Towards Application Aligned Synthetic Surgical Image Synthesis

Danush Kumar Venkatesh[1,2] **Stefanie Speidel**[2,3]
[1]Department of Translational Surgical Oncology, NCT/UCC Dresden, a partnership between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden,HZDR, Germany
[2]Department of Translational Surgical Oncology, NCT/UCC Dresden, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden Germany
[3]The Centre for Tactile Internet with Human-in-the-Loop (CeTI), TUD Dresden
danushkumar.venkatesh@nct-dresden.de

Abstract

The scarcity of annotated surgical data poses a significant challenge for developing deep learning systems in computer-assisted interventions. While diffusion models can synthesize realistic images, they often suffer from data memorization, resulting in inconsistent or non-diverse samples that may fail to improve, or even harm, downstream performance. We introduce *Surgical Application-Aligned Diffusion* (SAADi), a new framework that aligns diffusion models with samples preferred by downstream models. Our method constructs pairs of *preferred* and *non-preferred* synthetic images and employs lightweight fine-tuning of diffusion models to align the image generation process with downstream objectives explicitly. Experiments on three surgical datasets demonstrate consistent gains of 7–9% in classification and 2–10% in segmentation tasks, with the considerable improvements observed for underrepresented classes. Iterative refinement of synthetic samples further boosts performance by 4–10%. Unlike baseline approaches, our method overcomes sample degradation and establishes task-aware alignment as a key principle for mitigating data scarcity and advancing surgical vision applications.

1 Introduction

Minimally invasive surgery (MIS) has gained increasing popularity in recent years due to its numerous benefits, including shorter recovery times, reduced postoperative pain, improved surgical dexterity and a lower risk of infection (Dagnino and Kundrat, 2024). The primary objective of MIS is to minimize the number and size of incisions; however, this also introduces challenges for surgeons. Procedures are typically performed by observing 2D endoscopic images on a monitor, which restricts the field of view and eliminates depth perception. These constraints highlight the need for computational methods that can function as assistive technologies, supporting surgeons during interventions. With the rapid advancements in deep learning (DL), there is an opportunity to develop such systems to improve surgical safety and efficiency. A key application area is surgical scene understanding, which involves tasks such as identifying anatomical structures, segmenting target tissues, or issuing warnings about critical structures (e.g., arteries) near surgical instruments. For instance, DL methods can help localize hidden tumor tissue by overlaying 3D anatomical structures onto the intra-operative scene, thereby providing surgeons with crucial real-time guidance. Such visualizations have proven to improve surgical outcomes (Wagner et al., 2012).

Despite their promise, current DL approaches in the surgical domain rely on supervised learning, which requires large and diverse annotated datasets. Acquiring such surgical datasets are particularly

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

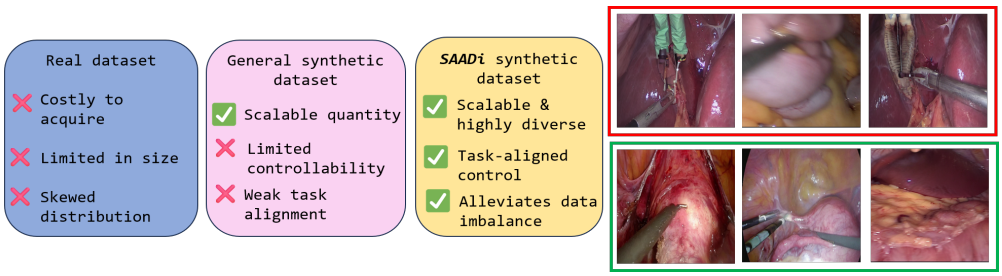

Figure 1: Comparison of real, general synthetic, and our (SAADi) datasets. Real datasets are often limited and imbalanced, and baseline generative models frequently yield inconsistent samples, such as blurred images (highlighted in the red box). In contrast, our method (SAADi) produces task-aligned, diverse, and realistic synthetic images that better capture anatomical structures (green box).

challenging due to patient privacy concerns, logistical constraints in the operating room, and strict data protection regulations across clinical centers (Maier-Hein et al., 2017). This creates a paradox in the field: while large-scale datasets are essential for building robust DL models, access to such data remains scarce.

To address the scarcity of real surgical data, increasing attention has been given to the use of synthetic datasets. Recent advances in generative modeling, particularly diffusion models (DMs) (Ho et al., 2020, Dhariwal and Nichol, 2021), have shown remarkable ability to produce high-quality, photorealistic images. In the surgical domain, DMs have been trained on real-world datasets to generate clinically plausible synthetic images, which can then be combined with real data to enhance the performance of downstream deep learning (DL) models (Nwoye et al., 2025, Frisch et al., 2023, Venkatesh et al., 2025). However, the use of synthetic data poses several key challenges. First, controlling the composition of generated images is crucial, as this directly influences the effectiveness of downstream models. Second, due to the typically small size of surgical datasets, DMs are prone to overfitting, often replicating samples from the training set (Somepalli et al., 2023, Chen et al., 2024) or producing undesirable configurations (see Fig. 1). Incorporating such low-quality or redundant samples into training pipelines has been shown to provide limited benefit, and in some cases may even degrade performance (Alaa et al., 2022, Azizi et al., 2023), thereby undermining the purpose of synthetic data augmentation. These limitations raise critical questions: how can we effectively control DMs to ensure that the generated data contributes meaningfully to downstream tasks, and how can we guarantee that the resulting synthetic data is indeed beneficial? This motivates the central goal of our work: to develop a diffusion-based framework that generates synthetic surgical data that is directly beneficial for downstream models.

To this end, we propose *SAADi*, **S**urgical **A**pplication-**A**ligned **Di**ffusion, a framework for synthetic image generation that produces not only realistic samples but also data explicitly aligned with downstream task performance. To the best of our knowledge, this is the first work to introduce application-aligned diffusion for surgical image synthesis. We build on Stable Diffusion (SD) (Rombach et al., 2022), a latent diffusion model, and introduce a framework in which the preferences of a downstream model explicitly guide image generation. Our approach is inspired by Diffusion-Direct Preference Optimization (DDPO) (Wallace et al., 2024), where DMs are fine-tuned on human preference data to improve aesthetic quality and prompt adherence. In contrast, we replace human supervision with automatically constructed preference pairs: we generate a large set of synthetic images using SD (trained on real surgical data), evaluate them with a downstream model (e.g., classification or detection), and retain or discard samples based on a predefined threshold. These preference pairs of *preferred* and *non-preferred* instances are constructed solely from the synthetic data and subsequently used to fine-tune SD with LoRA (Hu et al., 2022), introducing only minimal overhead. Our approach directly addresses the dataset scarcity paradox in surgical science by ensuring that synthetic data is diverse, clinically relevant, and actively improves downstream surgical computer vision tasks. We summarize our contributions as follows:

1. We introduce **SAADi**, the first application-aligned diffusion framework for surgical image synthesis, explicitly designed to improve downstream performance.

2. We introduce an innovative preference-pairing strategy that uses only synthetic images, enabling alignment of diffusion models through lightweight LoRA fine-tuning without additional human supervision.
3. We provide extensive experiments on three surgical datasets and two key downstream tasks (classification and segmentation), demonstrating consistent improvements, with performance gains of up to 15%, particularly in underrepresented classes.

2 Related Work

2.1 Diffusion models

Diffusion models (DMs) (Sohl-Dickstein et al., 2015) have revolutionized image synthesis with their superior image quality compared to generative adversarial networks (GANs) (Goodfellow et al., 2014). Particularly, Latent Diffusion Models (LDMs) (Rombach et al., 2022) extend this framework by performing the diffusion process in a compressed latent space, thereby significantly reducing computational costs while maintaining high image fidelity. Stable Diffusion (SD) (Rombach et al., 2022) is a large-scale implementation of LDMs trained on natural image datasets, where image generation is conditioned on text prompts. This conditioning is achieved by encoding text inputs into latent vectors using pre-trained language models such as CLIP (Radford et al., 2021). Owing to its strong generative capabilities and open-source availability, SD has emerged as one of the most widely adopted LDM variants. In this work, we build our framework on top of the SD model.

2.2 Synthetic surgical images

Laparoscopic image synthesis has been focused predominantly on image-to-image (I2I) translation methods. For example, computer-simulated surgical images, phantom data, and segmentation maps have been employed with GANs to synthesize realistic surgical images and videos (Chen et al., 2019, Sankaranarayanan et al., 2018, Pfeiffer et al., 2019, Rivoir et al., 2021, Venkatesh et al., 2024a, Yoon et al., 2022, Sharan et al., 2021, Marzullo et al., 2021). More recently, Stable Diffusion (SD)-based I2I methods were explored in (Kaleta et al., 2024, Venkatesh et al., 2024b, Martyniak et al., 2025, Venkatesh et al., 2025). Importantly, large quantities of synthetic images are generated, although their quality can sometimes be detrimental for the downstream task (Venkatesh et al., 2024a, Frisch et al., 2023). Beyond surgical applications, diffusion models have rapidly gained traction for medical image generation, particularly in MRI and CT (Dorjsembe et al., 2022, Khader et al., 2022, Lyu and Wang, 2022). However, these images differ broadly in modality from the surgical images.

2.3 Controllable generation

Effective control of diffusion models (DMs) is critical for customizing generated images. Text-based editing has been explored through prompt engineering and manipulation of CLIP features (Avrahami et al., 2022, Brooks et al., 2023, Gafni et al., 2022, Hertz et al., 2022, Kawar et al., 2023), but such approaches are less suitable in the surgical domain where detailed textual descriptions are scarce. In contrast, spatial control can be achieved with conditional images processed via adapter networks resembling the U-Net backbone in latent diffusion models (LDMs) (Zhang et al., 2023), or through lightweight adapters such as T2I-Adapter (Mou et al., 2023). While these methods provide strong controllability, they require substantial computational resources, large annotated datasets, and long training times for surgical adaptation. In this work, we take a simpler approach: we leverage existing datasets and implicitly guide the generation process by selecting task-relevant samples from the training distribution, thereby directly improving downstream performance.

3 Methodology

In this section, we present our approach (SAADi) for generating synthetic data that is explicitly aligned with downstream tasks. The framework consists of two main stages. In the first stage, we train a Stable Diffusion (SD) model on real-world surgical datasets to learn the underlying data distribution. In the second stage, we generate synthetic samples from the trained model and construct preference pairs to guide alignment with the downstream task. To achieve this, we employ a selection

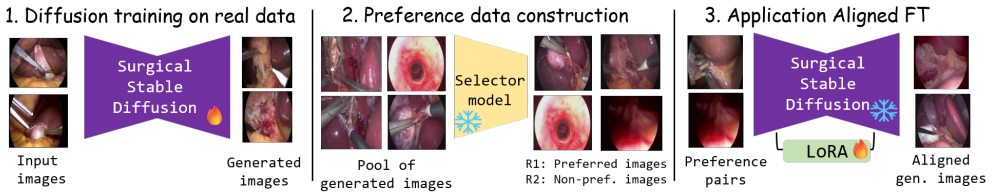

Figure 2: **Overview of the SAADi framework.** In Stage 1, we train a Surgical Stable Diffusion (SSD) model on real surgical images with text prompts and generate a large pool of synthetic images. In Stage 2, a selector model evaluates these synthetic images and separates them into *preferred* and *non-preferred* sets. In Stage 3, the resulting preference pairs are used to fine-tune the SSD model with LoRA adapters, aligning it with the preferences of the downstream task. The fine-tuned model is then used to sample task-aligned synthetic images.

model trained for the specific downstream task and run inference on the generated data. Based on a predefined threshold, each synthetic sample is categorized into a *preferred* or *non-preferred* set. Using this preference dataset, we fine-tune the SD model obtained from the first stage and subsequently sample from the refined model to obtain diverse and task-relevant synthetic images. An overview of the pipeline is shown in Fig. 2.

3.1 Diffusion models

Given samples from a data distribution $q(x_0)$, and a noise scheduling function α_t and σ_t (as defined in (Ho et al., 2020)), diffusion models are generative models $p_\theta(x_0)$ trained to progressively denoise corrupted data. The training objective is defined as

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{x_0,\epsilon,t,x_t}\left[\,\|\epsilon - \epsilon_\theta(x_t, t, P)\|_2^2\,\right], \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $t \sim \mathcal{U}(0, T)$, and $x_t \sim q(x_t \mid x_0) = \mathcal{N}(x_t; \alpha_t x_0, \sigma_t^2 \mathbf{I})$. Here, P denotes the text prompt. In the Stable Diffusion (SD) model (Rombach et al., 2022), an encoder E maps an input image x_0 into a latent space where the diffusion process is carried out, and a decoder D reconstructs the denoised latent back into the pixel space. We call this model *Surgical Stable Diffusion* (SSD) and sample images from it using P.

Similarly, we also employ the *Surgical Stable Inpaint (SSI)* model from Venkatesh et al. (2025), which is trained for inpainting-based synthesis. Given an image x_0 and a mask m, the model is trained to synthesize realistic texture within the masked region. Formally, in the forward process, the masked input $\tilde{x}_t$ is constructed as $\tilde{x}_t = x_t \odot m + x_0 \odot (1 - m)$, where x_t denotes the noised image at timestep t. The denoising network ϵ_θ is trained with a modified objective:

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

Since the training objective is localized to the masked regions, SSI learns to generate organ-specific textures conditioned jointly on the segmentation mask and the text signal.

3.2 Preference data creation

Let $\mathcal{R} = \{r_i\}_{i=1}^N$ denote the set of real surgical images, and $\mathcal{G}_s = \{g_j\}_{j=1}^M$ the pool of synthetic images generated by the diffusion model (SSD or SSI). A downstream model $f(\cdot)$ is trained on $\mathcal{R}$ to map images to task-specific outputs (e.g., class labels or segmentation masks). To construct preference pairs, each synthetic sample $g_j \in \mathcal{G}_s$ is evaluated using f, producing a score $s_j = f(g_j)$. Given a predefined threshold h, we separate $\mathcal{G}_s$ into preferred and non-preferred subsets:

$$\mathcal{G}_p = \{g_j \in \mathcal{G}_s \mid s_j \geq h\}, \quad \mathcal{G}_{np} = \{g_j \in \mathcal{G}_s \mid s_j < h\}. \tag{3}$$

From these partitions, we define the preference dataset as $\mathcal{D}_{\text{pref}} = \{(g_p, g_{np}) \mid g_p \in \mathcal{G}_p,\ g_{np} \in \mathcal{G}_{np}\}$, which is subsequently used to fine-tune the diffusion model for alignment.

3.3 Application aligned optimization

To generate synthetic data from the SSD or SSI models (ϵ_{ref}) that is beneficial for downstream tasks, it is necessary to align these models with the preferred samples identified by the downstream model. Our

Table 1: **Overview of the datasets**. The different anatomies, tools and diffusion baselines used for evaluation is listed here.

Dataset	Procedure	Train/Test	#Classes	Task: Anatomy	Task: Tools	Diffusion Baseline
LapGyn (Leibetseder et al., 2018)	Gynecological surgery	1014 / 438	5 organs + tools	Colon, Liver, Ovary, Oviduct, Uterus	Graspers, etc.	SSD (Surgical Stable Diffusion)
Endoscapes (Murali et al., 2023)	Cholecystectomy	343 / 100	5 organs + tools	Cystic plate, Calot's triangle, Cystic artery, Cystic duct, Gallbladder	Surgical tools	SSI (Surgical Stable Inpaint)
AutoLaparo (Wang et al., 2022)	Hysterectomy	1100 / 500	1 organ + 4 tools	Uterus	Grasping forceps, Dissecting forceps, LigaSure, Electric hook	SSI (Surgical Stable Inpaint)

approach, **SAADi** replaces human labels in DDPO with downstream task evaluations and construct preference pairs from synthetic data, as described in the previous step. The objective is to learn a new model $\epsilon_{p\theta}$ whose generations are explicitly aligned with these preferences by fine-tuning on pairs of *preferred* and *non-preferred* synthetic images. Let $(x_0^p, x_0^n) \sim \mathcal{D}_{\text{pref}}$ be a preferred/non-preferred pair. For $t \sim \mathcal{U}(0, T)$, sample $\epsilon^p, \epsilon^n \sim \mathcal{N}(0, \mathbf{I})$ and the forward diffusion is defined via,

$$x_t^p = \alpha_t x_0^p + \sigma_t \epsilon^p, \qquad x_t^n = \alpha_t x_0^n + \sigma_t \epsilon^n.$$

Let per-preference sample differences (relative to a fixed reference denoiser ϵ_{ref}) be defined as

$$\Delta_\theta^p(t) = \|\epsilon^p - \epsilon_{p\theta}(x_t^p, t)\|_2^2 - \|\epsilon^p - \epsilon_{\text{ref}}(x_t^p, t)\|_2^2, \quad \Delta_\theta^n(t) = \|\epsilon^n - \epsilon_{p\theta}(x_t^n, t)\|_2^2 - \|\epsilon^n - \epsilon_{\text{ref}}(x_t^n, t)\|_2^2.$$

With the logistic function $\sigma(u) = 1/(1 + e^{-u})$, we define the loss as

$$\mathcal{L}_{\text{SAADi}}(\theta) = -\,\mathbb{E}_{(x_0^p, x_0^n) \sim \mathcal{D}_{\text{pref}},\, t,\, \epsilon^p,\, \epsilon^n} \left[\log \sigma\big(-\beta \left(\Delta_\theta^p(t) - \Delta_\theta^n(t) \right) \big) \right], \tag{4}$$

where β is a weighting term. For additional details the readers can refer to Wallace et al. (2024). Once trained, the model generates synthetic data that are explicitly aligned with downstream task preferences, ensuring their utility for the downstream task.

4 Experiments

In this section, we present our experimental setup, including the datasets and downstream models. We evaluate our approach on three surgical datasets across two tasks: (i) multi-class classification of anatomical structures and surgical tools, and (ii) binary segmentation of anatomical structures and tools. Our primary objective is to assess the utility of the generated synthetic data by measuring its impact on downstream performance. An overview of the datasets and tasks is provided in Tab. 1.

Evaluation scheme. We design three evaluation settings to analyze the impact of synthetic datasets on downstream performance:

1. **Baseline comparison.** We generate synthetic data using the baseline models (SSD and SSI) and compare their performance against our approach (SAADi) across the downstream tasks described earlier. To ensure fairness, we add the same number of synthetic samples as real samples present in the training set.
2. **Data scaling.** We study the scaling behavior of synthetic data by adding multiples of the training set size ($2\times, 3\times, 4\times$) in synthetic samples from each method to the real dataset. This experiment assesses the impact of increasing synthetic data on downstream performance.
3. **Iterative refinement.** We further investigate whether synthetic data quality can be improved through refinement. In the first round, we train a downstream model on real data combined with synthetic images generated by SAADi. This trained model is then used as the selection model to re-score the initial pool of synthetic samples. Based on this updated scoring, we perform a second round of SAADi fine-tuning and generate a new set of images. The downstream models are subsequently evaluated on this refined dataset. This iterative process demonstrates that once useful synthetic data is introduced, downstream models improve, enabling the selection of stronger and more informative samples in subsequent rounds.

Surgical datasets For the multi-class classification task, we use the LapGyn (LG) dataset (Leibetseder et al., 2018), which consists of laparoscopic gynecological procedures. The dataset includes five anatomical structures: colon, liver, ovary, oviduct, and uterus. The training set contains 1014

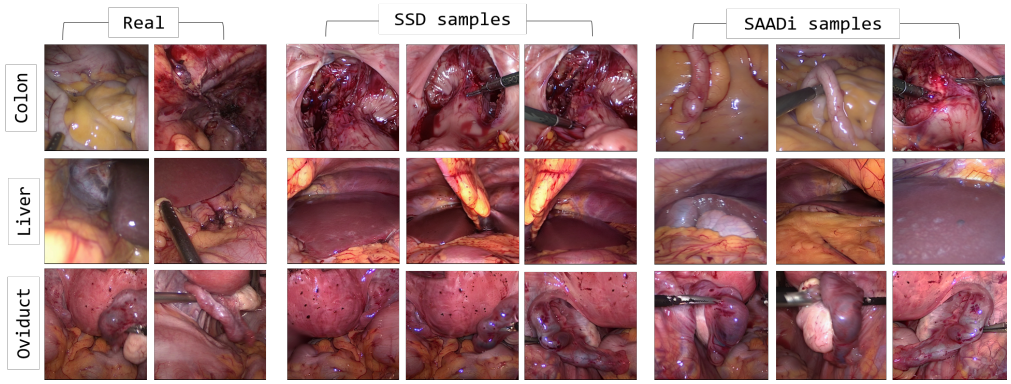

Figure 3: **Qualitative results** on the LapGyn dataset. Synthetic images from the baseline SSD model show limited diversity, with minimal variation across samples. In contrast, images generated by our method (SAADi) are diverse, application-aligned, and visually realistic, closely resembling real data.

images, while the test set contains 438 images. Notably, the dataset is imbalanced across classes. The task is defined as classifying anatomical structures and surgical tools in a given surgical scene.

For segmentation, we employ two datasets. First, the Endoscapes (Ed) dataset (Murali et al., 2023), where we use only the segmentation split comprising 343 training images and 100 test images from laparoscopic cholecystectomy. This dataset allows us to evaluate performance in a low-data regime. The annotated structures include the cystic plate, Calot's triangle, cystic artery, cystic duct, gallbladder, and surgical tools.

Second, the Autolaparo (AL) dataset (Wang et al., 2022), which contains data from 21 patients undergoing laparoscopic hysterectomy, with 1100 images for training and 500 images for testing. The annotated classes include surgical tools such as grasping forceps, dissecting forceps, LigaSure, and electric hook, along with the uterus as the anatomical structure. Since patient diversity is a crucial factor in the surgical domain, we utilize a held-out, patient-specific test set to assess the models.

These datasets were chosen to investigate the role of synthetic data under both class-imbalanced and resource-constrained conditions.

Baselines & Models As baselines for diffusion models, we trained the Surgical Stable Diffusion (SSD) model on the real surgical images from the LG dataset with the text prompts constructed as *"An image of <organ/tool> in laparoscopic gynecological surgery"*. We sample images from this model and add them with real dataset as a baseline against *SAADi*.

For the segmentation datasets, we used an inpainting model, *Surgical Stable Inpaint* (SSI) (Venkatesh et al., 2025), as the task involves generating organ and tool textures conditioned on masks. All diffusion models were trained using only the training splits of each dataset for 3000 steps with AdamW (Loshchilov and Hutter, 2017). For the inpainting models, we used 30 denoising steps during the generation process. Subsequently, *SAADi* fine-tuning was performed for 1500 steps, requiring approximately 8 minutes of training on a single 24GB GPU. We change the prompts correspondingly for each dataset and organ or tool.

For the downstream tasks, we adopted three architectures for classification: ResNet-50 (He et al., 2016), ConvNeXT-S (Liu et al., 2022), and ViT-S (Dosovitskiy et al., 2020); and three for segmentation: DeepLabV3 (DV3) (Chen et al., 2017), SegFormer (Xie et al., 2021), and UPerNet (Xiao et al., 2018). To mitigate class imbalance, we applied pixel weighting and inverse-frequency balancing, combined with standard data augmentations, during training on real datasets. Employing a diverse set of architectures allowed us to reduce bias toward any single model. For evaluation, we report the F1 and Dice score for classification and segmentation tasks respectively.

Table 2: **Classification of anatomies in the LapGyn dataset**. Reported values are F1 scores. Imbalanced classes are highlighted, and the best scores are shown in blue. The addition of synthetic images from our approach (SAADi) improves performance by approximately 4–9% compared to training on the real dataset alone.

Method	Training data	Colon	Liver	Ovary	Oviduct	Uterus	Mean
ResNet-50	Only Real	0.10	0.22	0.38	0.06	0.26	0.20
	Real + SSD	0.10	0.25	0.37	0.09	0.23	0.21 (↑1%)
	Real + SAADi	0.15	0.24	0.40	0.13	0.31	0.24 (↑4%)
ConvNeXT/S	Only Real	0.10	0.23	0.30	0.15	0.42	0.24
	Real + SSD	0.11	0.23	0.35	0.16	0.41	0.24 (–)
	Real + SAADi	0.16	0.26	0.41	0.22	0.48	0.31 (↑7%)
ViT/S	Only Real	0.10	0.25	0.32	0.11	0.37	0.23
	Real + SSD	0.14	0.23	0.34	0.22	0.39	0.26 (↑3%)
	Real + SAADi	0.19	0.26	0.45	0.24	0.38	0.32 (↑9%)

5 Results & Discussion

Addition of synthetic samples The qualitative results are shown in Fig. 3. The results for the classification task on the LG dataset are presented in Tab. 2. We observe that adding synthetic samples from the baseline model provides modest improvements in classification performance. In contrast, our approach, SAADi, yields substantial gains, with overall improvements of 7% and 9% for the ConvNeXT and ViT architectures, respectively. Notably, SAADi achieves significant improvements in the imbalanced classes, including an increase of more than 13% for the oviduct class. Although the same number of samples is added across all baselines, the results demonstrate that SAADi is particularly effective in addressing class imbalance and improving performance on underrepresented categories.

Table 3: **Anatomy segmentation in the Endoscapes dataset.** Synthetic samples generated by the baseline method (SSI) often introduce degenerate cases, leading to degraded performance. In contrast, images produced by our approach (SAADi) yield consistent improvements across all evaluated downstream models. Dice scores is reported.

Method	Training data	Cystic plate	Calot triangle	Cystic artery	Cystic duct	Gall bladder	Tool	Mean
DV3	Only Real	0.38	0.37	0.42	0.47	0.70	0.61	0.49
	Real + SSI	0.36	0.34	0.43	0.48	0.64	0.66	0.48 (↓1%)
	Real + SAADi	0.41	0.36	0.44	0.50	0.73	0.61	0.51 (↑2%)
Segformer	Only Real	0.40	0.36	0.30	0.41	0.68	0.62	0.46
	Real + SSI	0.39	0.36	0.38	0.37	0.64	0.55	0.44 (↓2%)
	Real + SAADi	0.41	0.49	0.40	0.52	0.70	0.63	0.52 (↑6%)
UPerNet	Only Real	0.33	0.27	0.37	0.38	0.54	0.41	0.38
	Real + SSI	0.41	0.42	0.32	0.39	0.52	0.48	0.42 (↑4%)
	Real + SAADi	0.40	0.44	0.46	0.40	0.57	0.61	0.48 (↑10%)

Tab. 3 and Tab. 4 show the results of the segmentation models on the Ed and AL datasets. We observe that adding synthetic samples from the SSI baseline leads to a decline in performance for both SegFormer and DeepLabV3, suggesting that image generation with structure-specific constraints alone is insufficient to meet the requirements of downstream models. A modest improvement of 4% is observed with UPerNet, further highlighting the importance of evaluating synthetic data across multiple architectures.

In contrast, synthetic samples from SAADi achieve the best performance in five out of six classes, with gains ranging from 2–10%. Similar trends are observed on the AL dataset, where SAADi provides consistent improvements while SSI degrades performance. These results highlight an important observation: simply adding the same number of synthetic samples as the training set may

Table 4: **Segmentation of tools in the Autolaparo dataset.** The SSI model fails to generate valid synthetic data, and its inclusion with the real dataset reduces performance, particularly for surgical tools. In contrast, our approach (SAADi) provides smaller yet consistent benefits when synthetic images are added.

Method	Training data	Grasping forceps	Liga Sure	Dissecting forceps	Electric hook	Uterus	Mean
DV3	Only Real	0.61	0.90	0.74	0.59	0.73	0.71
	Real + SSI	0.56	0.48	0.53	0.54	0.74	0.57 (↓14%)
	Real + SAADi	0.63	0.90	0.76	0.61	0.76	0.74 (↑3%)
Segformer	Only Real	0.64	0.92	0.76	0.65	0.77	0.75
	Real + SSI	0.65	0.87	0.62	0.62	0.78	0.71 (↓4%)
	Real + SAADi	0.67	0.91	0.77	0.64	0.80	0.76 (↑1%)
UPerNet	Only Real	0.60	0.91	0.74	0.62	0.71	0.72
	Real + SSI	0.63	0.58	0.54	0.55	0.74	0.61 (↓11%)
	Real + SAADi	0.61	0.93	0.76	0.63	0.76	0.74 (↑2%)

not always yield large performance gains. Instead, aligning synthetic data with the downstream task, as in SAADi, is crucial for maximizing its utility. *This further confirms that application-aware alignment is more critical for the surgical domain.*

Data scaling behavior The scaling behavior of synthetic samples on the Ed dataset with the DV3 model is shown in Fig. 4, with additional results provided in the suppl. material. A consistent trend is that synthetic samples from SAADi yield steady improvements in performance across most classes. In contrast, samples from the SSI baseline lead to performance degradation beyond a specific scale for four classes, consistent with the observations of Azizi et al. (2023). Another key observation is the plateauing of performance when more than $3\times$ or $4\times$ synthetic data is added. This saturation effect can be attributed to the fact that the diversity within the training distribution has already been extensively captured, and the generated samples largely reflect this existing diversity. To further improve performance, future work could explore incorporating variations in shape and texture depending on the inductive biases of the downstream model. *Our findings highlight the need for both task-alignment and data diversity in synthetic data for surgical applications.*

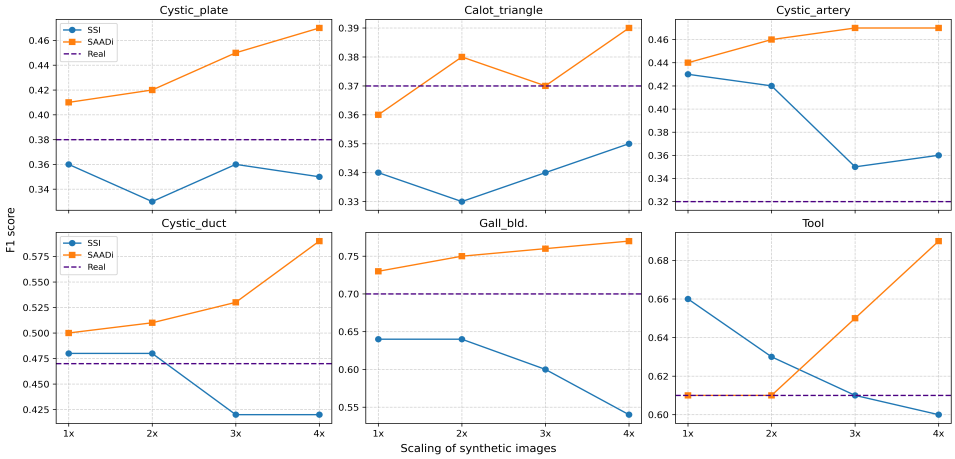

Figure 4: **Scaling of synthetic samples** for anatomy segmentation in the Endoscapes dataset. Adding synthetic samples from SAADi results in continuous performance improvements across classes. In contrast, samples from the SSI baseline lead to a decline in dice scores, reflecting inconsistencies from application-agnostic image generation.

Iterative refinement of synthetic data Tab. 5 indicates the results of a second round of refinement of synthetic samples generated by our approach, SAADi, on the Ed dataset. This additional step yields average performance improvements of 4–10% across the different downstream models. The gains are modest for surgical tools, while the largest improvements are observed for the cystic plate

Table 5: **Iterative refinement of SAADi samples.** Results are reported on the Endoscapes dataset. 1^{st} denotes the first round of SAADi fine-tuning, and 2^{nd} indicates the second round with iterative refinement of synthetic samples. Refinement consistently improves performance across different models, with gains in the range of 4–10%.

Method	Training data	Cystic plate	Calot triangle	Cystic artery	Cystic duct	Gall bladder	Tool	Mean
DV3	Only Real	0.38	0.37	0.42	0.47	0.70	0.61	0.49
	Real + SAADi (1^{st})	0.41	0.36	0.44	0.50	0.73	0.61	0.51 (↑2%)
	Real + SAADi (2^{nd})	0.42	0.36	0.47	0.53	0.73	0.64	0.53 (↑4%)
Segformer	Only Real	0.40	0.36	0.30	0.41	0.68	0.62	0.46
	Real + SAADi (1^{st})	0.41	0.49	0.40	0.52	0.70	0.63	0.52 (↑6%)
	Real + SAADi (2^{nd})	0.43	0.52	0.41	0.52	0.73	0.64	0.54 (↑8%)
Upernet	Only Real	0.33	0.27	0.37	0.38	0.54	0.41	0.38
	Real + SAADi (1^{st})	0.40	0.44	0.46	0.40	0.57	0.61	0.48 (↑4%)
	Real + SAADi (2^{nd})	0.41	0.45	0.48	0.46	0.61	0.64	0.51 (↑13%)

and cystic duct classes. These findings suggest that refinement can be model-dependent, as each downstream architecture may exhibit its own inductive biases. Further exploration of multi-stage refinement could provide deeper insights into the limitations of aligning synthetic data generation with downstream tasks, which we leave for future work. Additional results are provided in the suppl. material. *Overall, these results highlight iterative refinement as a promising strategy for enhancing the effectiveness of application-aligned synthetic data.*

Limitations Although our approach is capable of generating synthetic images that benefit downstream tasks, certain limitations remain. First, SAADi requires a base diffusion model to generate the initial cohort of synthetic data. As a result, any biases present in the base model are propagated during fine-tuning and cannot be eliminated. Second, our approach relies on a selection model for classifying or segmenting synthetic images, which in turn requires annotated data. While self-supervised models may help alleviate this dependency, further investigation is needed. Third, although we employ lightweight fine-tuning, this step still adds to the computational cost of generation, which may hinder real-time applications. Future work could explore integrating feedback-guided approach (Askari-Hemmat et al., 2025) with application alignment to reduce these overheads and further improve real-time applicability.

6 Conclusion

In this work, we presented **SAADi**, an application-aligned diffusion framework for surgical image synthesis that explicitly adapts generation to downstream tasks. Instead of relying on human feedback, SAADi leverages downstream model evaluations to fine-tune diffusion models on pairs of preferred and non-preferred samples, producing synthetic data that is both realistic and task-relevant. Comprehensive experiments on three surgical datasets demonstrate consistent improvements in both classification and segmentation tasks, with notable gains for underrepresented classes. Furthermore, iterative refinement yields additional improvements, highlighting the importance of alignment beyond simple dataset scaling. Taken together, these results establish preference alignment as a promising direction for generating clinically indicative synthetic data and mitigating the challenge of data scarcity in surgical data science.

Acknowledgement This work is partly supported by BMFTR (Federal Ministry of Research, Technology and Space) in DAAD project 57616814 (SECAI, School of Embedded Composite AI, https://secai.org/) as part of the program Konrad Zuse Schools of Excellence in Artificial Intelligence. Also partially funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) 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 Technische Universität Dresden.

References

A. Alaa, B. Van Breugel, E. S. Saveliev, and M. Van Der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International conference on machine learning*, pages 290–306. PMLR, 2022.

R. Askari-Hemmat, M. Pezeshki, E. Dohmatob, F. Bordes, P. Astolfi, M. Hall, J. Verbeek, M. Drozdzal, and A. Romero-Soriano. Improving the scaling laws of synthetic data with deliberate practice. *arXiv preprint arXiv:2502.15588*, 2025.

O. Avrahami, D. Lischinski, and O. 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.

S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

T. Brooks, A. Holynski, and A. 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.

C. Chen, D. Liu, and C. Xu. Towards memorization-free diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8434, 2024.

L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Y. Chen, W. Li, X. Chen, and L. V. 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.

G. Dagnino and D. Kundrat. Robot-assistive minimally invasive surgery: trends and future directions. *International Journal of Intelligent Robotics and Applications*, 8(4):812–826, 2024.

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

Z. Dorjsembe, S. Odonchimed, and F. Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical Imaging with Deep Learning*, 2022.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Y. Frisch, M. Fuchs, A. Sanner, F. A. Ucar, M. Frenzel, J. Wasielica-Poslednik, A. Gericke, F. M. Wagner, T. Dratsch, and A. 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.

O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

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

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

J. Kaleta, D. Dall'Alba, S. Płotka, and P. 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.

B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. 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.

F. Khader, G. Mueller-Franzes, S. T. Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baessler, S. Foersch, et al. Medical diffusion: denoising diffusion probabilistic models for 3d medical image generation. *arXiv preprint arXiv:2211.03364*, 2022.

A. Leibetseder, S. Petscharnig, M. J. Primus, S. Kletz, B. Münzer, K. Schoeffmann, and J. Keckstein. Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology. In *Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018*, pages 357–362. ACM, 2018. doi: 10.1145/3204949.3208127. URL https://doi.org/10.1145/3204949.3208127.

Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Q. Lyu and G. Wang. Conversion between ct and mri images using diffusion and score-matching models. *arXiv preprint arXiv:2209.12104*, 2022.

L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017.

S. Martyniak, J. Kaleta, D. Dall'Alba, M. Naskręt, S. Płotka, and P. Korzeniowski. Simuscope: Realistic endoscopic synthetic dataset generation through surgical simulation and diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4268–4278. IEEE, 2025.

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.

C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.08453.

A. Murali, D. Alapatt, P. Mascagni, A. Vardazaryan, A. Garcia, N. Okamoto, G. Costamagna, D. Mutter, J. Marescaux, B. Dallemagne, et al. The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. *arXiv preprint arXiv:2312.12429*, 2023.

C. I. Nwoye, R. Bose, K. Elgohary, L. Arboit, G. Carlino, J. L. Lavanchy, P. Mascagni, and N. Padoy. Surgical text-to-image generation. *Pattern Recognition Letters*, 190:73–80, 2025.

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, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*, pages 119–127. Springer, 2019.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3343–3353, October 2021.

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.

S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. 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.

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.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.

D. K. Venkatesh, D. Rivoir, M. Pfeiffer, F. Kolbinger, M. Distler, J. Weitz, and S. 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, 2024a.

D. K. Venkatesh, D. Rivoir, M. Pfeiffer, and S. Speidel. Surgical-cd: Generating surgical images via unpaired image translation with latent consistency diffusion models. In *European Conference on Computer Vision*, pages 218–235. Springer, 2024b.

D. K. Venkatesh, D. Rivoir, M. Pfeiffer, F. Kolbinger, and S. Speidel. Data augmentation for surgical scene segmentation with anatomy-aware diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2280–2290. IEEE, 2025.

O. Wagner, M. Hagen, A. Kurmann, S. Horgan, D. Candinas, and S. Vorburger. Three-dimensional vision enhances task performance independently of the surgical method. *Surgical endoscopy*, 26 (10):2961–2968, 2012.

B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

Z. Wang, B. Lu, Y. Long, F. Zhong, T.-H. Cheung, Q. Dou, and Y. Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 486–496. Springer, 2022.

T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

J. Yoon, S. Hong, S. Hong, J. Lee, S. Shin, B. Park, N. Sung, H. Yu, S. Kim, S. 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.

L. Zhang, A. Rao, and M. 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.

A Supplementary Material

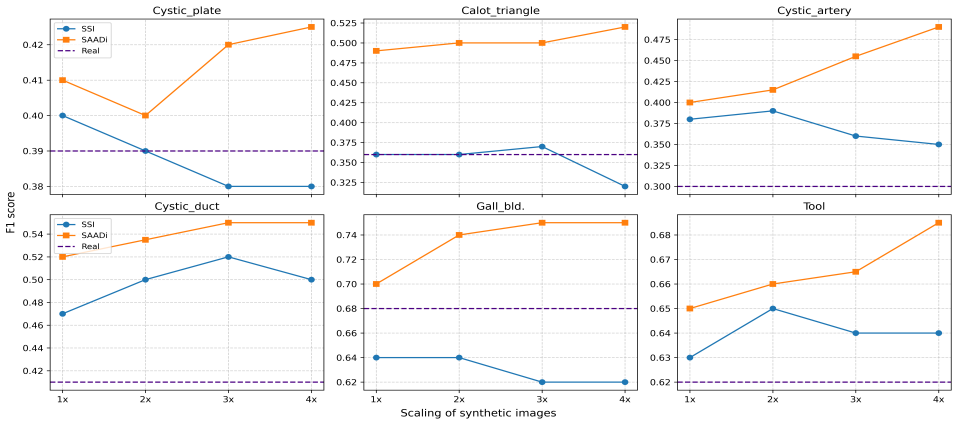

Figure 5: Scaling of synthetic samples for anatomy segmentation in the Endoscapes dataset with Segformer model. The addition of synthetic samples from SSI model leads to performance improvement in four out of six classes. The synthetic samples from SAADi leads to continuous gains in dice scores across different classes.

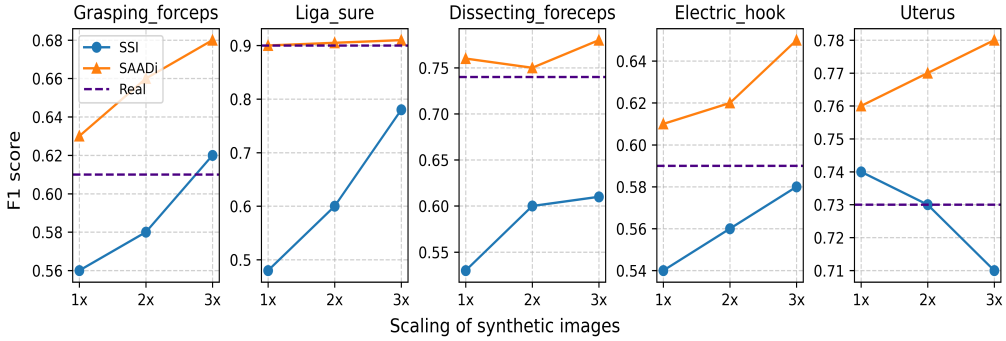

Figure 6: Scaling of synthetic samples for tool and anatomy segmentation on the Autolaparo dataset with DV3 model. For this dataset, in constract to Fig. 4 and Fig. 5 the synthetic samples from the baseline model leads to continuous increase of F1 scores for all the tools. However, the samples from our appraoch SAADi are more aligned to task and hence they outperform the baseline and shows the best scores for all the classes. For the Liga sure instrument we notice the scores to plateua beyond $1\times$.

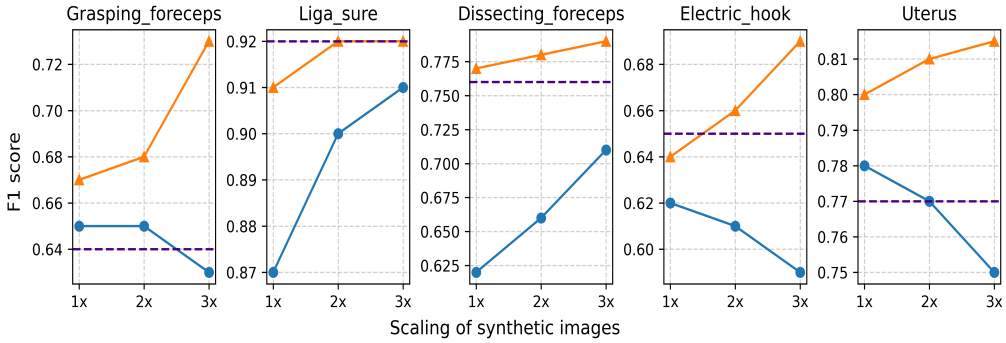

Figure 7: Scaling of synthetic samples for tool and organ segmentation on the Autolaparo dataset.