ControlPolypNet: Towards Controlled Colon Polyp Synthesis for Improved Polyp Segmentation

Vanshali Sharma[1], Abhishek Kumar[2], Debesh Jha[3], M.K. Bhuyan[1], Pradip K. Das[1], and Ulas Bagci[3]

[1]Indian Institute of Technology Guwahati, India
[2]Eli Lilly and Company, India
[3]Northwestern University, USA

{vanshalisharma, mkb, pkdas}@iitg.ac.in, abhishek.nkumar@lilly.com, {debesh.jha, ulas.bagci}@northwestern.edu

Abstract

*In recent years, generative models have been very popular in medical imaging applications because they generate realistic-looking synthetic images, which is crucial for the medical domain. These generated images often complement the hard-to-obtain annotated authentic medical data because acquiring such data requires expensive manual effort by clinical experts and raises privacy concerns. Moreover, with recent diffusion models, the generated data can be controlled using a conditioning mechanism, simultaneously ensuring diversity within synthetic samples. This control can allow experts to generate data based on different scenarios, which would otherwise be hard to obtain. However, how well these models perform for colonoscopy still needs to be explored. Do they preserve clinically significant information in generated frames? Do they help in downstream tasks such as polyp segmentation? Therefore, in this work, we propose ControlPolypNet, a novel stable diffusion based framework. We control the generation process (polyp size, shape and location) using a novel custom-masked input control, which generates images preserving important endoluminal information. Additionally, our model comprises a detection module, which discards some of the generated images that do not possess lesion-characterizing features, ensuring clinically relevant data. We further utilize the generated polyp frames to improve performance in the downstream task of polyp segmentation. Using these generated images, we found an average improvement of **6.84**% and **1.3**% (Jaccard index) on the CVC-ClinicDB and Kvasir-SEG datasets, respectively. The source code is available at* https://github.com/Vanshali/ControlPolypNet.

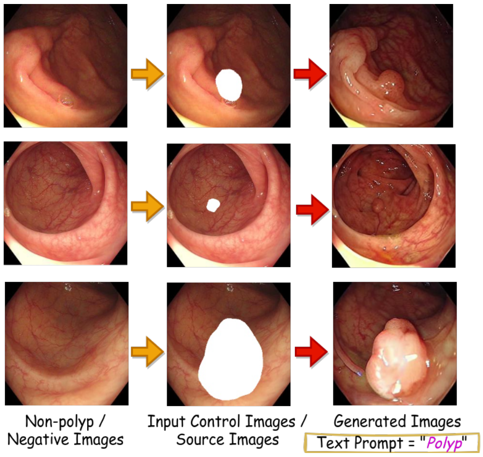

Figure 1. ***Controlling*** polyp generation using ***custom masks*** while leveraging largely accessible non-polyp/negative images. We turned negative samples into positive ones with controlled polyp shape, size and location.

1. Introduction

Colon polyps are abnormal growths with a high risk of developing into the third most common malignancy, colorectal cancer (CRC). Early detection of these polyps is crucial to reduce the associated mortality rate; hence, various screening tests are used in clinical practices, among which colonoscopy is the most widely used medical procedure. It has been reported that colonoscopy can reduce CRC incidence by about 30% [8]. However, despite this fact, the inter-class similarities and intra-class variations among polyps lead to increased miss-rate and make the process largely operator-dependent [14]. To alleviate such issues of misdetection, automated diagnosis systems are incorpo-

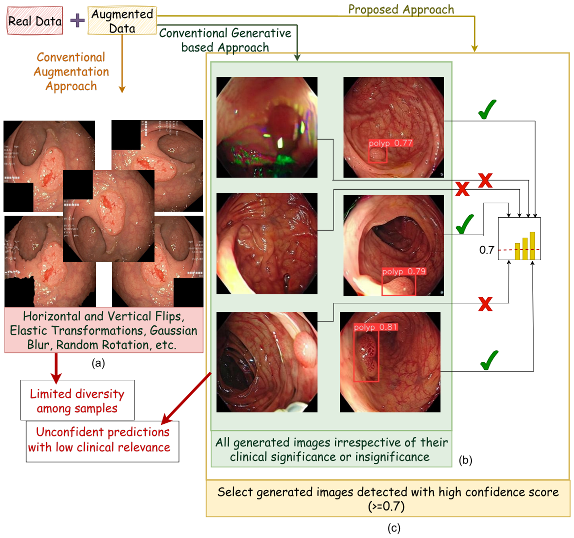
Real Data
Augmented Data
Proposed Approach
Conventional Generative based Approach
Conventional Augmentation Approach
polyp 0.77
polyp 0.79
polyp 0.81
0.7
Horizontal and Vertical Flips, Elastic Transformations, Gaussian Blur, Random Rotation, etc.
(a)
Limited diversity among samples
Unconfident predictions with low clinical relevance
All generated images irrespective of their clinical significance or insignificance
(b)
Select generated images detected with high confidence score (>=0.7)
(c)

Figure 2. Augmentation strategies; (a) Conventional augmentation techniques present limited diversity among samples, (b) Conventional generative approaches use all generated images irrespective of their clinical relevance, and (c) Our approach has an additional detection step that selects generated images which are detected with a high confidence score, ensuring clinical relevance.

rated into the traditional colonoscopy process.

Several colonoscopy analysis tasks, including polyp detection [29], segmentation [14], and classification [19] have shown remarkable improvement using deep learning based systems compared to conventional techniques. However, the performance of these systems on unseen examples significantly relies on the diversity and quantity of examples in the training data. Acquiring a large amount of varied data, especially with associated annotations, is challenging in the medical domain. This procedure involves data privacy concerns, and manual labeling by clinical experts becomes a bottleneck. To overcome this issue, existing automated methods typically use conventional augmentation techniques, such as rotation, vertical and horizontal flips, etc. Simply relying on such techniques restricts the scale-up of the dataset to a certain extent, depending on the dataset size, and limits diversity among samples. This augmentation practice is illustrated in Fig. 2(a).

Considering the above-mentioned scenarios, one possible solution is to expand the training data by incorporating synthetic data. This solution is viable and offers various benefits: *(1) It does not require the time-consuming task of manual labeling. (2) It eliminates the long process of obtaining data privacy informed consent, accelerating dataset development. (3) It provides an opportunity to obtain hard-to-find anomalies that are difficult to observe during routine colonoscopy.* To generate realistic-looking synthetic data, in recent years, generative adversarial networks (GANs) have been widely used in various fields, including medical imaging [9, 28]. Despite the improved performance in the downstream tasks, the issue of convergence instability of GANs and their limited contributions in such tasks resulted in the development of currently trending diffusion models [5, 10]. Diffusion models are expected to generate more realistic images and support text-to-image generation, thus facilitating automated systems with text prompts for better control. These models have been explored in many medical applications, such as image-to-image translation [18], reconstruction [20], image generation [6], segmentation [2], and classification [33], especially using radiology images. However, colonoscopy images have not been explored much and require validations on the diffusion models' ability to learn and generate complex patterns. Besides visually satisfactory image formations, these models must be evaluated on their ability to retain clinically significant information and the usefulness of generated data for downstream tasks such as polyp segmentation. Conventional generative approaches often skip such clinical relevance validations, using all generated images for augmentations, as shown in Fig. 2(b).

In this work, we propose *ControlPolypNet*, based on ControlNet [34] architecture and diffusion concept to generate realistic-looking polyp frames. Our framework has a novel input control map, which converts non-polyp frames with normal mucosa (relatively easy to obtain) to polyp frames (hard to obtain). This process is summarized in Fig. 1. Additionally, we employ a detector module in *ControlPolypNet* that discards frames that do not carry lesion-characterizing features. This step prioritizes polyp-specific characteristics, emphasizing them before proceeding with augmentation, as shown in Fig. 2(c). Also, we evaluated the impact of the generated data on the downstream task of polyp segmentation. Our method offers a more practical approach to data augmentation, which is expected to represent clinically relevant data with diverse characteristics. Using the generated data, polyp segmentation shows an average improvement of **6.84**% and **1.3**% Jaccard index on CVC-ClincDB and Kvasir-SEG, respectively. The contributions of our work can be summarised as follows:

- **Framework with novel user-configurable input control map:** We propose an approach using novel user-configurable input control to generate polyps while leveraging the largely accessible non-polyp frames. This control map can control the endoluminal objects and polyp generation (in terms of shape, size and location) using customized masks and non-polyp frames.
- **Additional examination to avoid irrelevant synthetic information:** We employ a detector module that verifies the quality of generated polyps and selects clinically appropriate synthetic polyps that carry lesion-characterizing features. The detector eliminates the risk of adding noise

and irrelevant information to the generated data.

- **Improved polyp segmentation performance:** We report enhanced polyp segmentation performance by augmenting two publicly available datasets using our synthetic images. This has been achieved without additional expensive manual annotation requirements.

2. Related Work

In recent years, various generative artificial intelligence strategies have been adopted in the colonoscopy domain. However, diffusion based models are still rarely explored in the domain. One such work is reported in [16], where Machacek et al. adopted a latent diffusion model to generate synthetic polyps using segmentation masks. Besides diffusion models, various GAN-based architectures have been employed to generate colon polyps to circumvent the issue of limited labeled data. For example, Shin et al. [30] utilized conditional adversarial networks and combined edge map with polyp binary mask as a conditioned input image. Further, they introduced dilated convolutions in the generator encoder. Sasmal et al. [26] adopted DCGAN to augment the polyp dataset, boosting the classifier's ability to differentiate between hyperplastic and adenomatous polyps. To expand the data distribution, He et al. [9] integrated a generator, a detector, and an attacker to produce false negative samples. They reported that instead of merely using a generator to produce synthetic images, using adversarial samples can enhance the performance of a re-trained detector.

Unlike the above methods that altogether generated a new image, Qadir et al. [21] employed a simple conditional GAN framework to convert polyp images into negative images initially and then used a controllable binary mask to convert them into polyp images. Sams and Shomee [25] produced random masks using StyleGAN2-ada and integrated them with normal colon images to obtain a composite image. This composite image then acts as the input for the conditional GAN. Thomaz et al. [4] adopted an approach to copy polyp regions to non-polyp images. Additionally, they used a conditional GAN to synthesize new polyps. The generated images are transferred to the target image using an algorithm based on image processing techniques. Such existing GAN-based techniques suffer from convergence issues and produce less realistic images with limited diversity. Also note that unlike Machacek et al. [16], our work focuses on generating more realistic images by controlling the background details, reducing the possibility of uninformative content and artifacts.

3. Proposed Method

3.1. Overview

The objective of the proposed method is to generate polyp frames to increase the sample count for training and enhancing deep learning models' performance. Given two subsets of images, polyp/positive (P) and non-polyp/negative (N), our goal is to utilize images in N to expand the subset P. This is achieved by transforming images $N = \{n_1, n_2, ..., n_s\}$ into $P' = \{p'_i | p'_i$ is similar in distribution to $p_j\}$, where $p_j \in P$. Moreover, during this transformation, polyp shape, location, and size are user-configurable and integrating P' with P should diversify the overall set. This signifies that the synthetic polyp set P' should be diverse and possess qualities similar to real images in set P.

3.2. Preliminaries

Stable Diffusion Models (SD): SD is a text-to-image model built upon the basic functionality of latent diffusion models (LDM) [23]. In this model, an encoder E encodes a given image $a \in \mathbb{R}^{H\times W\times 3}$ into a latent representation a_l. Like DPM [31], it gradually introduces Gaussian noise in the image but is done in the latent space on a_l, resulting in a noisy image a_{l_t} at time step t. Subsequently, a U-Net with ResNet blocks, incorporating the time step t, is employed for denoising. $\epsilon_\theta(a_{lt}, t)$ acts as a sequence of denoising autoencoders to predict the denoised version. The final output representation is reconstructed into a'_l using a decoder D. The corresponding objective is given below:

$$L_{LDM} := \mathbb{E}_{E(a),\epsilon,t}[||\epsilon - \epsilon_\theta(a_{lt}, t)||_2^2] \quad (1)$$

where $\epsilon \sim \mathcal{N}(0,1)$. The SD further utilizes a text encoder, which is a pre-trained CLIP [22]. It allows encoding the text prompts into embeddings. These text embeddings are then fused with the encoder and decoder of U-Net using cross-attention layers. This cross-attention mechanism helps condition the model using a text prompt b after processing it through an encoder $\mathcal{Z}$. The objective can be defined as:

$$L_{LDM_b} := \mathbb{E}_{E(a),b,\epsilon,t}[||\epsilon - \epsilon_\theta(a_{lt}, t, \mathcal{Z}_\theta(b))||_2^2] \quad (2)$$

ControlNet: ControlNet is designed to control the diffusion models to enable them to perform a specific downstream task. It uses an input control map that provides an opportunity to manipulate the generated output. ControlNet, in its standard form, supports control maps with different conditions, such as edge maps, scribbles, segmentation maps, pose, etc. It preserves the weights of the SD by making a locked copy of it. Simultaneously, it uses a trainable copy with task-specific conditional control for a downstream task. These two copies are connected via 1×1 zero convolution layers with both bias and weight initialized as zero. The convolutional weights of these layers gradually optimize, which gives the benefit of no extra added noise with faster training at the same time. Let locked and trainable copy parameters be denoted as α and α_c, respectively. If zero convolution operation is $\mathcal{C}(.;.)$ which uses two instances of parameters $\{\alpha_{c1}, \alpha_{c2}\}$, then combining it into the

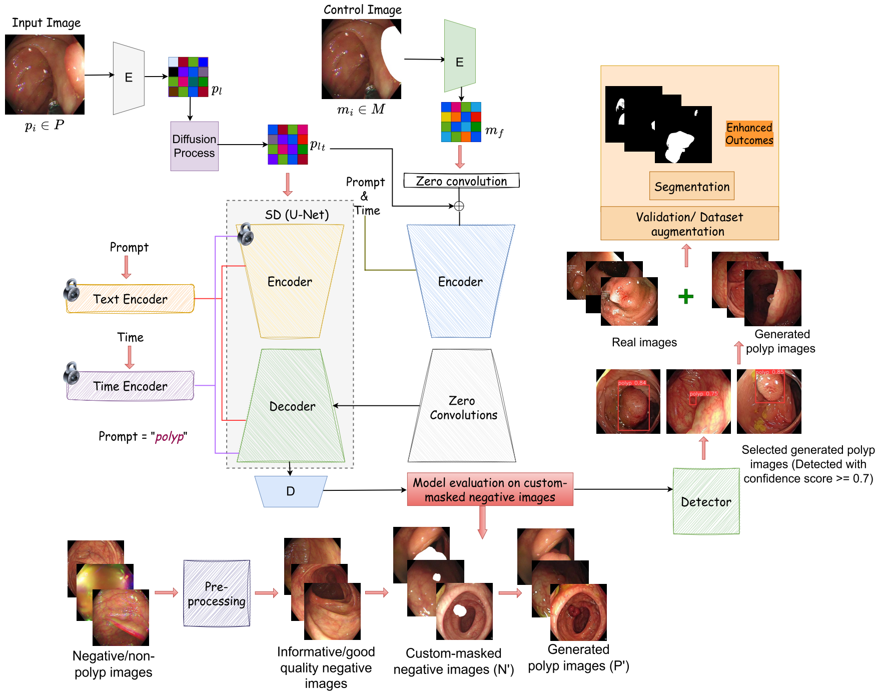

Figure 3. The proposed framework uses custom-masked images as input control with a "polyp" text prompt. The pre-processing pipeline shows the elimination of uninformative negative frames. Custom masks are used to generate polyps during the evaluation phase of *ControlPolypNet*.

ControlNet network blocks $\mathcal{H}(.;.)$ could be represented as:

$$y_c = \mathcal{H}(x, \alpha) + \mathcal{C}(\mathcal{H}(x + \mathcal{C}(c, \alpha_{c1}); \alpha_c); \alpha_{c1}) \quad (3)$$

where y_c is the output and x is the input feature map. The overall objective after including the downstream task can be modified as shown below:

$$L_{CN} := \mathbb{E}_{E(a),b,b'_f,\epsilon,t}[\|\epsilon - \epsilon_\theta(a_{lt}, t, \mathcal{Z}_\theta(b), b'_f)\|_2^2] \quad (4)$$

where b'_f is the intermediate representation of the task-specific condition.

3.3. ControlPolypNet

The architectural details of the proposed approach are shown in Fig. 3. *ControlPolypNet* consists of three main parts: (a) an SD U-Net architecture loaded with pre-trained weights of SD v1-5, (b) ControlNet, and (c) YOLOv8 [32], a detector pre-trained on the polyp images. The decoder part of the SD U-Net is kept unlocked, and only its encoder part is left locked during the complete training process. This unlocking is done to obtain better performance on medical imaging tasks like ours, as the initial weights are more inclined toward general images. Instead of adopting standard control map options presented by ControlNet, we tailored the input condition map to fit the necessary requirements.

We utilized the negative colonoscopy frames N, which are relatively easily accessible in sufficiently large amounts. We overlapped these frames with random custom masks to obtain N', which are the regions targeted for polyp generation to obtain P'. To make the model learn the mapping $N' \rightarrow P'$, we prepared our training set such that initially, it learns $M \rightarrow P$, where M is obtained by overlapping P with its binary mask ground truth. By providing P as the target image and M as the source image (control image), the model learns the mapping $M \rightarrow P$. While learning this mapping, the model learns the complex patterns in data, and when given a random mask over non-polyp image n'_i, it transforms it into p'_i when given the text prompt "polyp". This mapping allows the usage of custom masks with controllable positions and shapes of polyps. Also, this reduces the probability of obtaining unwanted structures or noise in the background/endoluminal scene.

When given a polyp image p_i, the standard diffusion process progressively adds noise to the image in its latent representation p_l to obtain a noisy version p_{lt}. This input is combined with conditions in the form of mask-overlapped image $m_i \in M$ and text prompt b, i.e., "polyp". m_i is further converted into an intermediate representation m_f by performing encoding on m_i to match the input size of SD. The objective of *ControlPolypNet* can be defined as:

$$L_{CPN} := \mathbb{E}_{E(p_i),b,m_f,\epsilon,t}[\|\epsilon - \epsilon_\theta(p_{lt}, t, \mathcal{Z}_\theta(b), m_f)\|_2^2] \quad (5)$$

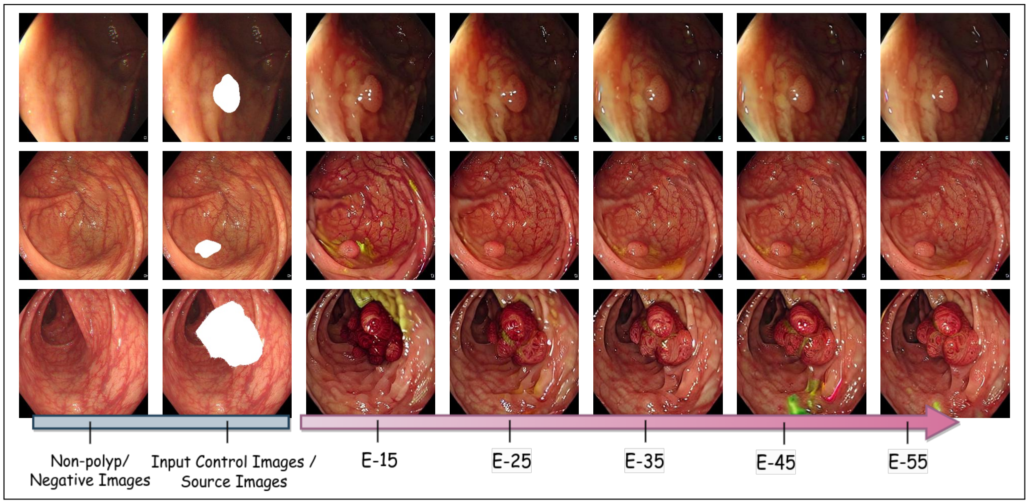

Figure 4. Epoch-wise sample images along with their corresponding negative images and input control images (custom-masked negative samples). E stands for epochs.

The proposed input control ensures that the other endoluminal scene remains intact, which could be beneficial to capturing and differentiating polyp regions during downstream tasks. As stated in [19], considering some regions from background aids in improving classification results. This outcome could be attributed to polyps exhibiting a distinct color and texture, setting them apart from the normal mucosal regions. Unwanted noise and irrelevant objects in the generated outputs' background create unrealistic data, deviating a model from the intended task. Therefore, we utilized the negative frames instead of relying on the standard binary masks. However, these negative frames can have some artifacts, as colonoscopy videos are prone to motion blur, interlacing, ghost colors, etc. Hence, we used an approach given by Sharma et al. [27] to filter out uninformative negative frames before their use in the translation.

Pathological Validation Setup: Although generative models are now common in the medical imaging domain, various studies [3, 7] show that they are liable to generate unrealistic medical conditions or structures. As pathological patterns are significantly crucial, we performed an elimination step instead of directly integrating them into the segmentation task training. This elimination step validates the presence of lesion-characterizing features in the synthetic images and simultaneously prepares a clinically valid set of images appropriate for data augmentation. We integrated a polyp detector, YOLOv8, in the proposed framework for this process. This detector is pre-trained on polyp images with a confidence score set in the range of 0.7 and 0.8 for inference. This integration helps choose only valid, visually appealing frames with lesion-characterizing features. We used these selected synthetic polyp frames to augment the training set for the segmentation task.

4. Experiments and Results

4.1. Dataset Details and Training Settings

We used three publicly available datasets, namely, SUN Database [17] (49,136 polyp frames and 109,554 non-polyp frames), CVC-ClinicDB [1] (612 polyp images) and Kvasir-SEG [12] (1000 polyp images), to validate the performance of our framework. The segmentation ground truth of the SUN Database, released in the form of SUN-SEG [15], is also used. The SUN Database and SUN-SEG are used in the training of *ControlPolypNet*, whereas CVC-ClinicDB and Kvasir-SEG are used to validate generated image quality in the downstream task of polyp segmentation.

During *ControlPolypNet* training, we used 38,284 polyp images; the rest were used for validation. To translate non-polyp images into polyp images, we custom-masked 10,000 negative images after pre-processing non-polyp video sequence cases with the informative/uninformative frame detector given by Sharma et al. [27]. The official split of CVC-ClinicDB and Kvasir-SEG is used. The implementation is done using PyTorch and PyTorch lightning frameworks. *ControlPolypNet* and downstream task training are executed using NVIDIA A100 and NVIDIA Titan-Xp GPU, respectively. *ControlPolypNet* is trained for 55 epochs with a batch size of 32 and a learning rate of 2e^{-6}.

4.2. Evaluation Metrics

The quality of the generated images is accessed using three metrics: Frechet inception distance (FID), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM). FID quantifies the quality of synthetic data for realism and diversity. PSNR is focused on the reconstruction quality of images, and SSIM quantifies the similarity be-

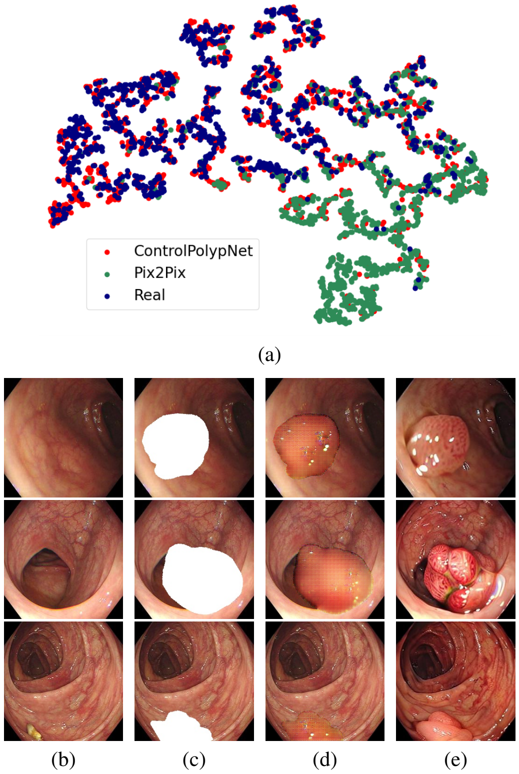

Figure 5. (a) Two-dimensional t-SNE embedding pertaining to real polyp images, and images generated by Pix2Pix and *ControlPolypNet*, (b)-(e) show negative images, masked negative images, synthetic images obtained using Pix2Pix and *ControlPolypNet*, respectively.

Table 1. Quantitative comparison of synthetic polyp images with different sets of real images over different epochs. **Bold** values represent the 'best' metrics score, and E, P, NP stand for 'epoch', 'polyp', and 'non-polyp', respectively. ↓ and ↑ denote 'lower is best' and 'higher is best', respectively.

Metrics	Trend	Comparsion (with)	E-15	E-25	E-35	E-45	E-55
FID	↓	Real P images	104.52	106.70	102.46	99.35	**94.07**
		Real NP images	92.10	93.77	91.16	89.22	**82.95**
PSNR	↑	Masked NP images	67.70	67.22	67.66	67.57	**68.39**
SSIM	↑	Masked NP images	0.9987	0.9984	0.9986	0.9986	**0.9988**

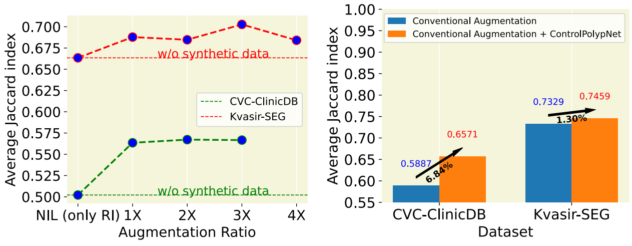

Figure 6. Analyzing the average Jaccard index across three segmentation models in various scenarios. (a) Impact of different data augmentation ratios. (b) Comparing the average Jaccard index using conventional augmentations with and without images generated by our approach.

tween two images. Additionally, we used task-specific segmentation metrics, including precision, recall, F1-score and Jaccard index. The Jaccard index determines the overlap between the ground truth and prediction masks.

4.3. Performance Evaluation

We evaluated our model on different epochs and examined the quality of the generated images using the quality assessment metrics (see Table 1). While using FID, we considered two scenarios: synthetic vs. real polyp images and synthetic vs. real non-polyp images. As expected, the latter case presented a better score because the related non-polyp images are translated into synthetic polyp with background details substantially preserved. It can be observed that the quality of images in both cases gradually increased with the epoch counts. Due to the high computational requirements of diffusion models, we considered training till the point where visually appealing results were obtained. We further explored the structure and information-preserving ability of our approach using PSNR and SSIM. We masked the generated images' polyp region and compared them with the masked non-polyp images. The results show that the quality of the endoluminal scene is satisfactorily preserved and is improved with the increasing epochs.

Besides quantitative outcomes, we observed the qualitative results, shown in Fig. 4. In the initial epochs, especially in epoch 15, the image details are not precisely generated and are obstructed by artifacts. Moreover, the color transfer ability from the input control image to synthetic images is higher in the later epochs. The randomness in polyp color and close mapping of the polyp shape and its location with the custom mask demonstrates our approach's potential to achieve data diversity and successful control over synthetic polyp shape, size and location. Although the results demonstrate the scope of improvement in color-preservation ability, structural-preservation outcomes are impressive. Further, we compared the outcomes of *ControlPolypNet* with that of Pix2Pix [11]. We selected Pix2Pix because it uses a mechanism to translate images from one domain to another, suitable for our objective to translate $N' \rightarrow P'$. A qualitative comparison is shown in Fig. 5 where the images in Fig. 5(b)-(e) clearly show that although both *ControlPolypNet* and Pix2Pix retained the polyp location and shape, more realistic polyp images with texture were generated by the former. However, compared to our model, Pix2Pix was better at retaining the original colors of background regions. Additionally, we generated a t-SNE plot (shown in Fig. 5(a)) using a DenseNet-201 that is trained to differentiate polyp and non-polyp images [27]. Feature embedding plots of real

Table 2. Performance of the U-Net [24], ColonSegNet [13], and TransNetR[14] models on the downstream task of polyp segmentation. RI stands for Real Images. The best results are highlighted in **bold** and the second best are underlined.

Dataset: CVC-ClinicDB

Training sample count (X = 490)	**U-Net**				**ColonSegNet**				**TransNetR**			
	Jaccard	**Recall**	**Precision**	**F1-score**	**Jaccard**	**Recall**	**Precision**	**F1-score**	**Jaccard**	**Recall**	**Precision**	**F1-score**
RI (X)	0.4682	0.5211	0.8509	0.5523	0.3429	0.3834	0.8256	0.4424	0.6952	0.7431	0.9399	0.7737
RI + Random Rotation (X+X)	0.4748	0.5244	0.8909	0.5568	0.4352	0.4859	0.8161	0.5312	0.7015	0.7450	0.9468	0.7805
RI + Gaussian Blur (X+X)	0.4447	0.4809	0.8705	0.5215	0.3467	0.3779	0.8291	0.4453	0.6960	0.7433	0.9357	0.7762
RI + Vertical Flip (X+X)	0.4589	0.5027	**0.9218**	0.5354	0.3666	0.3976	0.8412	0.4585	0.6675	0.7094	0.9283	0.7442
RI + Horizontal Flip (X+X)	0.4348	0.5138	0.8447	0.5198	0.4296	0.4696	0.8991	0.5080	0.6991	0.7581	0.9279	0.7823
RI + Elastic Transformation (X+X)	0.4296	0.4696	0.8991	0.5080	0.3867	0.4275	0.8019	0.4874	0.5907	0.6197	0.9439	0.6691
RI + Pix2Pix Synthetic Images (X+X)	0.4493	0.4964	0.7917	0.5474	0.3872	0.4019	0.8343	0.4661	0.7076	0.7406	0.9469	0.7872
RI + ControlPolypNet Synthetic Images (X+X)	0.5356	0.5781	0.9096	0.6232	0.4360	0.4831	0.8211	0.5359	0.7191	0.7731	0.9366	0.7967
RI + Pix2Pix Synthetic Images (X+2X)	0.3363	0.4323	0.6736	0.4429	0.4196	0.4465	0.7680	0.5065	0.6953	0.7299	**0.9570**	0.7719
RI + ControlPolypNet Synthetic Images (X+2X)	0.5424	0.6390	0.8292	0.6365	0.4272	0.4828	0.7782	0.5267	0.7322	0.7837	0.9366	0.8113
RI + Pix2Pix Synthetic Images (X+3X)	0.4763	0.4975	0.8752	0.5570	0.4283	0.4531	0.8683	0.5192	0.6875	0.7174	0.9571	0.7599
RI + ControlPolypNet Synthetic Images (X+3X)	0.5375	0.5802	0.8660	0.6149	0.4726	0.5432	0.8093	0.5760	0.6900	0.7287	0.9505	0.7628
RI + 5 aug. (X+5X)	0.5518	0.6252	0.9002	0.6353	0.4928	0.5307	0.8623	0.5855	0.7214	0.7639	0.9426	0.7963
RI + 5 aug. + ControlPolypNet Synthetic Images (X+5X+2X)	**0.6298**	**0.7132**	0.8900	**0.7160**	**0.5928**	**0.6308**	**0.9167**	**0.6874**	**0.7486**	**0.7968**	0.9365	**0.8198**

Dataset: Kvasir-SEG

Training sample count (X = 880)	**U-Net**				**ColonSegNet**				**TransNetR**			
	Jaccard	**Recall**	**Precision**	**F1-score**	**Jaccard**	**Recall**	**Precision**	**F1-score**	**Jaccard**	**Recall**	**Precision**	**F1-score**
RI (X)	0.6668	0.7796	0.8420	0.7508	0.5782	0.7148	0.7610	0.6869	0.7454	0.8273	0.9058	0.8267
RI + Random Rotation (X+X)	0.6852	0.7679	**0.8702**	0.7669	0.6143	0.7280	0.8045	0.7148	0.7469	0.8289	0.9005	0.8298
RI + Gaussian Blur (X+X)	0.6704	0.7736	0.8521	0.7563	0.5677	0.7116	0.7705	0.6793	0.7596	0.8426	0.8956	0.8399
RI + Vertical Flip (X+X)	0.6738	0.7693	0.8614	0.7580	0.6129	0.7504	0.7965	0.7184	0.7749	0.8552	0.8946	0.8501
RI + Horizontal Flip (X+X)	0.6837	0.7984	0.8390	0.7743	0.6039	0.7202	0.8105	0.7115	0.7629	0.8357	0.9120	0.8370
RI + Elastic Transformation (X+X)	0.6667	0.7996	0.8239	0.7538	0.6163	0.7399	0.8088	0.7208	0.7369	0.8265	0.8806	0.8160
RI + Pix2Pix Synthetic Images (X+X)	0.6550	0.7516	0.8353	0.7357	0.5757	0.6976	0.7920	0.6824	0.7659	0.8482	0.9020	0.8425
RI + ControlPolypNet Synthetic Images (X+X)	0.6795	0.8032	0.8498	0.7688	0.6262	0.7532	0.8098	0.7345	0.7579	0.8497	0.8801	0.8373
RI + Pix2Pix Synthetic Images (X+2X)	0.6127	0.7258	0.8103	0.7060	0.5820	0.7123	0.7783	0.6887	0.7651	0.8539	0.8984	0.8439
RI + ControlPolypNet Synthetic Images (X+2X)	0.6680	**0.8465**	0.7971	0.7640	0.6065	0.7508	0.7913	0.7209	0.7797	0.8665	0.9010	0.8523
RI + Pix2Pix Synthetic Images (X+3X)	0.6580	0.7624	0.8440	0.7441	0.6048	0.7353	0.7916	0.7113	0.7747	0.8524	0.9109	0.8497
RI + ControlPolypNet Synthetic Images (X+3X)	0.6997	0.8331	0.8464	0.7879	0.6326	0.7603	0.8121	0.7379	0.7760	**0.8677**	0.8938	0.8517
RI + Pix2Pix Synthetic Images (X+4X)	0.6720	0.7665	0.8633	0.7564	0.6021	0.7231	0.7961	0.6986	0.7346	0.8550	0.8441	0.8208
RI + ControlPolypNet Synthetic Images (X+4X)	0.6750	0.8126	0.8339	0.7651	0.6341	0.7835	0.7967	0.7440	0.7432	0.8139	0.9039	0.8245
RI + 5 aug. (X+5X)	0.7069	0.8131	0.8465	0.7912	0.6958	0.8086	0.8515	0.7907	**0.7960**	0.8518	**0.9366**	**0.8641**
RI + 5 aug. + ControlPolypNet Synthetic Images (X+5X+3X)	**0.7301**	0.8368	0.8657	**0.8153**	**0.7215**	**0.8191**	**0.8638**	**0.8129**	0.7861	0.8622	0.9024	0.8584

and synthetic polyp images clearly depict the closeness of our model's outcomes with real images. Contrarily, the images generated by Pix2Pix barely overlap with the real data.

4.3.1 Clinical Significance Validation and Downstream Task Evaluation

The clinical significance validation step employs a detector, as discussed in Section 3.3. The synthetic images that YOLOv8 detected with confidence scores in the range of 0.7 and 0.8 are used to augment the dataset of the downstream task. This augmentation approach provides two-fold benefits: a) It validates the synthetic images' quality and clinical significance, and b) It allows enhancing segmentation outcomes. We experimented with different proportions of synthetic images and five general augmentations: random rotation, Gaussian blur, elastic transformation and horizontal and vertical flips. We used three state-of-the-art polyp segmentation models, U-Net [24], ColonSegNet [13], and TransNetR [14] to experiment with different augmentation combinations. The associated results are shown in Table 2.

During augmentation, we increased the ratio of synthetic images as a multiple of X, where X is the original training set size. It can be observed that adding synthetic images in X proportion performs comparable to adding single conventional augmentation. We gradually increased synthetic images in iX proportion, where i={1,2,3,4}. The results show that the polyp segmentation performance achieves a significant increase with small ratios, and then, with increasing ratios, the improvement is either minimal or absent. The same can be inferred from Fig. 6(a). The value of i is incremented until the metrics values start to decrease. The proportion iX that performs the best is combined further with conventional augmentations. The outcomes from this integration show that synthetic images complement conventional augmentation techniques as the average performance increased compared to cases where only conventional augmentations were used. Additionally, we compared *ControlPolypNet* with Pix2Pix using the same proportion of their generated data for augmentation. An average Jaccard index over all the different proportions (X, 2X, 3X or 4X) is 5.61% and 2.3% higher using *ControlPolypNet* compared to Pix2Pix on CVC-ClinicDB and Kvasir-SEG, respectively. This increase can be observed in Fig. 6(b). Moreover, the individual performance with different data proportions and models has reported enhanced performance using our augmentation

Table 3. Quality assessment of images generated using Pix2Pix and *ControlPolypNet*. This assessment is conducted using U-Net [24], ColonSegNet [13], and TransNetR[14] models trained on real images. The best results are highlighted in **bold**.

Training Dataset	Generation Method	U-Net				ColonSegNet				TransNetR			
		Jaccard	Recall	Precision	F1-score	Jaccard	Recall	Precision	F1-score	Jaccard	Recall	Precision	F1-score
CVC-ClinicDB	Pix2Pix	0.2048	0.6943	0.2562	0.3054	0.2323	0.7129	0.2934	0.3414	0.4517	0.6510	**0.6219**	0.5662
	ControlPolypNet	**0.2613**	**0.7792**	**0.3088**	**0.3802**	**0.2633**	**0.7353**	**0.3328**	**0.3876**	**0.4761**	**0.7729**	0.5991	**0.6149**
Kvasir-SEG	Pix2Pix	0.5802	0.6994	**0.7450**	0.6597	**0.4778**	0.7508	**0.5814**	**0.5814**	0.6037	0.6414	**0.9109**	0.6657
	ControlPolypNet	**0.6285**	**0.8128**	0.7394	**0.7362**	0.4039	**0.7973**	0.4842	0.5354	**0.7580**	**0.8537**	0.8749	**0.8454**

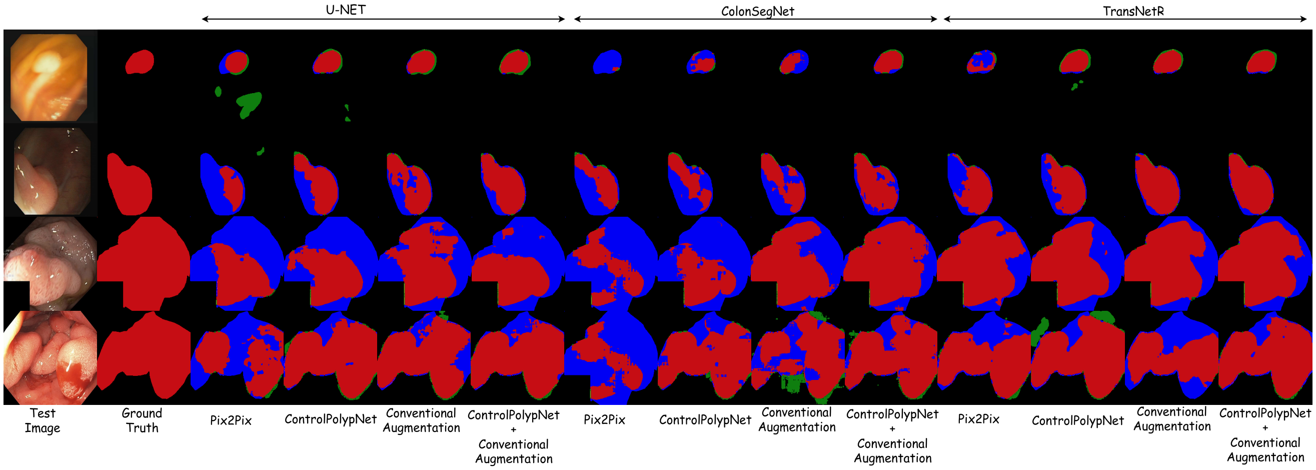
U-NET
ColonSegNet
TransNetR
Test Image
Ground Truth
Pix2Pix
ControlPolypNet
Conventional Augmentation
ControlPolypNet + Conventional Augmentation
Pix2Pix
ControlPolypNet
Conventional Augmentation
ControlPolypNet + Conventional Augmentation
Pix2Pix
ControlPolypNet
Conventional Augmentation
ControlPolypNet + Conventional Augmentation

Figure 7. Qualitative results of polyp segmentation outcomes. The figure illustrates that in most cases, when *ControlPolypNet*'s output is combined with conventional augmentation techniques, it predicts a mask closer to ground truth. Also, the masks obtained using *ControlPolypNet*'s generated images are more appropriate than those obtained using Pix2Pix's generated images.

approach. It is noteworthy that even though the synthetic images are generated using a different larger dataset, they are performing effectively on a small out-of-distribution dataset. This observation supports both quality and diverse information possessed by the generated images. Adopting traditional augmentation techniques is limited by the actual size of the dataset as they can only be scaled up by its multiple. Also, this scaling up produces redundant information in some form. Contrarily, adding our diverse set of synthetic images can complement this information and is independent of real dataset size, thus providing enhanced segmentation outcomes. These results are supported by some qualitative outcomes, shown in Fig. 7. It can be observed that, in most cases, combining conventional techniques with *ControlPolypNet*'s synthetic data provides results closer to the ground truth. We further tested the synthetic images obtained using *ControlPolypNet* and Pix2Pix using the three segmentation models (trained using only real data). The results shown in Table 3 signify that our approach generates more realistic images with polyp-specific characteristics.

Although our proposed approach provides an opportunity to obtain customized polyp images using negative images, some lingering gaps still need to be addressed as control over colors remains unexplored. In medical images, color is one of the criteria considered for domain shift issues, as color variations across inter-hospital and inter-patient data bring performance drops. Control over colonoscopy image color can expand the possibility of domain transfer and even enhance segmentation outcomes.

5. Conclusion

In this work, we proposed a stable diffusion based framework, *ControlPolypNet*, to generate polyp frames utilizing non-polyp frames. We showed that the polyp generation process can be customized, and a user-configurable control can be used to get more fine-grained data. The generated frames also capture pathological features with visually impressive results and help enhance the downstream task of polyp segmentation. A detector, introduced in our proposed framework, ensures the retention of pathological features. Our approach achieved an average increase of 6.84% and 1.3% (Jaccard index) over three models on the CVC-ClinicDB and Kvasir-SEG datasets, respectively.

Acknowledgments

V. Sharma is supported by the INSPIRE fellowship scheme (IF190362), DST, Govt. of India. D. Jha is supported by the NIH funding: R01-CA246704, R01-CA240639, U01 DK127384-02S1, and U01-CA268808.

References

[1] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 5

[2] Afshin Bozorgpour, Yousef Sadegheih, Amirhossein Kazerouni, Reza Azad, and Dorit Merhof. Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 146–158. Springer, 2023. 2

[3] Maria JM Chuquicusma, Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In *Proceedings of the 15th international symposium on biomedical imaging (ISBI 2018)*, pages 240–244, 2018. 5

[4] Victor de Almeida Thomaz, Cesar A Sierra-Franco, and Alberto B Raposo. Training data enhancements for improving colonic polyp detection using deep convolutional neural networks. *Artificial Intelligence in Medicine*, 111:101988, 2021. 3

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

[6] Zolnamar Dorjsembe, Sodtavilan Odonchimed, and Furen Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical Imaging with Deep Learning*, 2022. 2

[7] August DuMont Schütte, Jürgen Hetzel, Sergios Gatidis, Tobias Hepp, Benedikt Dietz, Stefan Bauer, and Patrick Schwab. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ digital medicine*, 4(1):141, 2021. 5

[8] Fatima A Haggar and Robin P Boushey. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, 22(04):191–197, 2009. 1

[9] Fan He, Sizhe Chen, Shuaiyi Li, Lu Zhou, Haiqin Zhang, Haixia Peng, and Xiaolin Huang. Colonoscopic image synthesis for polyp detector enhancement via gan and adversarial training. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1887–1891, 2021. 2, 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6

[12] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM*, pages 451–462, 2020. 5

[13] Debesh Jha, Sharib Ali, Nikhil Kumar Tomar, Håvard D Johansen, Dag Johansen, Jens Rittscher, Michael A Riegler, and Pål Halvorsen. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *Ieee Access*, 9:40496–40510, 2021. 7, 8

[14] Debesh Jha, Nikhil Kumar Tomar, Vanshali Sharma, and Ulas Bagci. TransNetR: Transformer-based Residual Network for Polyp Segmentation with Multi-Center Out-of-Distribution Testing. *MIDL*, 2023. 1, 2, 7, 8

[15] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022. 5

[16] Roman Macháček, Leila Mozaffari, Zahra Sepasdar, Sravanthi Parasa, Pål Halvorsen, Michael A. Riegler, and Vajira Thambawita. Mask-conditioned latent diffusion for generating gastrointestinal polyp images. In *Proceedings of the 4th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, page 1–9. Association for Computing Machinery, 2023. 3

[17] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Kinichi Hotta, Kazuo Ohtsuka, Takahisa Matsuda, Shoichi Saito, Toyoki Kudo, Toshiyuki Baba, Fumio Ishida, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal endoscopy*, 93(4):960–967, 2021. 5

[18] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023. 2

[19] Krushi Patel, Kaidong Li, Ke Tao, Quan Wang, Ajay Bansal, Amit Rastogi, and Guanghui Wang. A comparative study on polyp classification using convolutional neural networks. *PloS one*, 15(7):e0236452, 2020. 2, 5

[20] Cheng Peng, Pengfei Guo, S Kevin Zhou, Vishal M Patel, and Rama Chellappa. Towards performant and reliable undersampled MR reconstruction via diffusion model sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–633. Springer, 2022. 2

[21] Hemin Ali Qadir, Ilangko Balasingham, and Younghak Shin. Simple U-net based synthetic polyp image generation: Polyp to negative and negative to polyp. *Biomedical Signal Processing and Control*, 74:103491, 2022. 3

[22] 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 *Proceedings of the International conference on machine learning*, pages 8748–8763, 2021. 3

[23] 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

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmen-

tation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241, 2015. 7, 8

[25] Ataher Sams and Homaira Huda Shomee. GAN-based realistic gastrointestinal polyp image synthesis. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2022. 3

[26] Pradipta Sasmal, MK Bhuyan, Sourav Sonowal, Yuji Iwahori, and Kunio Kasugai. Improved endoscopic polyp classification using GAN generated synthetic data augmentation. In *2020 IEEE Applied Signal Processing Conference (ASPCON)*, pages 247–251, 2020. 3

[27] Vanshali Sharma, Pradipta Sasmal, MK Bhuyan, and Pradip K Das. Keyframe selection from colonoscopy videos to enhance visualization for polyp detection. In *2022 26th International Conference Information Visualisation (IV)*, pages 426–431, 2022. 5, 6

[28] Vanshali Sharma, MK Bhuyan, and Pradip K Das. Can adversarial networks make uninformative colonoscopy video frames clinically informative? (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16322–16323, 2023. 2

[29] Vanshali Sharma, Pradipta Sasmal, MK Bhuyan, Pradip K Das, Yuji Iwahori, and Kunio Kasugai. A multi-scale attention framework for automated polyp localization and keyframe extraction from colonoscopy videos. *IEEE Transactions on Automation Science and Engineering*, 2023. 2

[30] Younghak Shin, Hemin Ali Qadir, and Ilangko Balasingham. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access*, 6:56007–56017, 2018. 3

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

[32] Ultralytics. Yolov8. https://github.com/ultralytics/ultralytics/, 2023. Accessed: 01 Aug 2023. 4

[33] Yijun Yang, Huazhu Fu, Angelica I Aviles-Rivero, Carola-Bibiane Schönlieb, and Lei Zhu. Diffmic: Dual-guidance diffusion network for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 95–105. Springer, 2023. 2

[34] 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