ArSDM: Colonoscopy Images Synthesis with Adaptive Refinement Semantic Diffusion Models

Yuhao Du[1,2][†], Yuncheng Jiang[1,2,3][†], Shuangyi Tan[1,2], Xusheng Wu[6],

Qi Dou[5], Zhen Li[2,3][⋆], Guanbin Li[4][⋆], and Xiang Wan[1,2]

[1] Shenzhen Research Institute of Big Data, Shenzhen, China
[2] SSE, CUHK-Shenzhen, Shenzhen, China
[3] FNii, CUHK-Shenzhen, Shenzhen, China
[4] School of Computer Science and Engineering, Research Institute of Sun Yat-sen University in Shenzhen, Sun Yat-sen University, Guangzhou, China
[5] The Chinese University of Hong Kong, Hong Kong, China
[6] Shenzhen Health Development Research and Data Management Center, China
`lizhen@cuhk.edu.cn`
`liguanbin@mail.sysu.edu.cn`

Abstract. Colonoscopy analysis, particularly automatic polyp segmentation and detection, is essential for assisting clinical diagnosis and treatment. However, as medical image annotation is labour- and resource-intensive, the scarcity of annotated data limits the effectiveness and generalization of existing methods. Although recent research has focused on data generation and augmentation to address this issue, the quality of the generated data remains a challenge, which limits the contribution to the performance of subsequent tasks. Inspired by the superiority of diffusion models in fitting data distributions and generating high-quality data, in this paper, we propose an **A**daptive **R**efinement **S**emantic **D**iffusion **M**odel (**ArSDM**) to generate colonoscopy images that benefit the downstream tasks. Specifically, ArSDM utilizes the ground-truth segmentation mask as a prior condition during training and adjusts the diffusion loss for each input according to the polyp/background size ratio. Furthermore, ArSDM incorporates a pre-trained segmentation model to refine the training process by reducing the difference between the ground-truth mask and the prediction mask. Extensive experiments on segmentation and detection tasks demonstrate the generated data by ArSDM could significantly boost the performance of baseline methods.

Keywords: Diffusion models · Colonoscopy · Polyp segmentation · Polyp detection.

1 Introduction

Colonoscopy is a critical tool for identifying adenomatous polyps and reducing rectal cancer mortality. Deep learning methods have shown powerful abilities

---

[†] Equal contributions.

[⋆] Corresponding authors.

Fig. 1. Overview of the pipeline of our proposed approach, where details of **ArSDM** are described in section 2.

in automatic colonoscopy analysis, including polyp segmentation [5,22,26,27,29] and polyp detection [19,24]. However, the scarcity of annotated data due to high manual annotation costs results in poorly trained and low generalizable models. Previous methods have relied on generative adversarial networks (GANs) [9,25] or data augmentation methods [3,13,28] to enhance learning features, but these methods yielded limited improvements in downstream tasks. Recently, diffusion models [6,15] have emerged as promising solutions to this problem, demonstrating remarkable progress in generating multiple modalities of medical data [4,10,12,21].

Despite recent progress in these methods for medical image analysis, existing models face two major challenges when applied to colonoscopy image analysis. Firstly, the foreground (polyp) of colonoscopy images contains rich pathological information yet is often tiny compared with the background (intestine wall) and can be easily overwhelmed during training. Thus, naive generative models may generate realistic colonoscopy images but those images seldom contain polyp regions. In addition, in order to generate high-quality annotated samples, it is crucial to maintain the consistency between the polyp morphologies in synthesized images and the original masks, which current generative models struggle to achieve.

To tackle these issues and inspired by the remarkable success achieved by diffusion models in generating high-quality CT or MRI data [8,11,23], we creatively propose an effective adaptive refinement semantic diffusion model (ArSDM) to generate polyp-contained colonoscopy images while preserving the original annotations. The pipeline of the data generation and downstream task training is shown in Fig. 1. Specifically, we use the original segmentation masks as conditions to train a conditional diffusion model, which makes the generated samples share the same masks with the input images. Moreover, during diffusion model training, we employ an adaptive loss re-weighting method to assign loss weights for each input according to the size ratio of polyps and background, which addresses the overfitting problem for the large background. In addition, we fine-tune the diffusion model by minimizing the distance between the original ground

truth masks and the prediction masks from synthesis images via a pre-trained segmentation network. Thus the refined model could generate samples better aligned with the original masks.

In summary, our contributions are three-fold: (1) **Adaptive Refinement SDM**: Based on the standard semantic diffusion model [21], we propose a novel ArSDM with the adaptive loss re-weighting and the prediction-guided sample refinement mechanisms, which is capable of generating realistic polyp-contained colonoscopy images while preserving the original annotations. To the best of our knowledge, this is the first work for adapting diffusion models to colonoscopy image synthesis. (2) **Large-Scale Colonoscopy Generation**: The proposed approach can be used to generate large-scale datasets with no/arbitrary annotations, which significantly benefits the medical image society, laying the foundation for large-scale pre-training models in automatic colonoscopy analysis. (3) **Qualitative and Quantitative Evaluation**: We conduct extensive experiments to evaluate our method on five public benchmarks for polyp segmentation and detection. The results demonstrate that our approach could help deep learning methods achieve better performances. The source code is available at `https://github.com/DuYooho/ArSDM`.

2   Method

Background Denoising diffusion probabilistic models (DDPMs) [6] are classes of deep generative models, which have forward and reverse processes. The forward process is a Markov Chain that gradually adds Gaussian noise to the original data. This process can be formulated as the joint distribution $q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)$:

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) := \prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right), q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right), \ (1)$$

where $q\left(\mathbf{x}_0\right)$ is the original data distribution with $\mathbf{x}_0 \sim q\left(\mathbf{x}_0\right)$, $\mathbf{x}_{1:T}$ are latents with the same dimension of $\mathbf{x}_0$ and β_t is a variance schedule.

The reverse process is aiming to learn a model to reverse the forward process that reconstructs the original input data, which is defined as:

$$p_\theta\left(\mathbf{x}_{0:T}\right) := p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right), p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) := \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \sigma_t^2\mathbf{I}\right),$$
$$(2)$$

where $p\left(\mathbf{x}_T\right)$ is the noised Gaussian transition from the forward process at timestep T. In this case, we only need to use deep-learning models to represent $\boldsymbol{\mu}_\theta$ with θ as the model parameters. According to the original paper [6], the loss function can be simplified as:

$$\mathcal{L}_{\text{simple}} := \mathbb{E}_{t,\mathbf{x}_t,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\|^2\right]. \tag{3}$$

Thus, instead of training the model $\boldsymbol{\mu}_\theta$ to predict $\tilde{\boldsymbol{\mu}}_t$, we can train the model $\boldsymbol{\epsilon}_\theta$ to predict $\tilde{\boldsymbol{\epsilon}}$, which is easier for parameterization and learning.

Fig. 2. The overall architecture of **ArSDM**.

In this paper, we propose an adaptive refinement semantic diffusion model, a variant of DDPM, which has three key parts, *i.e.*, mask conditioning, adaptive loss re-weighting, and prediction-guided sample refinement. The overall illustration of our framework is shown in Fig. 2.

2.1   Mask Conditioning

Unlike the previous generative methods, our work aims to generate a synthetic image with an identical segmentation mask to the original annotation. To accomplish this, we adapt the widely used conditional U-Net architecture [21] in the reverse process, where the mask is fed as a condition. Specifically, for an input image $\mathbf{x}_0 \in \mathbb{R}^{H \times W \times C}$, $\mathbf{x}_t$ can be sampled at any timestep t with the closed form:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. It will be fed into the encoder $\mathcal{E}$ of the U-Net, and its corresponding mask annotation $\mathbf{c}_0 \in \mathbb{R}^{H \times W}$ will be injected into the decoder $\mathcal{D}$. The model output can be formulated as:

$$\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t, \mathbf{c}_0\right) = \mathcal{D}\left(\mathcal{E}\left(\mathbf{x}_t\right), \mathbf{c}_0\right). \tag{5}$$

Thus, the U-Net model $\boldsymbol{\epsilon}_\theta$ in Eq. 3 becomes $\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t, \mathbf{c}_0\right)$, and the loss function in Eq. 3 is changed to:

$$\mathcal{L}_{\text{condition}} = \mathbb{E}_{t, \mathbf{x}_t, \mathbf{c}_0, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t, \mathbf{c}_0\right)\right\|^2\right]. \tag{6}$$

Algorithm 1: One training iteration of ArSDM

Input: $t \sim \text{Uniform}(\{1, ..., T\})$, $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, $\mathbf{c}_0$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Output: $\tilde{\boldsymbol{\epsilon}}$, $\tilde{\mathbf{c}}_0$

1 $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$; $\tilde{\mathbf{x}}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c}_0)$

2 **for** $i = t, ..., 1$ **do**

3    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $i > 1$, else $\mathbf{z} = \mathbf{0}$; $\tilde{\mathbf{x}}_{i-1} = \frac{1}{\sqrt{\alpha_i}}\left(\tilde{\mathbf{x}}_i - \frac{1-\alpha_i}{\sqrt{1-\bar{\alpha}_i}}\boldsymbol{\epsilon}_\theta(\tilde{\mathbf{x}}_i, i, \mathbf{c}_0)\right) + \sigma_i\mathbf{z}$

4 **end for**

5 $\tilde{\mathbf{c}}_0 = \mathcal{P}(\tilde{\mathbf{x}}_0)$

6 Take gradient descent step on $\nabla_\theta \mathcal{L}_{\text{total}}$

2.2 Adaptive Loss Re-Weighting

The polyp regions in the colonoscopy images differ from the background regions, which contain more pathological information and should be adequately treated to learn a better model. However, training the diffusion models using the original loss function ignores the difference between different regions, where each pixel shares the same weights when calculating the loss. This would lead to the model generating more background-like polyps since the large background region will easily overwhelm the small foreground polyp regions during training. A simple way to alleviate this problem is to apply a weighted loss function that assigns the polyp and background regions with different weights. However, most polyps vary a lot in size and shape. Thus assigning constant weights for all polyps exacerbated the imbalance problem. In this case, to tackle this problem, we propose an adaptive loss function that vests different weights according to the size ratio of the polyp over the background. Specifically, we define a pixel-wise weights matrix $W^\lambda \in \mathbb{R}^{H \times W}$ with each entry $w^\lambda_{i,j}$ to be:

$$w^\lambda_{i,j} = \begin{cases} 1 - r & , \ p = 1 \\ r & , \ p = 0 \end{cases}, \qquad r = \frac{\#(p = 1)}{H \times W}, \tag{7}$$

where $p = 1$ means the pixel p at (h, w) belongs to the polyp region and $p = 0$ means it belongs to the background region. Thus, the loss function becomes:

$$\mathcal{L}_{\text{adaptive}} = \mathbb{E}_{t, \mathbf{x}_t, \mathbf{c}_0, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}\left[W^\lambda \cdot \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c}_0)\|^2\right]. \tag{8}$$

2.3 Prediction-Guided Sample Refinement

The downstream tasks of polyp segmentation and detection require rich semantic information on polyp regions to train a good model. Through extensive experiments, we found inaccurate sample images with coarse polyp boundary that is not aligned properly with the original masks may introduce large biases and noises to the datasets. The model can be confused by several conflicting training images with the same annotation. To this end, we design a refinement strategy that uses

Table 1. Comparisons of different settings applied on three polyp segmentation baselines.

Methods	EndoScene		ClinicDB		Kvasir		ColonDB		ETIS		Overall	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
PraNet	87.1	79.7	89.9	84.9	89.8	84.0	70.9	64.0	62.8	56.7	74.0	67.5
+LDM	83.7	76.9	88.2	83.5	88.4	83.0	62.6	56.0	56.2	50.3	67.8	61.7
+SDM	**89.9**	**83.2**	89.2	83.7	88.4	82.6	74.2	66.5	66.4	60.3	76.4	69.6
+**Ours**	89.7	82.7	**93.3**	**88.5**	**89.9**	**84.5**	**76.1**	**68.9**	**75.5**	**68.1**	**80.0**	**73.2**
SANet	88.8	81.5	**91.6**	85.9	90.4	84.7	75.3	67.0	75.0	65.4	79.4	71.4
+LDM	72.7	60.5	88.8	82.8	88.7	82.7	64.3	55.4	58.0	49.2	68.3	59.8
+SDM	**90.2**	83.0	89.9	84.1	90.9	85.4	77.6	69.3	74.7	66.8	80.4	72.9
+**Ours**	**90.2**	**83.2**	91.4	**86.1**	**91.1**	**85.6**	**77.7**	**70.0**	**78.0**	**69.5**	**81.5**	**74.1**
PVT	**90.0**	**83.3**	93.7	88.9	**91.7**	**86.4**	80.8	72.7	78.7	70.6	83.3	76.0
+LDM	88.2	81.2	92.3	87.1	91.2	85.7	78.7	70.4	78.0	69.6	81.9	74.2
+SDM	88.8	81.7	**93.9**	**89.2**	91.2	86.1	81.3	73.5	78.7	71.1	83.4	76.3
+**Ours**	88.2	81.2	92.2	87.5	91.5	86.3	**81.7**	**73.8**	**80.6**	**72.9**	**84.0**	**76.7**

the prediction of a pre-trained segmentation model on the sampled images to guide the training process and restore the proper polyp boundary information. Specifically, at each iteration of training, the output $\tilde{\epsilon} = \epsilon_\theta\left(\mathbf{x}_t, t, \mathbf{c}_0\right)$ will go into the sampler to generate sample image $\tilde{\mathbf{x}}_0$. Then, we take the sample image as the input of the segmentation model to predict the pseudo masks $\tilde{\mathbf{c}}_0$. We propose the following refinement loss based on IoU loss and binary cross entropy (BCE) loss between $\tilde{\mathbf{c}}_0$ and $\mathbf{c}_0$. The refinement loss is:

$$\mathcal{L}_{\text{refine}} = \mathcal{L}(\mathbf{c}, \tilde{\mathbf{c}}_g) + \sum_{i=3}^{i=5} \mathcal{L}\left(\tilde{\mathbf{c}}_i\right),$$
$$\tilde{\mathbf{c}}_0 = \{\tilde{\mathbf{c}}_3, \tilde{\mathbf{c}}_4, \tilde{\mathbf{c}}_5, \tilde{\mathbf{c}}_g\} = \mathcal{P}\left(\mathcal{S}\left(\tilde{\epsilon}\right)\right), \tag{9}$$

where $\mathcal{L} = \mathcal{L}_{IoU} + \mathcal{L}_{BCE}$ is the sum of the IoU loss and BCE loss, $\tilde{\mathbf{c}}_0$ is the collection of the three side-outputs $(\tilde{\mathbf{c}}_3, \tilde{\mathbf{c}}_4, \tilde{\mathbf{c}}_5)$ and the global map $\tilde{\mathbf{c}}_g$ as described in [5]. $\mathcal{P}(\cdot)$ represents the PraNet model and $\mathcal{S}(\cdot)$ is the DDIM [16] sampler. The detailed procedure of one training iteration is shown in Algorithm 1 and the overall loss function is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adaptive}} + \mathcal{L}_{\text{refine}}. \tag{10}$$

3   Experiments

3.1   ArSDM Experimental Settings

We conducted our experiments on five public polyp segmentation datasets: EndoScene [20], CVC-ClincDB/CVC-612 [1], CVC-ColonDB [18], ETIS [14] and Kvasir [7]. Following the standard of PraNet, 1,450 image-mask pairs from Kvasir

Table 2. Comparisons of different settings applied on three polyp detection baselines.

Methods	EndoScene		ClinicDB		Kvasir		ColonDB		ETIS		Overall	
	AP	F1	AP	F1	AP	F1	AP	F1	AP	F1	AP	F1
Center.	86.9	**91.4**	84.7	89.2	75.6	81.4	62.2	72.3	62.7	70.1	56.6	76.0
+LDM	84.1	84.4	**90.4**	89.9	81.3	81.8	73.4	74.5	65.2	71.7	62.0	76.9
+SDM	**87.8**	86.9	88.7	**91.0**	77.0	82.8	71.8	78.1	68.2	72.6	61.8	79.1
+**Ours**	85.0	89.1	86.1	90.8	**77.3**	**84.7**	**74.2**	**80.2**	**68.7**	**75.6**	**65.7**	**81.3**
Sparse.	89.9	87.8	81.4	86.4	75.6	80.2	78.2	73.2	63.8	62.4	63.7	73.2
+LDM	87.4	76.3	**95.0**	**93.5**	81.5	58.8	80.0	71.0	64.4	54.3	65.3	66.3
+SDM	**94.5**	**90.5**	88.7	86.5	79.0	80.5	**81.4**	76.8	67.8	67.1	65.2	76.7
+**Ours**	92.8	86.2	92.2	90.6	**81.6**	**82.3**	80.1	**79.8**	**72.4**	**70.4**	**66.4**	**79.0**
Deform.	**98.1**	**94.4**	89.7	89.9	**80.2**	74.4	**82.2**	75.5	65.3	54.7	64.5	71.8
+LDM	94.6	90.5	91.6	89.5	79.3	73.4	78.0	73.2	69.0	64.0	63.4	73.3
+SDM	96.0	90.6	90.3	91.2	82.2	78.9	80.1	75.1	67.5	66.7	65.1	75.8
+**Ours**	94.7	94.3	**92.3**	**92.0**	80.0	**80.3**	81.4	**77.3**	**74.1**	**69.3**	**67.9**	**77.9**

and CVC-ClinicDB are taken as the training set. The evaluations are conducted on the five datasets separately to verify the learning and generalization capability. The training image-mask pairs are padded to have the same height and width and then resized to the size of 384×384. Experiments with prediction-guided sample refinement are trained with around one-half NVIDIA A100 days, while others are trained with approximately one day for convergence. We use the DDIM sampler with a maximum timestep of 200 for sampling images.

3.2   Downstream Experimental Settings

We conduct the evaluation of our methods and the state-of-the-art counterparts on polyp segmentation and detection tasks. We generated the same number of samples as the diffusion training set using the original masks, and then combined them to create a new downstream training set. We employed PraNet [5], SANet [22], and Polyp-PVT [2] as baseline segmentation models with default settings, and evaluated them using mean Intersection over Union (IoU) and mean Dice metrics. For detection, we selected CenterNet [30], Sparse-RCNN [17], and Deformable-DETR [31] as baseline models with the same settings as the original papers, and evaluated them using Average Precision (AP) and F1-scores.

3.3   Quantitative Comparisons

The experimental results presented in Table 1 and 2 demonstrate the effectiveness of our proposed method in training better downstream models to achieve superior performance. Specifically, data generated by our approach assists the significant improvements for each model in mDice and mIoU, with increases of 6.0% and 5.7% over PraNet, 2.1% and 2.7% over SANet, and 0.7% and 0.7% over Polyp-PVT. We

Table 3. Ablation study of different components on polyp segmentation tasks.

Methods		PraNet		SANet	
Ada.	Ref.	mDice	mIoU	mDice	mIoU
✗	✗	76.4	69.6	80.4	72.9
✓	✗	79.1	72.4	80.5	72.8
✗	✓	78.5	71.5	81.1	73.2
✓	✓	**80.0**	**73.2**	**81.5**	**74.1**

Table 4. Ablation study of different components on polyp detection tasks.

Methods		CenterNet		Sparse.	
Ada.	Ref.	AP	F1	AP	F1
✗	✗	61.8	79.1	65.2	76.7
✓	✗	62.2	80.1	65.8	77.2
✗	✓	64.0	80.4	66.0	77.6
✓	✓	**65.7**	**81.3**	**66.4**	**79.0**

Fig. 3. Illustration of generated samples with the corresponding masks and original images for comparison reference.

also observe superior AP and F1-scores compared to CenterNet, Sparse-RCNN, and Deformable-DETR trained with original data, with gains of 9.1% and 5.3%, 2.7% and 5.8%, and 3.4% and 6.1%, respectively. Moreover, we conducted a comprehensive comparison with SOTA models, noting that these models were not specifically designed for colonoscopy images and may generate data that hinder the training process or lack the ability for effective improvement. Nevertheless, our experimental results confirm the superiority of our proposed method.

Ablation Study We conducted an ablation study to assess the importance of each proposed component. Table 3 and 4 report the overall accuracies on the test set. The results demonstrate both components contribute to the accuracy improvement of baseline models, indicating their essential roles in achieving the best final performance.

3.4   Qualitative Analyses

To further investigate the generative performance of our approach, we present visualization results in Fig. 3, which displays the generated samples and their corresponding masks, alongside the original images for reference. The generated samples demonstrate differences from the original images in both the polyp regions and the backgrounds while maintaining alignment with the masks. Additionally, we sought evaluations from medical professionals to assess the authenticity of the generated samples, and non-medical professionals to locate polyps in the images, which yielded positive feedback on the quality of the generated samples.

4   Conclusion

Automatic generation of annotated data is essential for colonoscopy image analysis, where the scale of existing datasets is limited by the expertise and time required for manual annotation. In this paper, we propose an adaptive refinement semantic diffusion model (ArSDM) for generating colonoscopy images while preserving annotations by introducing innovative adaptive loss re-weighting and prediction-guided sample refinement mechanisms. To evaluate our approach comprehensively, we conduct polyp segmentation and detection experiments on five widely used datasets, where experimental results demonstrate the effectiveness of our approach, in which model performances are greatly enhanced with little synthesized data.

Acknowledgement

References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015) 6
2. Bo, D., Wenhai, W., Deng-Ping, F., Jinpeng, L., Huazhu, F., Ling, S.: Polyp-pvt: Polyp segmentation with pyramidvision transformers (2021) 7
3. Chaitanya, K., Karani, N., Baumgartner, C.F., Erdil, E., Becker, A., Donati, O., Konukoglu, E.: Semi-supervised task-driven data augmentation for medical image segmentation. Medical Image Analysis **68**, 101934 (2021) 2
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021) 2
5. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020) 2, 6, 7
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020) 2, 3
7. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. pp. 451–462. Springer (2020) 6

8. Kim, B., Ye, J.C.: Diffusion deformable model for 4d temporal medical image generation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I. pp. 539–548. Springer (2022) 2

9. Ma, Y., Liu, Y., Cheng, J., Zheng, Y., Ghahremani, M., Chen, H., Liu, J., Zhao, Y.: Cycle structure and illumination constrained gan for medical image enhancement. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23. pp. 667–677. Springer (2020) 2

10. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019) 2

11. Pinaya, W.H., Graham, M.S., Gray, R., Da Costa, P.F., Tudosiu, P.D., Wright, P., Mah, Y.H., MacKinnon, A.D., Teo, J.T., Jager, R., et al.: Fast unsupervised brain anomaly detection and segmentation with diffusion models. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII. pp. 705–714. Springer (2022) 2

12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022) 2

13. Sandfort, V., Yan, K., Pickhardt, P.J., Summers, R.M.: Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. Scientific reports **9**(1), 16884 (2019) 2

14. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International journal of computer assisted radiology and surgery **9**, 283–293 (2014) 6

15. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015) 2

16. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021) 6

17. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14454–14463 (2021) 7

18. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging **35**(2), 630–644 (2015) 6

19. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Transactions on Medical Imaging (2016) 2

20. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering **2017** (2017) 6

21. Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., Li, H.: Semantic image synthesis via diffusion models. arXiv preprint arXiv:2207.00050 (2022) 2, 3, 4

22. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: MICCAI. pp. 699–708. Springer (2021) 2, 7

23. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: MICCAI. pp. 35–45. Springer (2022) 2
24. Wu, L., Hu, Z., Ji, Y., Luo, P., Zhang, S.: Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation. Lecture Notes in Computer Science (2021) 2
25. Xu, J., Anwar, S., Barnes, N., Grimpen, F., Salvado, O., Anderson, S., Armin, M.A.: Ofgan: Realistic rendition of synthetic colonoscopy videos. In: MICCAI. pp. 732–741. Springer (2020) 2
26. Zhang, R., Lai, P., Wan, X., Fan, D.J., Gao, F., Wu, X.J., Li, G.: Lesion-aware dynamic kernel for polyp segmentation. In: MICCAI. pp. 99–109. Springer (2022) 2
27. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: MICCAI. pp. 253–262. Springer (2020) 2
28. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8543–8553 (2019) 2
29. Zhao, X., Wu, Z., Tan, S., Fan, D.J., Li, Z., Wan, X., Li, G.: Semi-supervised spatial temporal attention network for video polyp segmentation. In: MICCAI. pp. 456–466. Springer (2022) 2
30. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019) 7
31. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020) 7