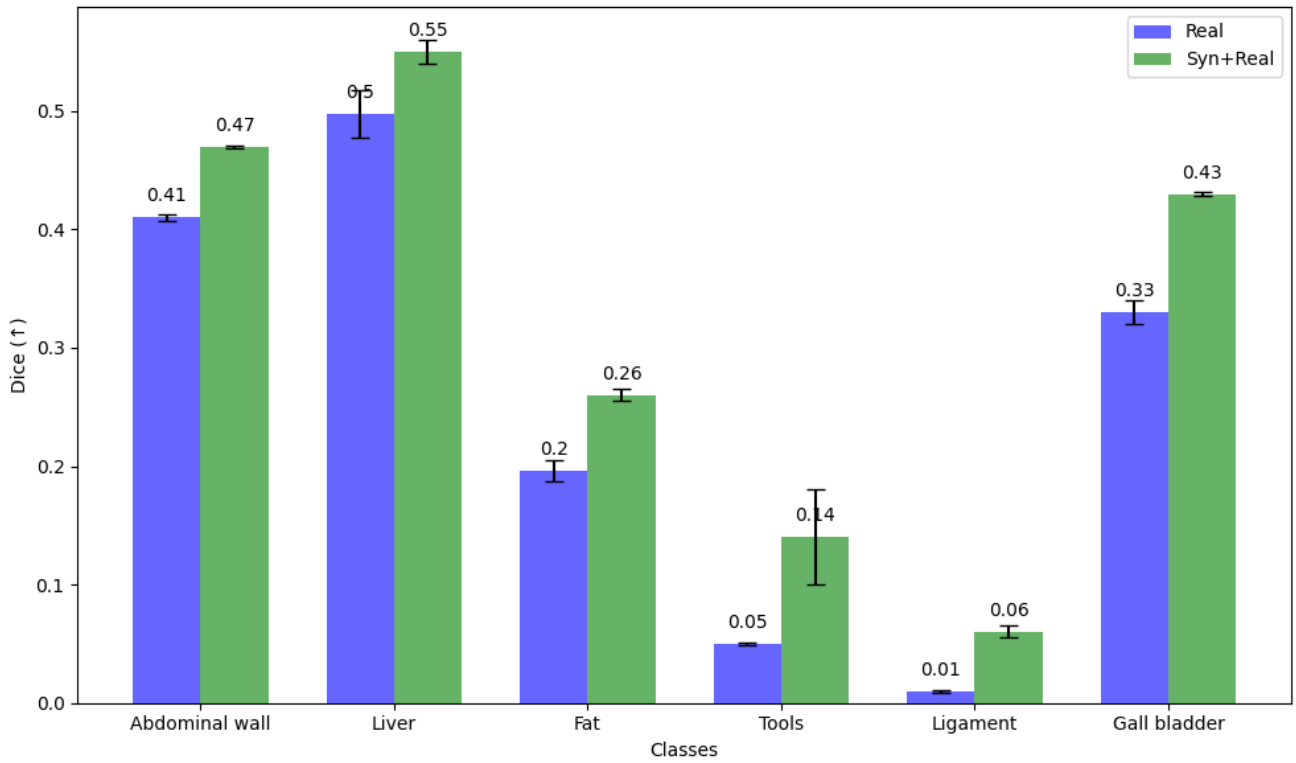
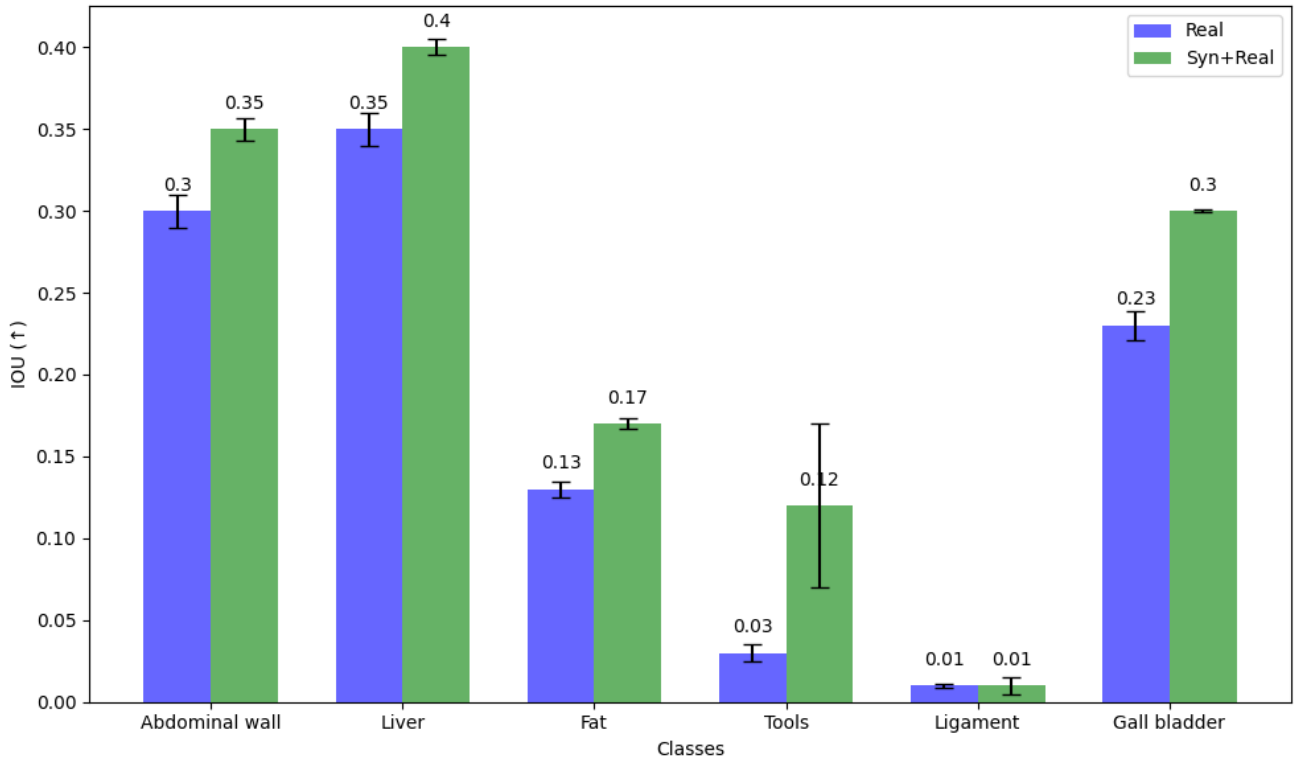


Figure 6. The conditioning signal for the pre-trained ControlNet model in Stage-2 of our approach is edge images. Naturally, these edges can be extracted from the real images using a edge detector. However, as shown in column 3 and 7, the extracted edges include the tools and other edges which does not correspond to the particular organ. Using such an extracted edge images would lead to inconsistent generation of the organ. Hence, we used the segmentation mask as the input to the extract the edges. As seen in column 4 and 8, the extracted edges correspond better to the segmentation mask. In our method, we simultaneously use the same segmentation mask to mask the region for inpainting and also to extract the conditioning signals for the ControlNet.

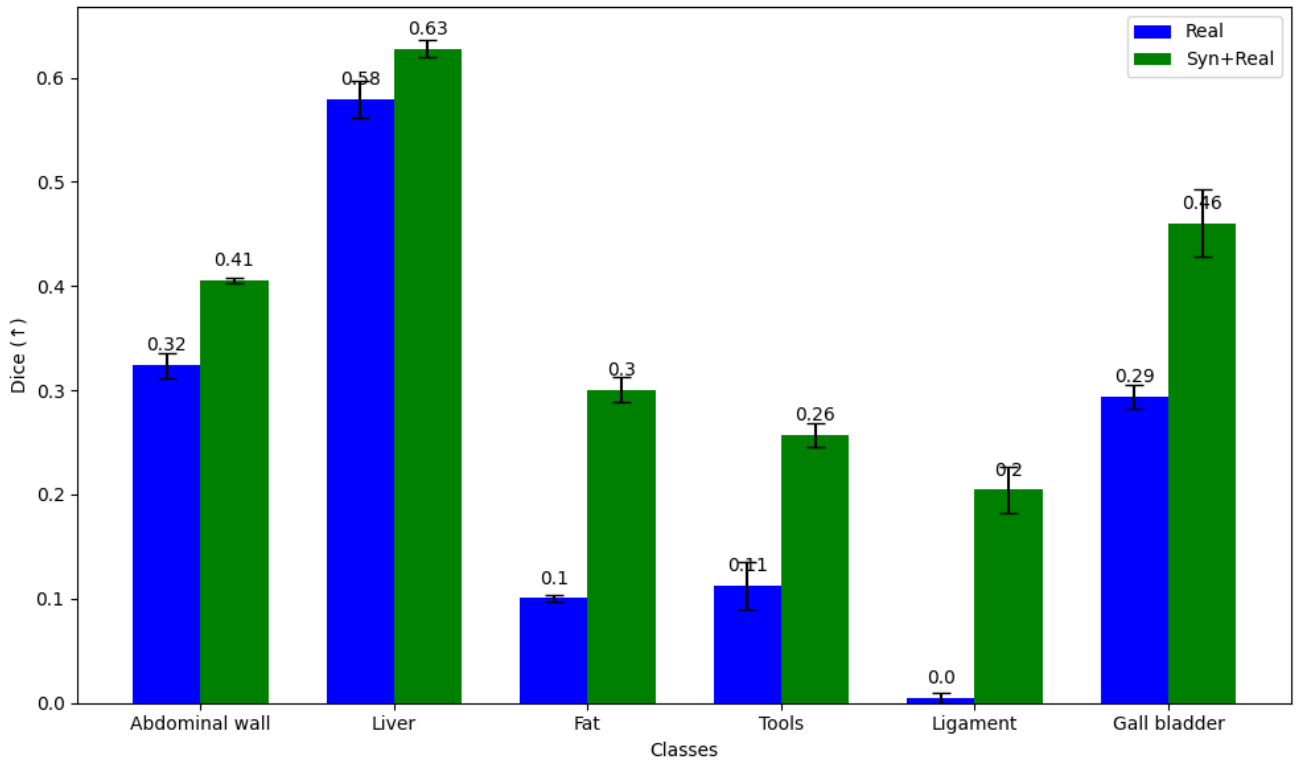


(a) The dice score on the Cholec80 dataset.

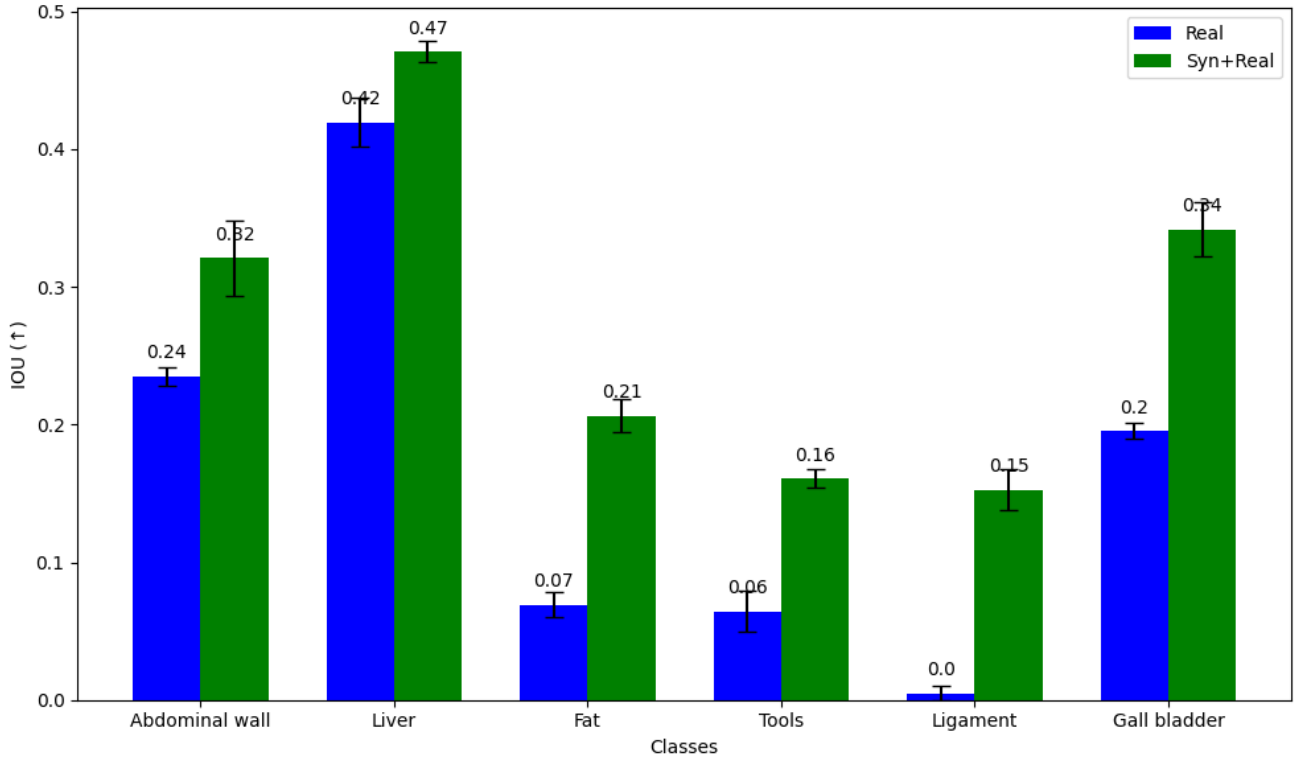


(b) The IOU score on the Cholec80 dataset.

Figure 7. The dice and IOU scores for each organ on the Cholec80 dataset. Adding our *Syn* datasets clearly show an imporvement in scores across each organ. Especially, the ligament and gall bladder seems to be segmented particularly well once our *Syn* datasets are added.

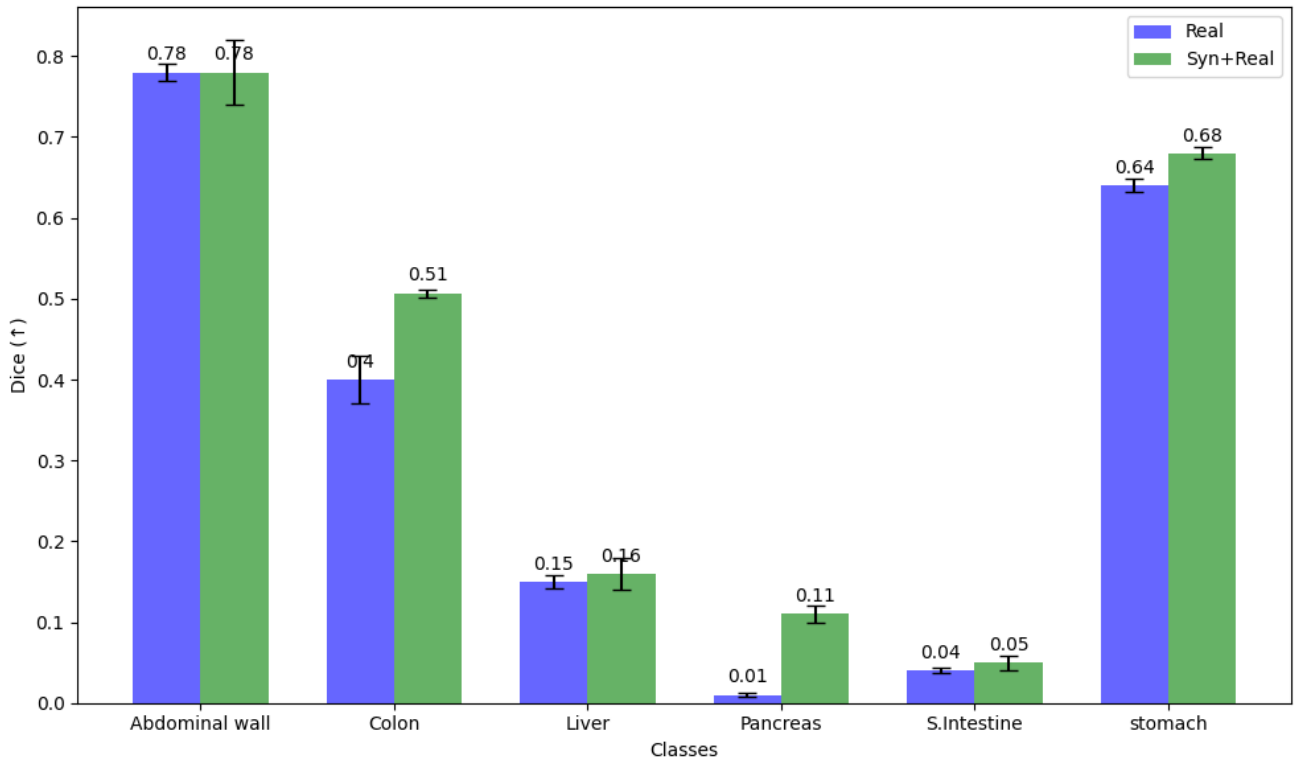


(a) The dice score on the HeiSurf dataset.

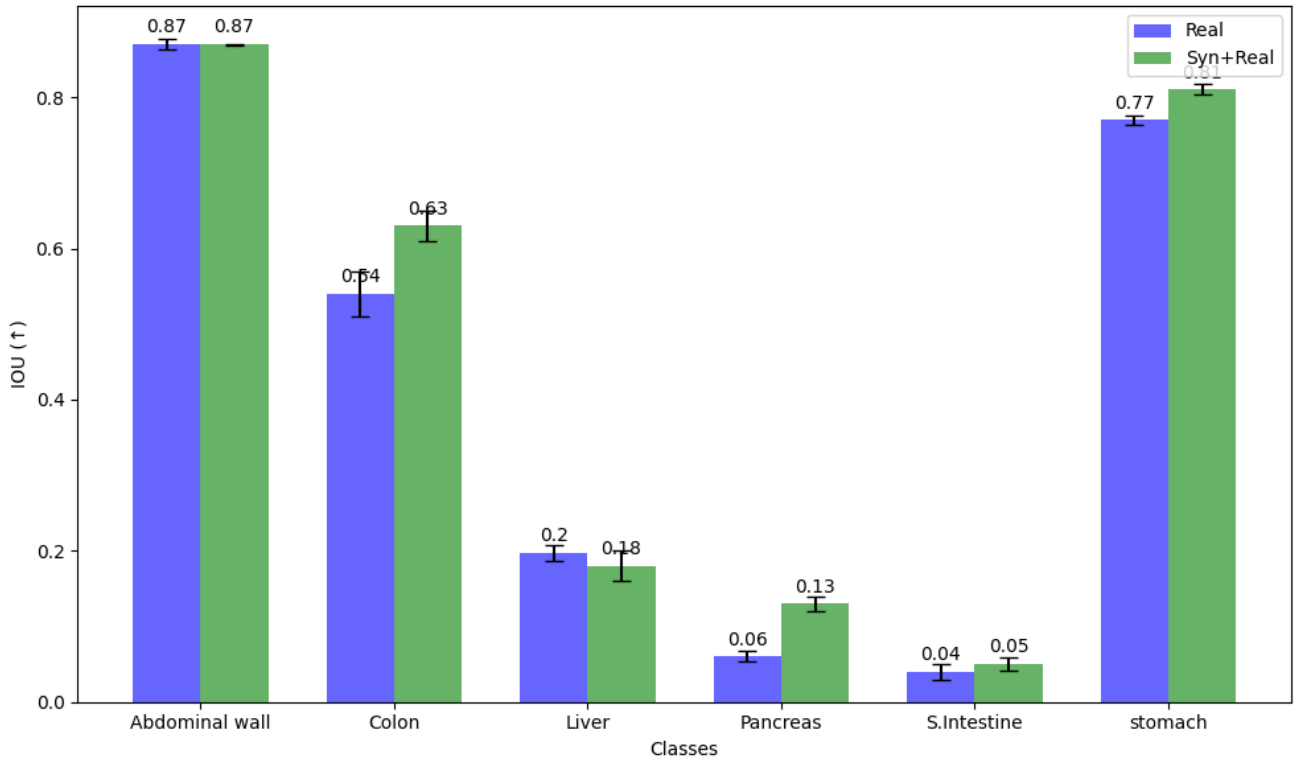


(b) The IOU score on the HeiSurf dataset.

Figure 8. The dice and IOU scores for each organ on the HeiSurf dataset. It is evident that combining our *Syn* datasets leads to improved segmentation across six different classes.



(a) The dice score on the DSAD dataset.



(b) The IOU score on the DSAD dataset.

Figure 9. The dice and IOU scores for each organ on the DSAD dataset. We did not notice clear improvements for the abdominal wall and stomach, however, the smaller oragns like the colon, small intestine and pancreas gets segmented better by adding our *Syn* datasets.