This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# Data Augmentation for Surgical Scene Segmentation with Anatomy-Aware Diffusion Models

Danush Kumar Venkatesh[1,2], Dominik Rivoir[1,3], Micha Pfeiffer[1], Fiona Kolbinger[2,4], Stefanie Speidel[1,2,3]

[1]NCT/UCC Dresden, DKFZ Heidelberg, Faculty of Medicine & University Hospital Carl Gustav Carus TUD,HZDR Dresden
[2]Faculty of Medicine & University Hospital Carl Gustav Carus TUD Dresden, Germany
[3]The Centre for Tactile Internet with Human-in-the-Loop (CeTI), TUD Dresden, Germany
[4]Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA
{firstname.lastname}@nct-dresden.de, fkolbing@purdue.edu

## Abstract

*In computer-assisted surgery, automatically recognizing anatomical organs is crucial for understanding the surgical scene and providing intraoperative assistance. While machine learning models can identify such structures, their deployment is hindered by the need for labeled, diverse surgical datasets with anatomical annotations. Labeling multiple classes (i.e., organs) in a surgical scene is time-intensive, requiring medical experts. Although synthetically generated images can enhance segmentation performance, maintaining both organ structure and texture during generation is challenging. We introduce a multi-stage approach using diffusion models to generate multi-class surgical datasets with annotations. Our framework improves anatomy awareness by training organ specific models with an inpainting objective guided by binary segmentation masks. The organs are generated with an inference pipeline using pre-trained ControlNet to maintain the organ structure. The synthetic multi-class datasets are constructed through an image composition step, ensuring structural and textural consistency. This versatile approach allows the generation of multi-class datasets from real binary datasets and simulated surgical masks. We thoroughly evaluate the generated datasets on image quality and downstream segmentation, achieving a* $15\%$ *improvement in segmentation scores when combined with real images.*

## 1. Introduction

Computer-assisted surgery (CAS) aims to improve assistance for surgical teams and minimize complications during surgical procedures [31]. A comprehensive understanding of surgical anatomy is essential for implementing such context-aware guidance in laparoscopic surgeries. Moreover, anatomy segmentation is a challenging computer vision task that serves useful in other downstream surgical tasks such as action recognition [49], surgical skill assessment [37, 58] and for navigation with segmentation [12]. Hence, training machine learning models on surgical images to semantically segment each anatomical structure serves as a promising solution to improve interventional healthcare.

Despite the remarkable progress in deep learning for semantic segmentation, their application to surgical data science faces a hindrance due to the necessity for large-scale diverse and annotated data [31]. Generating a multi-class dataset requires annotating every pixel representing each anatomical structure within the surgical scene. In contrast, binary annotations pertain to only one subject in an image. The surgical field faces a challenge in that the expertise of medical professionals, i.e., doctors, are required to annotate the datasets, who have limited time resources [48]. Consequently, this has led to very limited open-sourced multi-class anatomical datasets such as HeiSurf [5] or the Dresden Surgical Anatomy dataset (DSAD) [7] with only a few thousand images.

Simulation environments offer the capability to generate synthetic surgical images. This process is advantageous because different labels, such as semantic masks, depth, and normal maps, can be rendered automatically. Prior works [9, 41, 44, 47, 59, 64] have demonstrated the usage of synthetic images along with generative models such as generative adversarial networks (GANs) [19] to improve segmentation. However, these methods often struggle to generate high-quality images with sufficient diversity.

Diffusion models (DMs) have emerged as a promising approach to generative modeling [21, 53] and have surpassed the state-of-the-art GAN-based methods in image

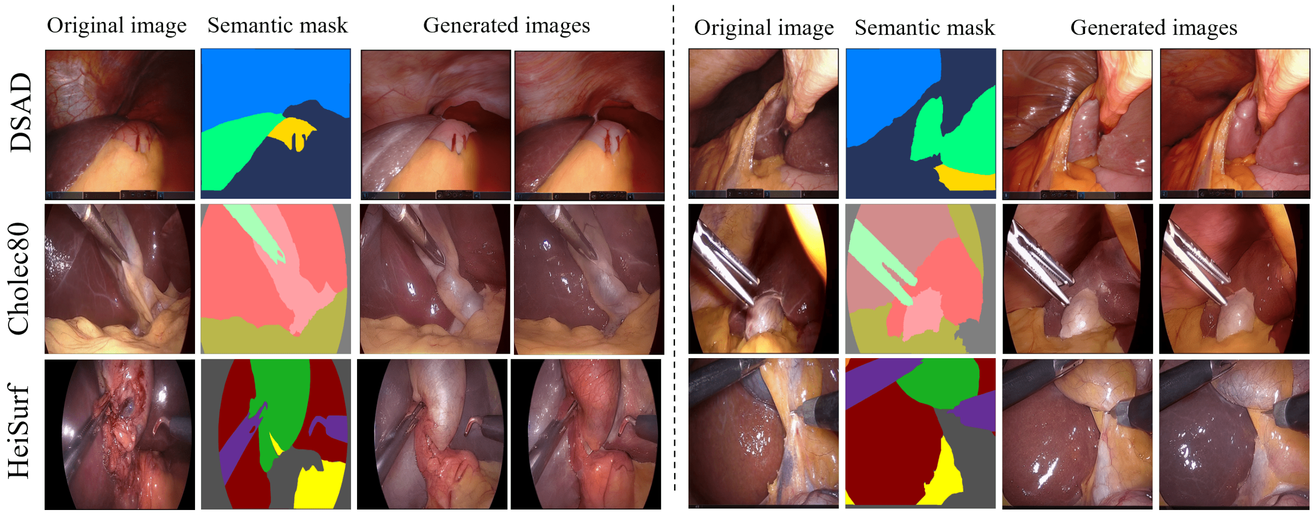


Figure 1. The generated multi-class surgical images (Generated images column) for three different surgical datasets (denoted by name on the left side) with their corresponding semantic masks using our diffusion approach. Our approach can generate realistic and diverse organ textures using the segmentation masks as masking and conditioning signals.

synthesis [11]. For surgical applications, precise control over the shape and structure of organs is essential. However, achieving spatial alignment using only text prompts in DMs is notably challenging [39, 55]. Recent research efforts, such as ControlNet [65] and T2i-Adapters [35], have introduced methods to control the output of DMs using additional signals like segmentation masks and edge maps.

In this work, we create synthetic multi-class surgical datasets with full-scene annotations. Our pipeline is based on latent diffusion models conditioned by text prompts. We utilize segmentation masks and formulate a diffusion inpainting objective to make these diffusion models understand the texture properties of different anatomies (anatomy-aware). To precisely generate different anatomical structures, we introduce an inference pipeline consisting of pre-trained ControlNet [65] with extracted edge images as the controlling signal. We address the challenges of multi-object compositionality by implementing an image-fusing method to produce multi-class surgical images containing diverse anatomical structures. We thoroughly evaluate the generated images and demonstrate their usefulness in the downstream segmentation of different anatomical structures and surgical tools.

A natural question arises: **What is the available segmentation mask?** We identify two possibilities. Firstly, we have minimal segmentation masks from the open-sourced (OS) surgical datasets (See Fig. 1). These masks accurately represent the shapes of various organs in the dataset. We use the OS masks to train the diffusion models and generate synthetic multi-class datasets through inpainting models, resulting in the first synthetic dataset (*Syn*). Secondly, surgical simulations (SS) can provide full-scene segmentation masks, although replicating the exact shapes of organs in laparoscopic setting is time-consuming (See Fig. 4). Consequently, the SS masks contain organs which approximate the true shapes similar to those in real datasets. We utilize the SS masks to generate images of different organs, leading to the synthetic dataset (*SS-Syn*) that incorporates the anatomical texture properties from real surgical datasets.

**Contributions**. We summarize our contributions as follows:

1. We introduce a multi-stage approach using diffusion models to generate high-quality, realistic surgical images with full annotations in an anatomy-controlled manner.

2. By modeling and integrating each semantic class through separate organ-specific diffusion inpainting models, our approach ensures organs' shape and texture consistency in the generated images, surpassing methods that rely on full-scene segmentation masks. An additional benefit of our method is its capacity to generate multi-class datasets using only real binary datasets and multi-class simulation masks. This is especially advantageous when multi-class segmentation labels are limited, but low-cost binary labels are available (see Sec. 5.4).

3. We thoroughly evaluate the generated images on image quality and downstream segmentation of organs. Our results show a 15% improvement in segmentation scores when models are trained on combined

datasets of our generated and real images. To support research in surgical computer vision, we are releasing the generated data along with their labels at https://gitlab.com/nct_tso_public/muli-class-image-synthesis.

## 2. Related Work

**Diffusion Models** DMs were initially introduced by Sohl-Dickstein et al. [53] and have recently gained significant attention for their superior image synthesis performance compared to GANs [11]. In Latent Diffusion Models (LDMs) [45], the diffusion process occurs in the latent image space [14], reducing computational costs. Stable Diffusion (SD) [45] is a large-scale implementation of LDM trained on natural images. The image generation was conditioned using text prompts by encoding the text inputs into latent vectors using pre-trained language models like CLIP [42]. SD remains a competitive open-source LDM model, and we use them to train on our surgical datasets.

**Controllable Image generation** To enable personalization or customization, controlling the DMs is necessary. Text-based editing has been achieved by adjusting text prompts or manipulating the CLIP features [3, 6, 16, 20, 26]. ControlNet [65] proposes to attain spatial conditioning via learning adapter networks similar to the Unet in LDMs. Conditioning signals were learned with additional smaller adapters that plug into DMs in T2i-Adapter [35]. Despite their promise, these methods demand extensive computational resources and prolonged training times for adaption to the surgical datasets. Additionally, generating multi-subjects (classes) together is still a challenge that has yet to be solved. In our approach, we use the advantages of ControlNet for spatial conditioning on each organ and generate the multi-class(subject) images via a simple image composition step.

**Surgical image synthesis** Studies on laparoscopic image synthesis have primarily focused on image-to-image (I2I) translation. Computer-simulated surgical images [41, 44, 59, 64], phantom data [50] and segmentation maps [33] were used with GANs to synthesize realistic surgical images or video data. Frisch et al. [15] generated rare cataract surgical images using DDIMs [54] and SD based I2I methods were explored in [25, 60]. Our method requires only segmentation masks as input, and hence, this further reduces the simulation modeling efforts. Allmendinger et al. [1] analyzed diffusion models like Dall-e2 [43], Imagen [46] for laparoscopic image synthesis. Diffusion models have become popular for generating medical data [27], especially MRI [13, 28] or CT images [30]. However, it is essential to note that they differ in imaging modality from surgical data. To our knowledge, this is the first study to focus on generating multi-class surgical image datasets.

**Segmentation strategy** Data augmentation is widely used to amplify existing datasets [51] and can be easily implemented during training. Both color and geometric augmentations have demonstrated improvements in segmentation performance for medical images [18]. In surgical context, Jenke et al. [24] suggested an implicit labeling method based on mutual information between classes for surgical images. We train segmentation models using various augmentations as baselines and evaluate their performance against our synthetically generated datasets.

## 3. Preliminaries

### 3.1. Stable Diffusion

Diffusion models [21, 53] (DMs) are probabilistic generative models that iteratively generate images by removing noise from an initial Gaussian noise image, $x_T \sim \mathcal{N}(0, I)$. This includes a forward process defined as

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} z, \tag{1}$$

where $z \sim \mathcal{N}(0, I)$, $\bar{\alpha_t}$ is the noise schedule and $x_0$ is the clean image. In the backward process a neural network $\epsilon_\theta(x_t, t, P)$ learns to predict the added noise $z_t$ at each step $t$ by optimizing the loss, $\mathcal{L}_{DM}$,

$$\mathcal{L}_{DM} = \mathbb{E}_{z \sim \mathcal{N}(0,I),t} \left[ \| z - \epsilon_\theta \left( x_t, t, P \right) \|_2^2 \right], \tag{2}$$

where P is the guiding text prompt. In this work, we leverage the pre-trained text-conditioned Stable Diffusion (SD) [45] model, in which the diffusion process occurs in the latent space using an image autoencoder.

### 3.2. ControlNet

ControlNet (CN) is a framework designed for controlling pre-trained DMs' image generation process by integrating additional conditioning signals such as sketches, key points, edges, and segmentation maps [65]. The model consists of two sets of U-Net weights derived from the pre-trained DM: with $\theta_c$, that undergoes training using task-specific datasets to accommodate the additional condition, $\mathrm{c}_f$ and the frozen copy, $\theta_L$. Let $S_f$ be the input feature map from SD, then the feature map $y_c$ from the ControlNet is defined as,

$$\boldsymbol{y}_c = \mathcal{B}(S_f; \theta_L) + \mathcal{C}\left(\mathcal{B}\left(S_f + \mathcal{C}\left(\boldsymbol{c_f}; \Theta_{\mathrm{s}1}\right); \theta_c\right); \Theta_{\mathrm{s}2}\right), \tag{3}$$

where $\mathcal{C}(\cdot;\cdot)$ denotes 1x1 zero-convolution layers with $\Theta_{(\cdot)}$ parameters that links pre-trained SD with ControlNet blocks and $\mathcal{B}(\cdot;\cdot)$ is a neural block with a set of parameters. We use pre-trained CN for spatial conditioning.

### 3.3. SDEdit

SDEdit is an image editing method that uses stochastic differential equations (SDE) to solve the reverse diffusion process [34]. A user-given image is firstly noised up to a specific limit depending on the specified noising strength, and denoising starts from this noisy image, which serves as a controlling signal, back to a realistic image. Text prompts can be added as additional guiding signals during the denoising process. This method is used in the final stage for image refinement in our pipeline.

# 4. Methodology

To generate the multi-class datasets, we begin with real surgical images and segmentation masks to train the SD model using an inpainting objective. An inference pipeline containing this trained SD model and a pre-trained CN model is utilized to generate different organs. The generated organs are then fused into a multi-class image during the image composition stage. Finally, image refinement is performed using the SDEdit approach. An overview of the process is illustrated in Fig. 2.

## 4.1. Stable Diffusion inpainting

Our Stage-1 comprises training a diffusion inpainting model. Given an image $x$ with a mask $m$, the inpainting model is trained to generate an object in the masked region. Existing inpainting models randomly mask regions within an image. As a result, only partial objects or background areas can be generated in the masked region. Instead, we use the existing segmentation mask of the individual organs with the organ type as the corresponding text signal for training. In the forward process, the mask $m$ and its corresponding text prompt $P$ for image $x_0$ is used to obtain $\tilde{x}_t$ with Eq. (1) as,

$$\tilde{x}_t = x_t \odot m + x_0 \odot (1 - m), \tag{4}$$

The model $\epsilon_\theta$ learns to predict the added noise $z_t$ with the objective

$$\mathcal{L}_{DM} = \mathbb{E}_{z \sim \mathcal{N}(0,I),t} \left[ \| z - \epsilon_\theta \left( \tilde{x}_t, t, P, m \right) ) \|_2^2 \right] , \tag{5}$$

We sample for $T$ steps starting from $x_T = \epsilon \odot m + x_0 \odot (1 - m)$ to obtain the inpainted result $x_0$. The inpainting objective focuses specifically on the masked region, which leads to the diffusion models learning the texture of each organ. We call this model **S**urgical **S**table **I**npaint (SSI). We split the segmentation mask into $N$ individual organs and train the SSI model $N$ times separately. In this manner, we make our approach aware of each individual anatomy. As an additional flexibility this allows the introduction of new organs into a multi-class scene and only that diffusion model needs to be trained. Our approach adds minimal overhead in training compared to other methods based on GANs or DMs. Additionally, we fine-tune a SD model with only Eq. (2) with all combined organs for the image refinement stage, which is explained below.

## 4.2. Inference with ControlNet

In Stage-2, anatomical structures are generated using CNs. Our preliminary results indicated that maintaining anatomical structures solely with segmentation masks and text prompts proved challenging. Hence, we opted for a simplified inference stage using pre-trained CN models. Training a CN model from scratch requires extensive computational resources and a large dataset. For instance, the CN model controlled by segmentation maps (CN-Seg) was trained with 164k images [65]. Hence, we circumvent this process by integrating a pre-trained CN model into the inpainting SSI (SSI-CN) model to control the shape and texture of the generated organs precisely. The number of classes for the pre-trained CN-Segmodel did not match our surgical datasets, so we opted for the pre-trained soft edge CN model, which uses extracted edge images from the segmentation masks as the conditioning signal. Given an input image and a mask, the new organ texture is inpainted only in the masked region leaving the background the same.

## 4.3. Image composition

In Stage-3, we generate the multi-class synthetic datasets. The different generated anatomical structures with SSI-CN model are cut out per organ from the generated image using the separate masks and combined to form the newly composed image. This results in an image comprising multiple classes with corresponding semantic labels.

## 4.4. Image enhancement stage

We noticed that the image composition operation introduced sharp edges between the organs and lighting artifacts, which is not present in real surgical images (see Fig. 3). Hence, in Stage-4, we perform an image enhancement step using SDEdit [34]. We use the SD model trained with all organs combined with SDEdit to remove the inconsistencies introduced in the previous Stage-3. Low levels of noise has shown to improve texture components in images [52] and hence this step can be optionally added to maintain the overall texture.

# 5. Experiments & Results

In this section, we explain our experimental setup and the evaluation procedure for the generated synthetic images. We evaluate the generated datasets on image quality and their utility as training data for downstream segmentation.

## 5.1. Data

For real surgical datasets, we used the CholecSeg8K [22] dataset, which is a labeled (multi-class) subset of Cholec80 [2] with 5080 training images and 2000 test images, and the HeiSurf [5] dataset consisting of 330 training and 110 test images. Both these datasets involve the

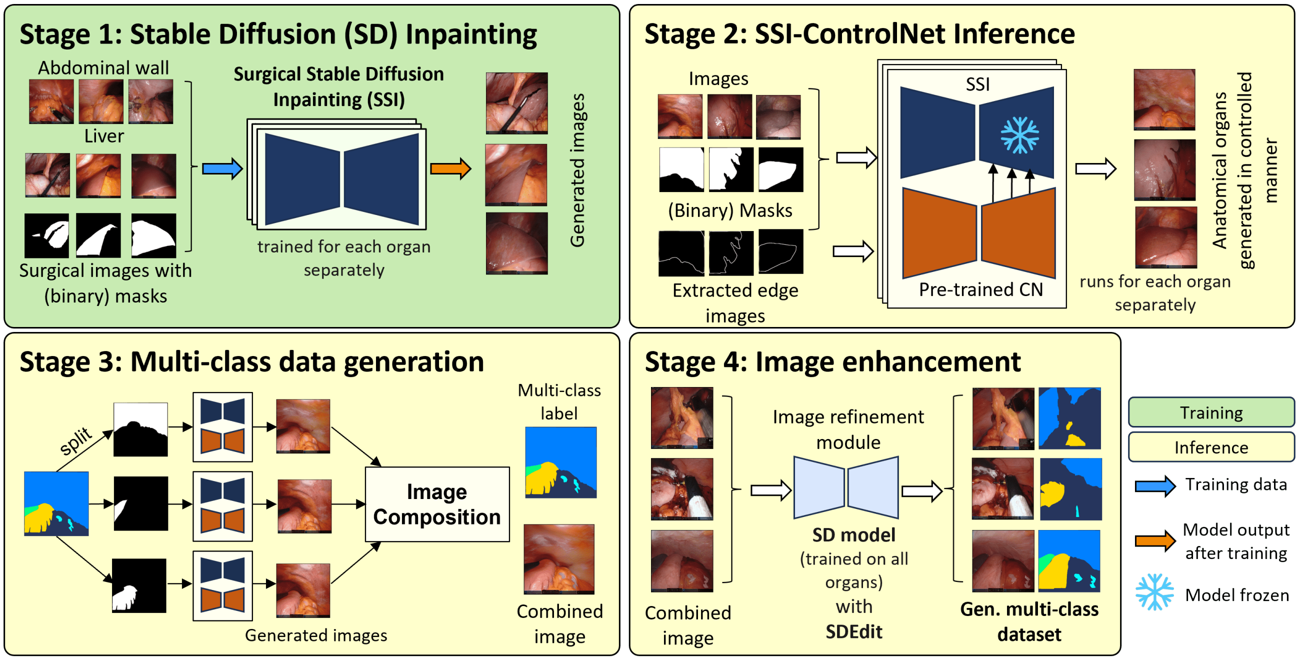


Figure 2. Overview of the diffusion approach to generate a multi-class dataset. Stage-1 involves training the SD inpainting model using the real images and masks for each organ separately. In stage-2, pre-trained ControlNet is plugged into the SSI model (SSI-CN) to precisely generate each anatomical structure using extracted edges from the segmentation mask. The image composition in stage-3 includes cutting out each organ from the generated image and combining them together to form the multi-class surgical dataset. Stage-4 (optional) includes an image refinement process using SDEdit [34] to rectify inconsistencies during the composition operation and generate the multi-class images. We skip stage-1 for the simulated masks and start directly with the inference stages to generate the synthetic datasets.

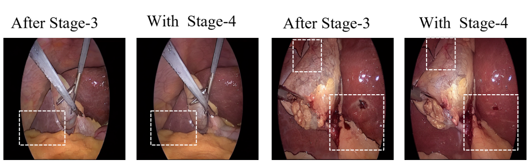


Figure 3. The generated images before and after Stage-4. White boxes show the inconsistent regions like sharp junctions between organs.

surgical removal of the gall bladder (Cholec.). Secondly, we utilize the DSAD dataset, which has binary (BN) segmentation masks of 1000 images for each 11 organs and a multi-class (MC) subset of 1400 images containing six organs. We use the labels (six organs) common to both the datasets. It is to be noted that the MC-subset is smaller in number of images compared to the BN datasets. We used simulated masks from from Pfeiffer et al. [41] and Rivoir et al. [44] for the Cholec. and DSAD datasets respectively.

### 5.2. Task implementation

We design two distinct tasks to showcase our method's ability to work with multi- and binary-class datasets.

**Task T1: Multi $\rightarrow$ Multi** In this task, we use the multi-class segmentation masks from the CholecSeg8k and HeiSurf datasets. As explained in Sec. 4.1, the masks are split into binary classes for training and images are generated via fusion in Stage-3.

**Task T2: Binary $\rightarrow$ Multi** We re-iterate in this task on how our diffusion approach can be used to train only on real binary datasets and construct multi-class datasets. We use only the binary dataset (BN) from the DSAD dataset to train the diffusion models. For Stages 2 and 3, we use this dataset's multi-class (MC) segmentation masks as inputs and generate the multi-class synthetic datasets. For both tasks, we also use the simulated masks directly from Stage-2.

**Implementation details** We use Stable-Diffusion inpainting v1.5 as the base diffusion model. The model was fine-tuned for 1500 steps with a $1e^{-5}$ learning rate. The Soft Edge CN model was used during inference (Stage 2) with a conditioning scale of 0.5. For the details on the text prompts, kindly refer to the SM. Our training setup for six organs takes 3.5 hrs, whereas fine-tuning a pre-trained CN model already takes 3.2 hrs on Nvidia RTX $A5000$ GPUs. An image is generated in 5.25s using the inference pipeline(Stage-2 to Stage-4). For Stage-4, the SD model is

| Method | CFID (↓) | KID (↓) | CMMD(↓) | LPIPS(↓) |
|---|---|---|---|---|
| SPADE [38] | $391.85_{\pm 6.53}$ | $0.39_{\pm 0.04}$ | $1.94_{\pm 0.13}$ | $0.60_{\pm 0.03}$ |
| Pix2PixHD [61] | $371.81_{\pm 8.81}$ | $0.50_{\pm 0.02}$ | $2.47_{\pm 0.06}$ | $0.72_{\pm 0.02}$ |
| ControlNet [65] | $360.22_{\pm 8.82}$ | $0.45_{\pm 0.05}$ | $2.22_{\pm 0.03}$ | $0.71_{\pm 0.03}$ |
| T2i-Adapter [35] | $358.50_{\pm 8.57}$ | $0.56_{\pm 0.03}$ | $4.17_{\pm 0.15}$ | $0.76_{\pm 0.03}$ |
| Ours-*SS-Syn* | $\mathbf{337.16_{\pm 3.23}}$ | $0.42_{\pm 0.01}$ | $1.75_{\pm 0.08}$ | $0.74_{\pm 0.01}$ |
| Ours-*Syn* | $348.04_{\pm 9.10}$ | $\mathbf{0.39_{\pm 0.03}}$ | $\mathbf{1.69_{\pm 0.10}}$ | $\mathbf{0.58_{\pm 0.03}}$ |

Table 1. **Image quality comparison on CholecSeg8K dataset**. Our synthetic datasets show better quality than GAN or diffusion based approaches.

trained for 4 mins and the inference took 1.1s, as only ten sampling steps were used for image refinement. We consider this to be a minimal overhead in comparison to annotating the surgical scenes.

### 5.3. Evaluation

**Baselines** We compare our approach to popular semantic image synthesis GANs such as SPADE [38], SPADE-vae [38] and Pix2PixHD [61]. For the diffusion approaches, we used the SD model trained on all organs as the base model and trained the ControlNet [65] and T2i-Adapter (T2i) [35] from scratch conditioned on the multi-class labels from the real surgical datasets. We also fine-tuned the pre-trained Soft edge CN (ControlNet-SE) and Canny T2i (T2i-Adapter-CY). These models serve as powerful baselines for mask-conditioned image generation.

**Image quality** Firstly, we evaluate the quality and diversity of the generated images. We employ different metrics to compare the generated image quality. We use the popular metric CFID [40] and KID [4] to measure the realism of the images. The CMMD score [23] measures the image quality on better-extracted features from CLIP and is suitable for smaller datasets as it is unbiased, unlike CFID. Finally, we compute the LPIPS [66] metric, highlighting the image's perceptual quality.

**Downstream semantic segmentation** We asses the utility of generated images by performing two evaluations: (1). To compare our approach to other image synthesis methods, we train a DV3+ [8] model with the generated images from different image synthesis methods and fine-tune them on the real images. The model performance is evaluated on the test set of the real dataset. (2). We train different state-of-the-art segmentation models such as Unet++ [67], UperNet [62] and Segformer [63] models. As baselines, we train each model using no augmentations, color augmentations, and a combination of color and spatial augmentations. We curated a list of these different augmentations from prior works [17, 18, 24, 29, 36, 56, 57]. For Task 2: Binary $\rightarrow$ Multi, we train implicit labeling method [24]

| Method | Dice (↑) | IOU (↑) | HD(↓) |
|---|---|---|---|
| SPADE [38] | $0.58_{\pm 0.01}$ | $0.46_{\pm 0.02}$ | $119.30_{\pm 1.75}$ |
| SPADE-vae [38] | $0.56_{\pm 0.01}$ | $0.44_{\pm 0.01}$ | $109.86_{\pm 1.94}$ |
| Pix2Pix-HD [61] | $0.58_{\pm 0.01}$ | $0.44_{\pm 0.01}$ | $110.58_{\pm 0.82}$ |
| ControlNet [65] | $0.57_{\pm 0.01}$ | $0.44_{\pm 0.01}$ | $115.59_{\pm 6.80}$ |
| ControlNet-SE [65] | $0.61_{\pm 0.01}$ | $0.48_{\pm 0.02}$ | $107.54_{\pm 3.49}$ |
| T2I-adapter [35] | $0.59_{\pm 0.02}$ | $0.46_{\pm 0.01}$ | $117.55_{\pm 1.63}$ |
| T2I-adapter-CY [35] | $0.60_{\pm 0.03}$ | $0.46_{\pm 0.02}$ | $110.01_{\pm 3.21}$ |
| Ours-*SS-Syn* | $0.64_{\pm 0.05}$ | $0.51_{\pm 0.05}$ | $\mathbf{95.86_{\pm 8.25}}$ |
| Ours-*Syn* | $\mathbf{0.68_{\pm 0.01}}$ | $\mathbf{0.56_{\pm 0.01}}$ | $95.93_{\pm 6.89}$ |

Table 2. **Segmentation comparison on the CholecSeg8K dataset** (T1:Multi→Multi). The results of the model trained using our synthetic data outperforms all the baselines.

and fine-tune this model on the MC subset as a baseline. We compare the performance of these methods against the models that are trained on our datasets. *Syn* denotes synthetic dataset using mask from the real surgical datasets and *SS-Syn* uses masks from surgical simulations. Following suggestions from [32], we chose Dice, IOU, and Hausdorff distance (HD) as the segmentation evaluation metrics and ignore the bg. as we inpaint only the masked organ region. The readers can refer to the suppl. material for more details on the training process.

### 5.4. Results

**(Task 1: Multi $\rightarrow$ Multi)** (1). *CholecSeg8K*: The image quality evaluation results are indicated in Tab. 1. Overall, the results indicate that our approach generates high-quality and diverse surgical images. Fig. 4 shows generated images from SS masks, further indicating that our approach can be used to generate organs with different shapes and textures. Tab. 2 indicates the segmentation results using the generated images from different approaches. The ControlNet-SE shows a dice score of 0.61 and IOU of 0.48, which outperforms the GAN methods. Our *Syn* dataset leads to 8% improvement in dice and outperforms the ControlNet-SE model. The *SS-Syn* dataset performs better than other baselines and falls slightly short of the *Syn* dataset.

The segmentation performance of different models using data augmentation is shown in Tab. 3. Overall, adding color and spatial augmentations improved the performance across three models on the real surgical datasets. The results show that only using the *Syn* dataset already matches the scores of the real images. We consistently see performance improvement across all the segmentation models when combined *Syn*+Real training is done. The *SS-Syn*+Real dataset scores are similar to those of the *Syn* datasets. We hypothesize that the remaining performance difference can be attributed to the organ shape between the domains.

(2). HeiSurf: The results on image quality in Tab. 4 in-

| Training scheme | Unet++ [67] | | | DV3+ [8] | | | UperNet-Tiny [62] | | |
|---|---|---|---|---|---|---|---|---|---|
| | Dice(↑) | IOU (↑) | HD (↓) | Dice (↑) | IOU (↑) | HD (↓) | Dice (↑) | IOU (↑) | HD (↓) |
| Real with no-aug | $0.50_{\pm 0.03}$ | $0.36_{\pm 0.01}$ | $101.03_{\pm 0.12}$ | $0.50_{\pm 0.01}$ | $0.36_{\pm 0.01}$ | $115.36_{\pm 3.82}$ | $0.56_{\pm 0.01}$ | $0.47_{\pm 0.02}$ | $118.37_{\pm 6.62}$ |
| Real with color-aug | $0.52_{\pm 0.01}$ | $0.38_{\pm 0.02}$ | $\mathbf{98.95}_{\pm \mathbf{2.05}}$ | $0.53_{\pm 0.01}$ | $0.39_{\pm 0.01}$ | $101.54_{\pm 0.19}$ | $0.59_{\pm 0.01}$ | $0.45_{\pm 0.01}$ | $110.93_{\pm 1.42}$ |
| Real with color+spatial-aug | $0.61_{\pm 0.05}$ | $0.49_{\pm 0.04}$ | $109.09_{\pm 0.52}$ | $0.58_{\pm 0.01}$ | $0.45_{\pm 0.01}$ | $108.14_{\pm 1.07}$ | $0.61_{\pm 0.04}$ | $0.50_{\pm 0.05}$ | $108.63_{\pm 1.51}$ |
| Ours only *Syn* | $0.53_{\pm 0.03}$ | $0.40_{\pm 0.01}$ | $110.65_{\pm 1.31}$ | $0.53_{\pm 0.01}$ | $0.41_{\pm 0.02}$ | $108.66_{\pm 1.18}$ | $0.56_{\pm 0.01}$ | $0.44_{\pm 0.01}$ | $109.41_{\pm 2.09}$ |
| Ours-*SS-Syn* + Real | $\mathbf{0.67}_{\pm \mathbf{0.01}}$ | $\mathbf{0.54}_{\pm \mathbf{0.01}}$ | $107.10_{\pm 0.49}$ | $0.64_{\pm 0.05}$ | $0.51_{\pm 0.05}$ | $\mathbf{95.86}_{\pm \mathbf{8.25}}$ | $0.65_{\pm 0.03}$ | $0.53_{\pm 0.02}$ | $\mathbf{95.76}_{\pm \mathbf{2.49}}$ |
| Ours-*Syn* + Real | $0.64_{\pm 0.03}$ | $0.51_{\pm 0.01}$ | $101.96_{\pm 1.43}$ | $\mathbf{0.68}_{\pm \mathbf{0.01}}$ | $\mathbf{0.56}_{\pm \mathbf{0.01}}$ | $95.93_{\pm 6.89}$ | $\mathbf{0.67}_{\pm \mathbf{0.01}}$ | $\mathbf{0.54}_{\pm \mathbf{0.01}}$ | $99.97_{\pm 2.24}$ |

Table 3. **Evaluation on CholecSeg8K dataset using different segmentation models** (T1:Multi→Multi). A 10% improvement was noticed in the segmentation scores with combined training *Syn*+Real. The best scores are highlighted in bold.

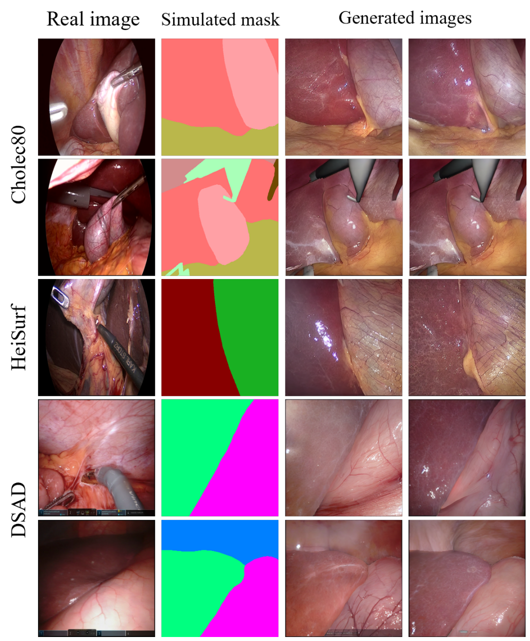


Figure 4. The generated images using simulated masks (SS). By using SS masks, we can generate surgical images other than the train datasets as the organ shapes differs with a similar organ texture to real datasets.

| Method | CFID (↓) | KID (↓) | CMMD(↓) | LPIPS(↓) |
|---|---|---|---|---|
| SPADE [38] | $382.55_{\pm 2.32}$ | $0.35_{\pm 0.05}$ | $2.08_{\pm 0.04}$ | $0.70_{\pm 0.04}$ |
| Pix2PixHD [61] | $\mathbf{347.75}_{\pm \mathbf{4.12}}$ | $0.39_{\pm 0.04}$ | $3.81_{\pm 0.10}$ | $0.82_{\pm 0.04}$ |
| ControlNet-SE [65] | $380.68_{\pm 3.41}$ | $0.32_{\pm 0.05}$ | $2.09_{\pm 0.03}$ | $0.85_{\pm 0.03}$ |
| T2i-Adapter-CY [35] | $409.16_{\pm 3.67}$ | $0.34_{\pm 0.05}$ | $1.07_{\pm 0.05}$ | $0.78_{\pm 0.03}$ |
| Ours-*SS-Syn* | $351.09_{\pm 2.70}$ | $0.39_{\pm 0.05}$ | $0.69_{\pm 0.03}$ | $0.69_{\pm 0.04}$ |
| Ours-*Syn* | $369.62_{\pm 8.15}$ | $\mathbf{0.31}_{\pm \mathbf{0.01}}$ | $\mathbf{0.57}_{\pm \mathbf{0.02}}$ | $\mathbf{0.68}_{\pm \mathbf{0.01}}$ |

Table 4. **Image quality comparison on HeiSurf dataset**. The images generated from our method shows better image quality than other methods.

dicates that our approach is better in maintaining realism of the images. For such smaller datasets, other image synthesis methods suffer to generate images suitable for the ap-

| Training scheme | Unet++ [67] | | | DV3+ [8] | | |
|---|---|---|---|---|---|---|
| | Dice(↑) | IOU (↑) | HD (↓) | Dice (↑) | IOU (↑) | HD (↓) |
| Real with no-aug | $0.40_{\pm 0.02}$ | $0.29_{\pm 0.01}$ | $\mathbf{148.36}_{\pm \mathbf{2.56}}$ | $0.30_{\pm 0.05}$ | $0.20_{\pm 0.07}$ | $239.57_{\pm 9.69}$ |
| Real with color-aug | $0.42_{\pm 0.01}$ | $0.30_{\pm 0.02}$ | $165.76_{\pm 2.02}$ | $0.42_{\pm 0.01}$ | $0.30_{\pm 0.02}$ | $180.41_{\pm 3.61}$ |
| Real with color+spatial-aug | $0.45_{\pm 0.01}$ | $0.32_{\pm 0.01}$ | $247.21_{\pm 8.15}$ | $0.40_{\pm 0.02}$ | $0.31_{\pm 0.01}$ | $206.34_{\pm 3.04}$ |
| Ours only *Syn* | $0.42_{\pm 0.01}$ | $0.30_{\pm 0.01}$ | $201.20_{\pm 9.01}$ | $0.35_{\pm 0.02}$ | $0.24_{\pm 0.01}$ | $205.34_{\pm 3.62}$ |
| Ours-*SS-Syn* + Real | $\mathbf{0.55}_{\pm \mathbf{0.01}}$ | $0.36_{\pm 0.02}$ | $211.81_{\pm 2.47}$ | $0.47_{\pm 0.01}$ | $0.33_{\pm 0.01}$ | $170.63_{\pm 2.19}$ |
| Ours-*Syn* + Real | $0.53_{\pm 0.01}$ | $\mathbf{0.40}_{\pm \mathbf{0.01}}$ | $207.65_{\pm 1.70}$ | $\mathbf{0.49}_{\pm \mathbf{0.01}}$ | $\mathbf{0.36}_{\pm \mathbf{0.01}}$ | $\mathbf{165.42}_{\pm \mathbf{3.04}}$ |

Table 5. **Comparison of data augmentations on HeiSurf dataset** (T1:Multi→Multi). Using the generated images leads to improved performance across the two models.

plication. Our *Syn* datasets leads to a 10% difference in scores compared to other models (in *suppl*). Furthermore as evidenced from Tab. 5, for the Unet++ architecture, the combined training *Syn*+Real shows a 8% improvement in both dice and IOU compared to the data augmentations on the real images. For the DV3+ model, dice score improved by 9% with a drastic improvement in HD scores when using the *Syn* dataset. These results further show that our approach is effective at capturing the texture of different organs, thereby allowing the generation of surgical datasets. Additonal qualitative results are in suppl. material.

**(Task 2: Binary → Multi)** The qualitative results are presented in Fig. 5. Our method precisely generates the organs according to the semantic mask. In contrast, the GAN-based method fails to maintain image quality, and the diffusion approaches fall short in maintaining spatial alignment. These results highlight the importance of the image composition stage, which aids in preserving the organ structures while the diffusion process effectively generates their textures. The segmentation scores shown in Tab. 6 indicate that our method outperforms the baselines, achieving an improvement of more than 8% in scores. Additionally, in Tab. 7, the results demonstrate that combining the generated synthetic data leads to a 5% improvement in dice and IOU for the DV3+ model, with a notable boost in HD scores observed for the Segformer model. Combining the generated datasets with the implicit labeling method showed smaller improvements.

**Ablation study** The results of the ablation study is shown in Tab. 8. In Tab. 8 Config A, we removed the pretrained CN from Stage-2 and the image enhancement stage (Stage-4) on the DSAD datasets. A decrease in segmenta-

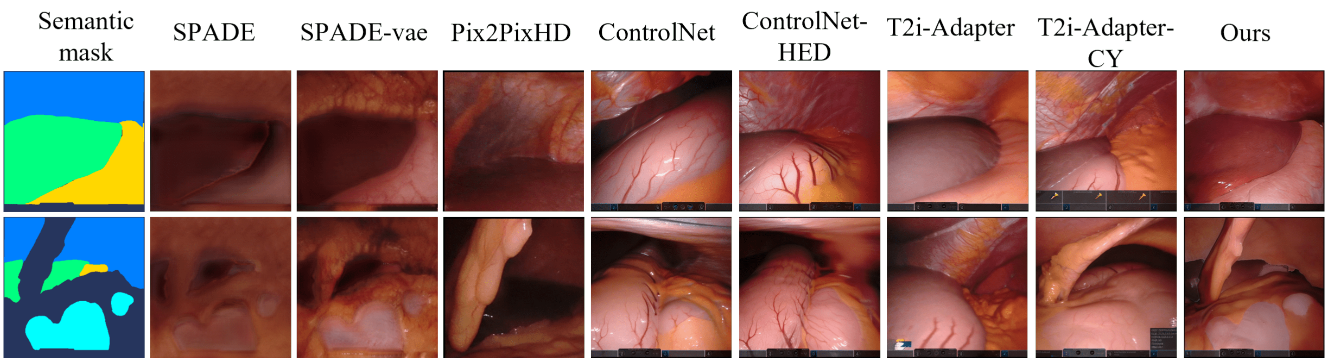


Figure 5. **Image quality comparison on the DSAD dataset.** The GAN methods (columns 2-4) fail to generate high quality images. The diffusion methods (columns 5-8) generate organs with realistic looking textures, however the spatial alignment to the semantic label is broken. Our method is able to maintain the shape and texture of different organs.

| Method | Dice (↑) | IOU (↑) | HD(↓) |
|---|---|---|---|
| SPADE [38] | $0.75_{\pm 0.01}$ | $0.67_{\pm 0.04}$ | $93.63_{\pm 0.33}$ |
| SPADE-vae [38] | $0.79_{\pm 0.02}$ | $0.69_{\pm 0.01}$ | $\mathbf{83.39}_{\pm \mathbf{4.50}}$ |
| Pix2Pix-HD [61] | $0.76_{\pm 0.02}$ | $0.66_{\pm 0.01}$ | $103.66_{\pm 1.02}$ |
| ControlNet [65] | $0.76_{\pm 0.03}$ | $0.67_{\pm 0.01}$ | $101.67_{\pm 3.71}$ |
| ControlNet-SE [65] | $0.78_{\pm 0.02}$ | $0.68_{\pm 0.01}$ | $95.53_{\pm 3.08}$ |
| T2I-adapter [35] | $0.78_{\pm 0.03}$ | $0.68_{\pm 0.01}$ | $89.15_{\pm 3.89}$ |
| T2I-adapter-CY [35] | $0.76_{\pm 0.01}$ | $0.67_{\pm 0.03}$ | $89.28_{\pm 2.36}$ |
| Ours-*SS-Syn* | $0.82_{\pm 0.01}$ | $0.73_{\pm 0.04}$ | $86.32_{\pm 4.15}$ |
| Ours-*Syn* | $\mathbf{0.83}_{\pm \mathbf{0.01}}$ | $\mathbf{0.74}_{\pm \mathbf{0.01}}$ | $92.50_{\pm 6.05}$ |

Table 6. **Segmentation eval. on DSAD dataset.** Our synthetic datasets show superior performance to other baselines.

| Training scheme | DV3+ [8] | | | Segformer [63] | | |
|---|---|---|---|---|---|---|
| | Dice(↑) | IOU (↑) | HD (↓) | Dice (↑) | IOU (↑) | HD (↓) |
| Real (MC) with no-aug | $0.78_{\pm 0.01}$ | $0.68_{\pm 0.03}$ | $218.80_{\pm 7.73}$ | $0.81_{\pm 0.01}$ | $0.74_{\pm 0.02}$ | $73.80_{\pm 2.52}$ |
| Real (MC) with color-aug | $0.78_{\pm 0.03}$ | $0.70_{\pm 0.02}$ | $206.07_{\pm 5.26}$ | $0.83_{\pm 0.01}$ | $0.75_{\pm 0.01}$ | $75.79_{\pm 2.01}$ |
| Real (MC) with color+spatial-aug | $0.79_{\pm 0.02}$ | $0.70_{\pm 0.02}$ | $196.24_{\pm 7.24}$ | $0.84_{\pm 0.02}$ | $0.76_{\pm 0.01}$ | $85.58_{\pm 7.49}$ |
| Implicit label [24] on Real (BN) | $0.22_{\pm 0.02}$ | $0.13_{\pm 0.01}$ | $296.45_{\pm 5.96}$ | $0.22_{\pm 0.01}$ | $0.15_{\pm 0.01}$ | $324.90_{\pm 8.85}$ |
| Implicit label on Real (BN) + Real (MC) | $0.80_{\pm 0.01}$ | $0.70_{\pm 0.01}$ | $83.56_{\pm 2.01}$ | $0.82_{\pm 0.01}$ | $0.74_{\pm 0.01}$ | $74.31_{\pm 6.71}$ |
| Ours only *Syn* | $0.60_{\pm 0.03}$ | $0.51_{\pm 0.01}$ | $95.19_{\pm 1.11}$ | $0.62_{\pm 0.01}$ | $0.51_{\pm 0.02}$ | $116.14_{\pm 1.08}$ |
| Ours *Syn* + Implicit label on Real (BN) + Real (MC) | $0.81_{\pm 0.01}$ | $0.70_{\pm 0.02}$ | $\mathbf{81.32}_{\pm \mathbf{4.59}}$ | $0.84_{\pm 0.02}$ | $0.76_{\pm 0.02}$ | $\mathbf{69.52}_{\pm \mathbf{3.35}}$ |
| Ours *SS-Syn* + Real (MC) | $0.82_{\pm 0.01}$ | $0.73_{\pm 0.04}$ | $86.32_{\pm 4.15}$ | $0.84_{\pm 0.01}$ | $0.75_{\pm 0.02}$ | $82.54_{\pm 5.20}$ |
| Ours *Syn* + Real (MC) | $\mathbf{0.83}_{\pm \mathbf{0.01}}$ | $\mathbf{0.74}_{\pm \mathbf{0.01}}$ | $92.50_{\pm 6.05}$ | $\mathbf{0.86}_{\pm \mathbf{0.01}}$ | $\mathbf{0.78}_{\pm \mathbf{0.02}}$ | $79.90_{\pm 1.04}$ |

Table 7. **Comparison of augmentations on DSAD dataset** (T2:Binary→Multi). BN denotes binary dataset and MC multi-class subset. The combined *Syn*+Real on (MC) training method shows the best performance across the two models.

| Method | Stage-2 CN | Stage-4 | Dice (↑) | IOU(↑) | HD(↓) |
|---|---|---|---|---|---|
| Config A | × | × | $0.52_{\pm 0.01}$ | $0.41_{\pm 0.01}$ | $110.55_{\pm 4.58}$ |
| Config B | × | ✓ | $0.55_{\pm 0.02}$ | $0.42_{\pm 0.01}$ | $102.10_{\pm 2.42}$ |
| Ours | ✓ | ✓ | $0.60_{\pm 0.03}$ | $0.51_{\pm 0.01}$ | $95.19_{\pm 1.11}$ |

Table 8. **Ablation study on the DSAD dataset.** Removing Stage-4 or the CN from Stage-2 leads to drop in performance.

tion metrics is seen in this case. Similarly, we added Stage-4 for Config B and saw minor improvements in the scores. Config C, our approach highlights the need for a combination of both the stages shown by higher dice and HD scores.

## 6. Conclusion

In this study, we present a diffusion approach for generating multi-class surgical datasets. Our findings show that guiding the generation process through inpainting and edge-conditioning organ-specific models preserves both the texture and shape of the organs. Moreover, our approach allows the creation of multi-class datasets using only binary real datasets and multi-class simulated masks. The evaluation results confirm that our synthetic datasets are high quality and valuable as downstream training datasets. As our generated images contain realistic textures they can be effectively utilized as pre-training datasets for other downstream surgical tasks like surgical target prediction or detection (Tab.5 in *suppl.*).

**Limitations**. Our approach produces realistic images, but it has limitations. We train diffusion models on each organ separately. Implementing spatial conditioning of subjects like bounded attention [10] could be a promising solution for multi-organ composition. Due to organ compositionality we fall behind on maintaining the depth and lighting variations within the generated images. As a promising alternative pseudo depth maps combined with a defined lighting for the scene could be integrated into the inference pipeline.

**Acknowledgements**. This work is partly supported by BMBF in DAAD project 57616814(SECAI) and partly by the German Research Foundation(DFG) as part of Germany's Excellence Strategy – EXC 2050/1 –Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of TUD. DK. thanks Isabel Funke for the insightful discussions.

## References


[1] Simeon Allmendinger, Patrick Hemmer, Moritz Queisner, Igor Sauer, Leopold Müller, Johannes Jakubik, Michael Vössing, and Niklas Kühl. Navigating the synthetic realm: Harnessing diffusion-based models for laparoscopic text-to-image generation. *arXiv preprint arXiv:2312.03043*, 2023. 3

[2] Twinanda Andru, Shehata Sherif, Mutter Didier, Marescaux Jacques, Michel De Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36, 02 2016. 4

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. 3

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[5] S Bodenstedt, S Speidel, M Wagner, J Chen, A Kisilenko, BP Müller, L Maier-Hein, B Oliveira, S Hong, J Zamora-Anaya, et al. Heichole surgical workflow analysis and full scene segmentation (heisurf). MICCAI, 2021. 1, 4

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[7] Matthias Carstens, Franziska M Rinner, Sebastian Bodenstedt, Alexander C Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R Kolbinger. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1):3, 2023. 1

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 6, 7, 8

[9] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1841–1850, 2019. 1

[10] Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. *arXiv preprint arXiv:2403.16990*, 2024. 8

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3

[12] Reuben Docea, Jinjing Xu, Wei Ling, Alexander C Jenke, Fiona R Kolbinger, Marius Distler, Carina Riediger, Jürgen Weitz, Stefanie Speidel, and Micha Pfeiffer. Seesaw: Learning soft tissue deformation from laparoscopy videos with gnns. *IEEE Transactions on Biomedical Engineering*, 2024. 1

[13] Zolnamar Dorjsembe, Sodtavilan Odonchimed, and Furen Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical Imaging with Deep Learning*, 2022. 3

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[15] Yannik Frisch, Moritz Fuchs, Antoine Sanner, Felix Anton Ucar, Marius Frenzel, Joana Wasielica-Poslednik, Adrian Gericke, Felix Mathias Wagner, Thomas Dratsch, and Anirban Mukhopadhyay. Synthesising rare cataract surgery samples with guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 354–364. Springer, 2023. 3

[16] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 3

[17] Fabio Garcea, Alessio Serra, Fabrizio Lamberti, and Lia Morra. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152:106391, 2023. 6

[18] Evgin Goceri. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56(11):12561–12605, 2023. 3, 6

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[22] 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. 4

[23] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. *arXiv preprint arXiv:2401.09603*, 2023. 6

[24] Alexander C Jenke, Sebastian Bodenstedt, Fiona R Kolbinger, Marius Distler, Jürgen Weitz, and Stefanie Speidel. One model to use them all: training a segmentation model with complementary datasets. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–9, 2024. 3, 6, 8

[25] Joanna Kaleta, Diego Dall'Alba, Szymon Płotka, and Przemysław Korzeniowski. Minimal data requirement for realistic endoscopic image generation with stable diffusion. *International journal of computer assisted radiology and surgery*, 19(3):531–539, 2024. 3

[26] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 3

[27] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023. 3

[28] Firas Khader, Gustav Mueller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baessler, Sebastian Foersch, et al. Medical diffusion: denoising diffusion probabilistic models for 3d medical image generation. *arXiv preprint arXiv:2211.03364*, 2022. 3

[29] 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. 6

[30] Qing Lyu and Ge Wang. Conversion between ct and mri images using diffusion and score-matching models. *arXiv preprint arXiv:2209.12104*, 2022. 3

[31] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. Surgical data science–from concepts toward clinical translation. *Medical image analysis*, 76:102306, 2022. 1

[32] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 21(2):195–212, 2024. 6

[33] Aldo Marzullo, Sara Moccia, Michele Catellani, Francesco Calimeri, and Elena De Momi. Towards realistic laparoscopic image generation using image-domain translation. *Computer Methods and Programs in Biomedicine*, 200:105834, 2021. 3

[34] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3, 4, 5

[35] 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, 3, 6, 7, 8

[36] Chinedu Innocent Nwoye, Deepak Alapatt, Tong Yu, Armine Vardazaryan, Fangfang Xia, Zixuan Zhao, Tong Xia, Fucang Jia, Yuxuan Yang, Hao Wang, et al. Cholectriplet2021: A benchmark challenge for surgical action triplet recognition. *Medical Image Analysis*, 86:102803, 2023. 6

[37] Dóra Papp, Renáta Nagyné Elek, and Tamás Haidegger. Surgical tool segmentation on the jigsaws dataset for autonomous image-based skill assessment. In *2022 IEEE 10th Jubilee International Conference on Computational Cybernetics and Cyber-Medical Systems (ICCC)*, pages 000049–000056. IEEE, 2022. 1

[38] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization, 2019. 6, 7, 8

[39] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[40] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022. 6

[41] Micha Pfeiffer, Isabel Funke, Maria R Robu, Sebastian Bodenstedt, Leon Strenger, Sandy Engelhardt, Tobias Roß, Matthew J Clarkson, Kurinchi Gurusamy, Brian 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, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*, pages 119–127. Springer, 2019. 1, 3, 5

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3

[44] Dominik Rivoir, Micha Pfeiffer, Reuben Docea, Fiona Kolbinger, Carina Riediger, Jürgen Weitz, and Stefanie Speidel. Long-term temporally consistent unpaired video translation from simulated surgical 3d data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3343–3353, October 2021. 1, 3, 5

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn 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. 3

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[47] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3752–3761, 2018. 1

[48] Paul Maria Scheikl, Stefan Laschewski, Anna Kisilenko, Tornike Davitashvili, Benjamin Müller, Manuela Capek,

Beat P Müller-Stich, Martin Wagner, and Franziska Mathis-Ullrich. Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery. In *Current Directions in Biomedical Engineering*, volume 6, page 20200016. De Gruyter, 2020. 1

[49] Lalithkumar Seenivasan, Sai Mitheran, Mobarakol Islam, and Hongliang Ren. Global-reasoned multi-task learning model for surgical scene understanding. *IEEE Robotics and Automation Letters*, 7(2):3858–3865, 2022. 1

[50] Lalith Sharan, Gabriele Romano, Sven Koehler, Halvar Kelm, Matthias Karck, Raffaele De Simone, and Sandy 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. 3

[51] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 3

[52] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4743, 2024. 4

[53] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1, 3

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[55] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2

[56] https://www.kaggle.com/c/prostate-cancer-grade-assessment/discussion/169108. Prostate cancer grade assessment (panda) challenge, 2020. 6

[57] https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/discussion/117330. Rsna intracranial hemorrhage detection, 2020. 6

[58] S Swaroop Vedula, Anand Malpani, Narges Ahmidi, Sanjeev Khudanpur, Gregory Hager, and Chi Chiung Grace Chen. Task-level vs. segment-level quantitative metrics for surgical skill assessment. *Journal of surgical education*, 73(3):482–489, 2016. 1

[59] Danush Kumar Venkatesh, Dominik Rivoir, Micha Pfeiffer, Fiona Kolbinger, Marius Distler, Jürgen Weitz, and Stefanie Speidel. Exploring semantic consistency in unpaired image translation to generate data for surgical applications. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–9, 2024. 1, 3

[60] Danush Kumar Venkatesh, Dominik Rivoir, Micha Pfeiffer, and Stefanie Speidel. Surgical-cd: Generating surgical images via unpaired image translation with latent consistency diffusion models. *arXiv preprint arXiv:2408.09822*, 2024. 3

[61] 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. 6, 7, 8

[62] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 6, 7

[63] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 6, 8

[64] Jihun Yoon, SeulGi Hong, Seungbum Hong, Jiwon Lee, Soyeon Shin, Bokyung Park, Nakjun Sung, Hayeong Yu, Sungjae Kim, SungHyun Park, et al. 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, 2022. 1, 3

[65] 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, 3, 4, 6, 7, 8

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver 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. 6

[67] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation, 2018. 6, 7